RESEARCH ARTICLE

Glycosyltransferase Family 43 Is Also Found in Early Eukaryotes and Has Three Subfamilies in Charophycean Green Algae

Rahil Taujale, Yanbin Yin*

Department of Biological Sciences, Northern Illinois University, DeKalb, Illinois, United States of America

* yyin@niu.edu

Abstract

The glycosyltransferase family 43 (GT43) has been suggested to be involved in the synthesis of xylans in plant cell walls and proteoglycans in animals. Very recently GT43 family was also found in Charophycean green algae (CGA), the closest relatives of extant land plants. Here we present evidence that non-plant and non-animal early eukaryotes such as fungi, *Haptophyceae*, *Choanoflagellida*, *Ichthyosporea* and *Haptophyceae* also have GT43-like genes, which are phylogenetically close to animal GT43 genes. By mining RNA sequencing data (RNA-Seq) of selected plants, we showed that CGA have evolved three major groups of GT43 genes, one orthologous to IRX14 (IRREGULAR XYLEM14), one orthologous to IRX9/IRX9L and the third one ancestral to all land plant GT43 genes. We confirmed that land plant GT43 has two major clades A and B, while in angiosperms, clade A further evolved into three subclades and the expression and motif pattern of A3 (containing IRX9) are fairly different from the other two clades likely due to rapid evolution. Our in-depth sequence analysis contributed to our overall understanding of the early evolution of GT43 family and could serve as an example for the study of other plant cell wall-related enzyme families.

Citation: 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(5): e0128409. doi:10.1371/journal.pone.0128409

Academic Editor: Joshua L Heazlewood, The University of Melbourne, Australia

Received: January 28, 2015

Accepted: April 27, 2015

Published: May 29, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the start-up funds and the research & artistry award to YY from Northern Illinois University. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Xylans are the second most abundant polysaccharides in plant cell walls. Much progress has been made in elucidating the biosynthesis of xylans in model plant organisms and has been most recently reviewed by [1]. Briefly using genetic screening mutants with irregular xylem phenotype, two GT43 proteins (AtIRX9 and AtIRX14) and one GT47 protein (AtIRX10) in *Arabidopsis thaliana* were suggested to be involved in the synthesis of the xylan backbone [2–6]. Further studies have shown that their close paralogous proteins (IRX9-like, IRX14-like and IRX10-like) might function redundantly [2, 3]. A recent study hypothesized that IRX9, IRX14 and IRX10 might work together as a large protein complex [1, 7]. Lastly, other GT proteins particularly those of GT47 and GT8 families, were also indicated to play a role in xylan backbone synthesis [2, 8, 9].

In addition to functional characterizations, phylogenetic and sequence motif studies of GT43 proteins have also been published to understand the evolution and origin of the GT43 family. Two reports have included non-plant proteins in their analyses and demonstrated that plant and animal proteins are phylogenetically well separated [10, 11]. One of the reports showed that three major GT43 clades (monophyletic clusters) exist in land plants [10]: A contains AtIRX9L, B contains AtIRX14/14L and C contains AtIRX9. More recent studies in early plants suggested that there are only two major land plant GT43 clades [12, 13], as AtIRX9 and AtIRX9L were grouped into one single clade named A. Early land plants (moss and spike moss) [12] and Charophycean green algae (CGA) [13] were shown to have orthologous genes in each of the two clades. We have previously identified a single chlorophyte green algal GT43 homolog (GenBank ID: XP_003063890.1) from *Micromonas pusilla* CCMP1545 using a hidden Markov model (HMM)-based approach [14], and have demonstrated that clade A could be further split into three subclades, with AtIRX9 and AtIRX9L in two separate subclades and the third subclade containing no Arabidopsis genes.

Clearly there exist some disagreements in the literature with regard to the phylogenetic classification of the GT43 family. One of the possible reasons might be that the analyzed data in previous reports were from different organisms: e.g. some only used land plants, while others also included CGA or even animals. We believe that including proteins of broader organismic groups in the phylogeny could help improve the classification resolution. In addition, although AtIRX14/14L clade (B) and AtIRX9/9L clade (A) must have already separated since CGA [13], it is unknown when AtIRX9 and AtIRX9L separated. CGA transcriptomes have been recently mined for different cell wall-related GT families [13]; however only a very small number of GT43 hits were phylogenetically analyzed due to short sequence lengths of assembled transcripts. Therefore it remains unanswered how many GT43 clades exist in CGA and how are they evolutionarily related to land plant clades. Furthermore, since a distant GT43 homolog was found in chlorophyte green algae [14] that evolved earlier than CGA, it is possible that GT43 genes might be present in other early organisms.

In recent years, the study of the evolution of plant cell walls has gained considerable interest [15–17], as it can assist in understanding the complexity of cell wall compositions across different plant lineages and the information used to develop transgenic plants with improved biomass. Here we sought to further the evolutionary study of the GT43 family among all cellular kingdoms, as it is difficult to believe that GT43's are absent in more ancient early non-plant/algal/animal eukaryotes and prokaryotes. We aimed to obtain a more robust phylogenetic classification of plant/algal GT43 genes by including more representative species especially those without complete genomes such as liverworts, hornworts, ferns and CGA. We used more sensitive bioinformatics approaches combining sequence homology searches, conserved motif scanning and expression profiles to better understand the evolutionary divergence of the GT43 family.

Results and Discussion

GT43-like proteins are found in non-plant/animal early eukaryotes but not in prokaryotes

In order to look for non-*Streptophyta* and non-*Metazoa* GT43 homologs, we have performed an exhaustive HMMER search (www.hmmer.org) against the NCBI non-redundant protein database (NCBI-nr), the largest protein sequence database containing both experimentally and computationally determined proteins. The query is an HMM profile of the GT43 family downloaded from dbCAN database [18]. In total 21 non-*Streptophyta* and non-*Metazoa* protein homologs (Table 1, including the previously reported XP_003063890.1 of chlorophyte *M. pusilla*

Table 1. 21 NCBI-nr proteins that are non-plant and non-animal proteins and hit the GT43 domain with an E-value < 1e-10; GenBank identifiers (IDs) having at least 3 motifs are shown in bold and italic.

GenBank ID	Length	Species	Taxonomic group	GT43 domain E-value	GenBank ID of best BLAST hit in CAZy database	E-value of best hit in CAZy database	Total # of motifs *
XP_004345176.1	**390**	***Capsaspora owczarzaki ATCC 30864***	***Ichthyosporea***	**1.50E-59**	**CAI62038.1**	**4.80E-25**	**6**
XP_001744836.1	**305**	***Monosiga brevicollis MX1***	***Choanoflagellida***	**3.60E-56**	**AFE77696.1**	**6.80E-42**	**6**
XP_001742263.1	**243**	***Monosiga brevicollis MX1***	***Choanoflagellida***	**1.80E-48**	**AEO35548.1**	**6.00E-39**	**6**
XP_001747805.1	**280**	***Monosiga brevicollis MX1***	***Choanoflagellida***	**7.50E-48**	**AAI69407.1**	**4.10E-37**	**6**
XP_007514541.1	**352**	***Bathycoccus prasinos***	***Chlorophyta***	**9.30E-28**	**CAI62044.1**	**2.00E-09**	**5**
XP_004343171.1	**458**	***Capsaspora owczarzaki ATCC 30864***	***Ichthyosporea***	**2.30E-51**	**CAI63873.1**	**2.90E-18**	**4**
XP_004989824.1	**324**	***Salpingoeca rosetta***	***Choanoflagellida***	**1.80E-50**	**AAH78400.1**	**3.00E-31**	**4**
EMS23236.1	**318**	***Rhodosporidium toruloides NP11***	***Fungi***	**1.90E-33**	**CAA15837.2**	**1.40E-14**	**4**
EGU10941.1	**318**	***Rhodotorula glutinis ATCC 204091***	***Fungi***	**1.70E-32**	**CAA15837.2**	**8.80E-15**	**4**
XP_005794278.1	**290**	***Emiliania huxleyi CCMP1516***	***Haptophyceae***	**5.00E-34**	**ACF84760.1**	**2.80E-08**	**3**
XP_003336254.2	**221**	***Puccinia graminis f sp tritici CRL 75-36-700-3***	***Fungi***	**2.40E-24**	**CAB04033.2**	**1.30E-07**	**3**
XP_003336238.1	472	*Puccinia graminis f sp tritici CRL 75-36-700-3*	*Fungi*	1.80E-23	CAB04033.2	4.40E-08	2
AIA87534.1	116	*uncultured bacterium*	*Bacteria*	1.90E-13	AAI21201.1	4.40E-13	2
XP_004345780.1	370	*Capsaspora owczarzaki ATCC 30864*	*Ichthyosporea*	6.30E-37	AAI21202.1	4.50E-14	1
EFW41058.2	407	*Capsaspora owczarzaki ATCC 30864*	*Ichthyosporea*	8.30E-37	AAI21202.1	8.10E-14	1
XP_005769385.1	547	*Emiliania huxleyi CCMP1516*	*Haptophyceae*	2.80E-31	CAI93173.1	1.90E-08	1
XP_003063890.1	227	*Micromonas pusilla CCMP1545*	*Chlorophyta*	2.30E-22	CAI63866.1	6.40E-05	1
AIA83381.1	112	*uncultured bacterium*	*Bacteria*	3.00E-14	CAI63870.1	3.20E-25	1
XP_003336242.2	158	*Puccinia graminis f sp tritici CRL 75-36-700-3*	*Fungi*	2.80E-11	ABF18200.1	1.80E-03	1
XP_003325987.2	529	*Puccinia graminis f sp tritici CRL 75-36-700-3*	*Fungi*	9.90E-21	CAI68027.1	1.10E-03	0
XP_007416303.1	119	*Melampsora larici-populina 98AG31*	*Fungi*	4.40E-13	CAI68027.1	1.20E-02	0

* Motif details are available in S1 Table

doi:10.1371/journal.pone.0128409.t001

CCMP1545 with E-value = 2.3e-22) were found using a very conservative cutoff E-value < 1e-10. A BLASTP search of these 21 proteins against the CAZy database [19] showed that all candidates had a GT43 protein as their best hit.

If using a more relaxed cutoff E-value < 1e-5, 428 additional non-*Streptophyta* and non-*Metazoa* hits were found (i.e. with E-value between 1e-5 and 1e-10), 380 of which are from prokaryotes, 27 are from fungi, 14 are from other eukaryotes (*Alveolata*, *Euglenozoa*, *Chlorophyta*) and 8 are from viruses. However a BLASTP search against the CAZy database showed that only four of the 428 proteins had a GT43 protein as their best hit, suggesting that most of

Table 2. PROSITE [20] style patterns of 8 conserved sequence motifs found in GT43 proteins.

Motif	PROSITE pattern	Position in human GlcAT-I *
1	T-P-[TI]-[YI]	81–84
2	W-[IL]-[ILV]-[VIA]-E-[DAKG]	108–113
3	[QMN]-R-[NL]	160–162
4	D-D-[DS]-N	194–197
5	[EQ]-[GA]-P	227–229
6	[ILVM]-[DEH]-[MWI]-[AS]-[GS]-F	251–256
7	[QLN]-[DE]-[SNT]	280–282
8	W-[HRNW]-[LT]-[RQKH]	307–310

* Details are available in S2 Table and Fig 1

doi:10.1371/journal.pone.0128409.t002

them are false positives. Furthermore a manual examination of the four proteins (all from *Trypanosoma cruzi* of *Euglenozoa*) by inspecting NCBI Blink (pre-computed BLAST result against NCBI-nr) pages and conserved domain search pages suggested that none of them seem to be GT43 like.

To be more conservative, we further scanned the 21 protein sequences (Table 1 and detailed version in S1 Table) for the presence of 8 PROSITE-style [20] motif patterns shown in Table 2. These 8 patterns were derived by referring to a previous report [11] covering 17 key residues (S2 Table and Fig 1) interacting with the substrates of a structurally solved human GT43 protein named GlcAT-I [21]. Because these motif patterns were defined very strictly, our automatic motif scanning approach is very conservative and having these motifs is a strong support for a protein to be a GT43 protein. Finally 11 out of the 21 proteins have at least 3 of the 8 motifs and all of the 11 proteins have GT43 domains with very significant E-values (between 2.4e-24 and 1.5e-59).

These 11 proteins (bold and italic GenBank IDs in Table 1) belong to eight species of five taxonomic groups: *Choanoflagellida*, *Ichthyosporea*, *Fungi*, *Chlorophyta* and *Haptophyceae*. The other 10 of the 21 proteins (Table 1) have less than 3 motifs but are also from these five taxonomic groups, except for two proteins (AIA87534.1 and AIA83381.1) that are annotated as being derived from uncultured bacteria. Both AIA87534.1 and AIA83381.1 are short in sequence (< 120 aa) and NCBI Blink pages show that the top 100 hits of both proteins are all from *Metazoa*. Given that the two proteins were derived from intestinal metagenomes of snout beetle (*Rhynchophorus ferrugineus*) [22], it is likely that they are actually encoded by beetle genes but were mis-annotated as deriving from uncultured intestinal bacteria. As mentioned above, 380 prokaryotic proteins have E-values between 1e-5 and 1e-10; however Blink and conserved domain examination indicated that their top hits are not GT43 proteins. Hence we can conclude that no prokaryotic GT43 homologs were found in our search and GT43 is a family restricted to eukaryotes.

We believe that the absence of GT43 homologs in prokaryotes can be explained as follows. GT43 might have evolved from other GT families in that many GT families, including GT43, GT2 and GT8, are classified into the GT-A superfamily in the CAZy database [19]. The reason that we do not see homology between eukaryotic GT43 proteins and prokaryotic GT-A proteins is because they have diverged too much in sequence although the structural similarity can still be detected. In fact, of the aforementioned 428 additional protein hits with E-value between 1e-5 and 1e-10, 67 bacterial proteins matched other GT proteins (excluding GT43) as their best hit in the CAZy database, among which 42 had a GT2 as their best hit.

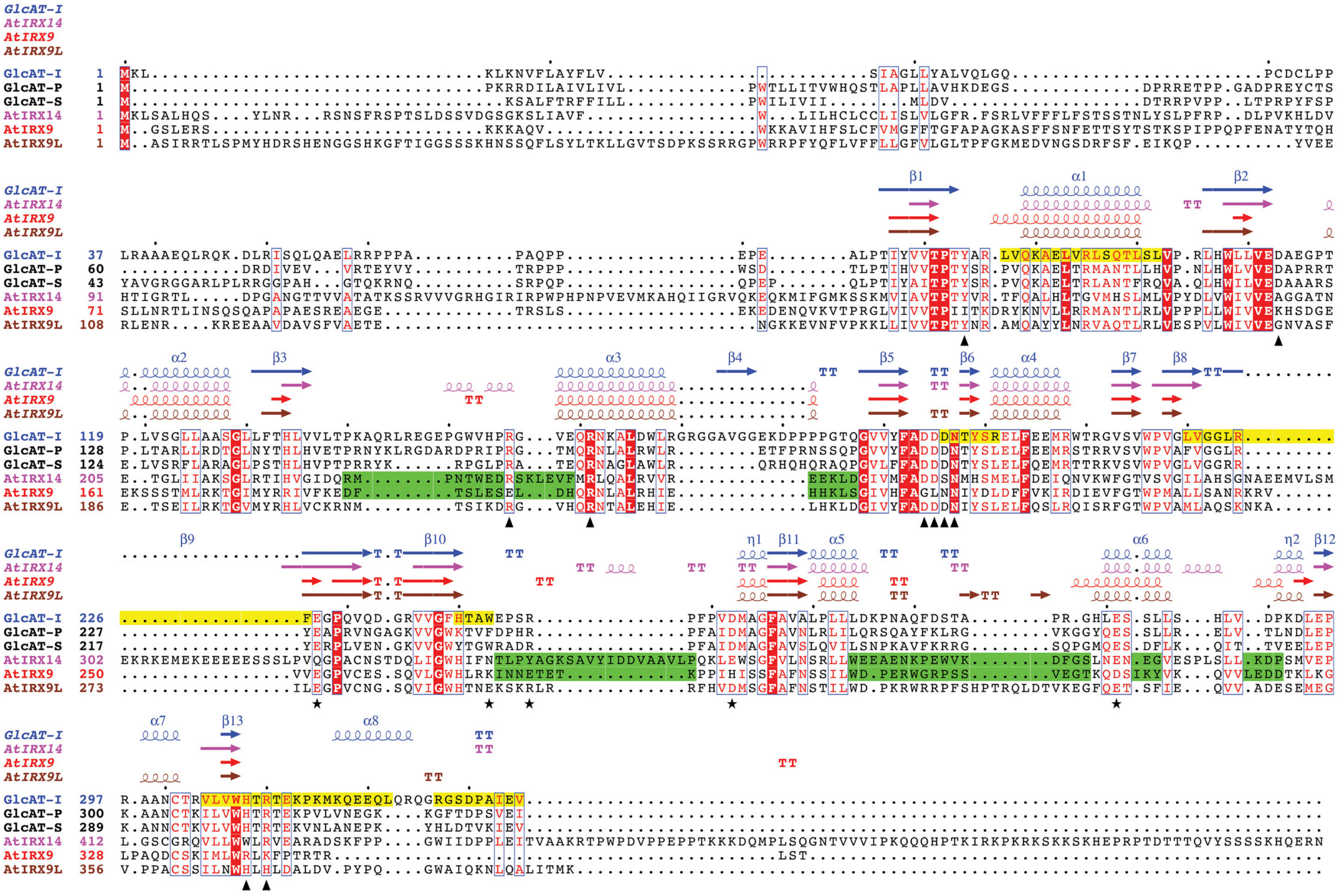

Fig 1. Sequence alignment of human and Arabidopsis GT43 proteins. Human proteins GlcAT-I (UniProt ID: O94766), GlcAT-P (Q9P2W7), GlcAT-S (Q9NPZ5) and Arabidopsis proteins AtIRX9, AtIRX9L and AtIRX14 were used to build the alignment using MAFFT program [41]. The alignment graph was generated using ESPript server [45], which also took the PDB format files of GlcAT-I (PDB ID: 1kws), AtIRX9 (predicted), AtIRX9L (predicted) and AtIRX14 (predicted) to display their secondary structures. Arrows and stars below the alignment indicate residues in GlcAT-I structure that interact with UDP-GlcA (sugar donor) and the trisaccharide Galβ1-3Galβ1-4Xyl (sugar acceptor). Yellow shaded regions indicate protein-protein interaction (PPI) regions reported in GlcAT-I structure [21]. Green shaded regions indicate weakly aligned regions that do not overlap with PPI regions.

doi:10.1371/journal.pone.0128409.g001

GT43 proteins are found in CGA, liverworts, hornworts and ferns

Using the same approach, we mined RNA sequencing (RNA-Seq) data in GenBank generated from 5 liverworts, 2 hornworts, 3 ferns and 15 CGA (S3 Table). Although these plants are expected to have GT43 genes, we wanted to prove that GT43 genes are indeed present in these plant taxa and then further study how they are phylogenetically related to flowering plant GT43 genes. We modified a computational pipeline that we developed recently [23] to assemble RNA-Seq reads species by species and then mined these assembled transcripts for GT43 homologs (S1 Fig).

Particularly for CGA, a previous report [13] used pre-assembled expressed sequence tag (EST) data. We on the other hand started with the raw RNA-Seq reads because pre-assembled ESTs usually only cover a subset of the original RNA-Seq reads. In addition, the previous report used ESTs sequenced mostly by Sanger and 454 technologies. We however employed ~166 million reads (S3 Table) sequenced using the Illumina technology, an approach that is known to have increased sequencing depth. These together determined that we would be able to retrieve more GT43 homologs that might have been missed by the previous analysis.

Using an E-value < 1e-10 in a HMMER search, 54 transcript hits were obtained. A BLASTP search of these 54 transcripts against the CAZy database showed that all of them hit a GT43 protein as the best match. After further applying the motif filter, 29 peptide sequences translated from the 54 transcripts (S4 Table), including 6 from liverworts, 3 from hornworts, 3 from ferns and 17 from CGA, have at least 3 of the 8 motifs. A total of 21 have an amino acid sequence length > 180.

Phylogenetic classification of GT43 family in all organismic groups

In order to understand how the newly identified GT43 homologs are phylogenetically related to land plant and animal proteins, we combined sequences of 12 early eukaryotic proteins (Table 1, 11 proteins with 3+ motifs plus XP_003063890.1 of the chlorophyte green alga), 29 early plant peptides (S4 Table) as well as 33 plant and animal proteins selected from CAZy and PlantCAZyme [24] databases to build a maximum likelihood phylogeny (see details in Methods). All but 3 of the 33 plant and animal proteins have at least 3 of the 8 motifs outlined in Table 2.

It should be noted that XP_003063890.1 hit the GT43 HMM with an E-value = 2.3e-22 and also CAI63866.1 (a GT43 protein) of the CAZy database as the best hit (6.4e-05) (Table 1). It has less than 3 of the 8 motifs due to its relatively short sequence length (227 aa). We included it in the phylogenetic analysis because it has already been reported [14]. As expected, it was stably clustered with another chlorophyte green algal protein (XP_007514541.1 from *Bathycoccus prasinos*) in the phylogeny (Fig 2). Since both proteins are computationally predicted, future experimental study is needed to confirm whether these chlorophyte proteins are indeed expressed and functional in these chlorophyte algae.

We have verified the reliability of this phylogeny (Fig 2) by including all 54 homologs in (S4 Table) and by editing the sequence alignment to remove regions containing large gaps and ambiguously aligned regions. This effectively eliminated the effect of short fragments on the resulting phylogeny. The new phylogeny (S2 Fig) looks very similar to the one in Fig 2, suggesting that including short fragments did not significantly change the phylogenetic clustering.

Classification of GT43 proteins of various organisms into four major clades

In order to follow the previous nomenclature used in the plant literature (i.e. A for AtIRX9/9L and B for AtIRX14/14L) and incorporate the newly identified GT43 proteins in various organisms, we suggest a classification of the GT43 family into four clades: A, B, C and D, according to Fig 2. Clade A and B contain land plants and CGA proteins (detailed below), following the two recent reports [12, 13]. These two clades are clustered together with a bootstrap value = 76%, with clade C basal to them. Clade C and D are newly designated clades. Clade C contains exclusively CGA proteins and two *Chlorophyta* proteins, although the grouping of the *Chlorophyta* proteins with CGA proteins is not well supported (bootstrap value < 50%). Clade D only contains non-plant proteins including those from animals, fungi, *Haptophyceae*, *Choanoflagellida*, *Ichthyosporea* and *Haptophyceae*. The function of clade D proteins might be related to the synthesis of proteoglycans, as human GT43 proteins have been known to be involved in the synthesis of a surface carbohydrate epitope HNK1 and glycosaminoglycan using UDP-glucuronic acid as the sugar donor [25]. In contrast proteins of the plant clades A and B are

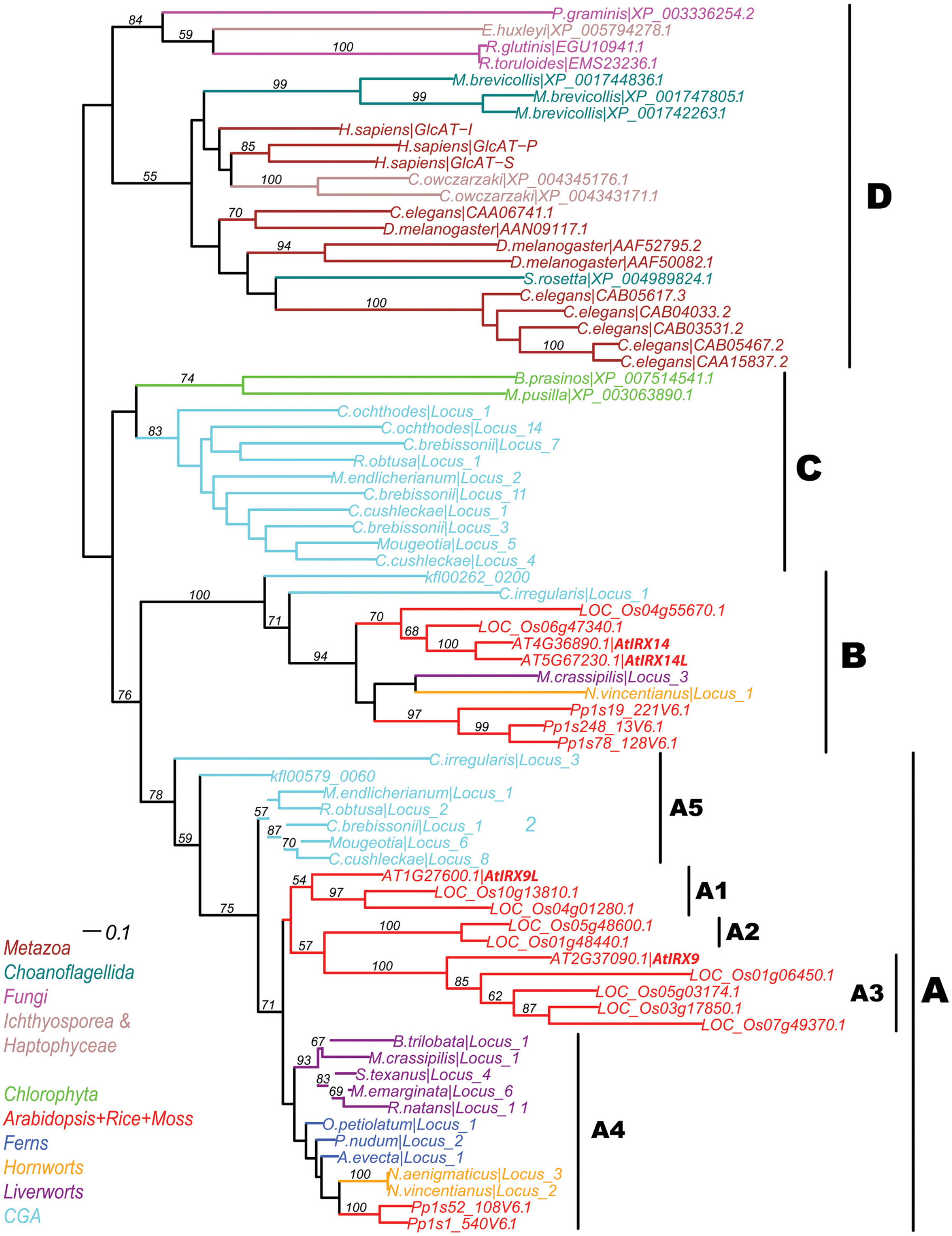

Fig 2. Phylogeny of 74 GT43 proteins. 21 plant proteins are from the PlantCAZyme database, 12 animal proteins are from the CAZy database, 12 early eukaryotic proteins are from Table 1 and 29 early plant transcriptome-derived peptides are from S4 Table. Sequences were aligned using the MAFFT program. Phylogeny was inferred using RAxML program (see Methods). Complete species names are available in S3 Table and S4 Table.

doi:10.1371/journal.pone.0128409.g002

likely to use completely different sugar donors (UDP-xylose) and acceptors (oligosaccharide with β-1,4 xylan backbone).

GT43 proteins are sporadically distributed in non-plant/animal groups

The early eukaryotic GT43 homologs in clade D hit the GT43 domain model with E-value < 1e-10 and have conserved sequence motifs that are signatures of the GT43 family. Therefore we may conclude that GT43 is more broadly distributed than was once believed. On the other hand, interrogating ~400 completed fungal genomes and a dozen *Chlorophyta* green algal genomes did not reveal more GT43 proteins. Hence the phyletic distribution of GT43 proteins is at most sporadic in early eukaryotes, while they are almost universally present in all sequenced animals and plants (land plants and CGA). Though phylogenetically closer to animal proteins, the origin of these early (non-plant and non-animal) eukaryotic proteins remains a puzzle, potentially as a result of complex evolutionary histories involving horizontal gene transfer and gene loss.

CGA have three major clades of GT43 proteins and clade C evolved earliest

A recent study has found 16 assembled GT43 EST transcripts in seven out of 13 surveyed CGA transcriptomes [13]. Only two of the 16 transcripts were used in their phylogenetic analysis (others are too short to use). The analysis showed that one was clustered in clade A and the other was clustered in clade B. In our phylogeny (Fig 2), we included 19 CGA sequences, 17 of which are from seven CGA transcriptomes. The other two (kfl00579_0060, kfl00262_0200) are from the recently sequenced *Klebsormidium flaccidum* genome [26]. While we confirmed that CGA proteins fell into both clades A and B, ten of the 19 sequences form the newly defined clade C, ancestral to both A and B. The ten sequences are from six CGA species, suggesting that clade C is broadly distributed in CGA. Although the fully sequenced *K. flaccidum* genome contains only A and B proteins, three CGA species including *Cylindrocystis cushleckae*, *Cylindrocystis brebissonii* and *Mougeotia sp*. have representative genes in all the three clades (S2 Fig). Taking transcripts from *C. brebissonii* as an example, its clade C transcript C.brebissonii|Locus_3 (translated peptide 254 aa) shares 27% identity with AtIRX9 and 26% identity with AtIRX14; as a comparison, its clade A transcript C.brebissonii|Locus_12 (translated peptide 211 aa) shares 40% identity with AtIRX9, and its clade B transcript C.brebissonii|Locus_4 (translated peptide 185 aa) shares 33% identity with AtIRX14. This supports the phylogeny that clade C is ancestral to both clade A and clade B and represents the most ancient subfamily of plant GT43 genes.

Biochemical evidence has been shown that xylan is present in CGA [27] and it was proposed that CGA have already evolved the enzymatic components for xylan synthesis [13]. Since clade A and B that are involved in xylan synthesis are present in CGA, it remains a highly interesting question what functional role the ancestral clade C might have, especially for species having all three forms of GT43 genes. Future experiments are needed to characterize clade C genes in order to understand their function. The clustering of the two *Chlorophyta* GT43 in clade C suggests they might be also involved in xylan synthesis, given that xylans are also found in *Chlorophyta* [28].

Angiosperm proteins of clade A could be further classified into three subclades

It is clear from Fig 2 that land plants have two major GT43 clades (A and B), which is consistent with previous reports [12–14]. Clade A contains AtIRX9/AtIRX9L orthologs and clade B contains AtIRX14/AtIRX14L orthologs in all plant taxa spanning from CGA to angiosperms. AtIRX14L is very similar (70% amino acid identity) to AtIRX14; the two proteins must have evolved from a recent duplication event. However AtIRX9 and AtIRX9L are quite different in sequence (39% amino acid identity); therefore in some studies AtIRX9 and AtIRX9L were placed into two different clades [10] or subclades [14].

Similar to many other GT families, GT43 is expanded in plants by gene duplication and sequence divergence. Such expansion is more evident in clade A than B, which makes it necessary to further classify A into subclades. In Fig 2, A is further classified into five subclades. A1, A2 and A3 are angiosperm-specific. A1 contains AtIRX9L, A3 contains AtIRX9 and A2 contains no Arabidopsis proteins. A4 contains proteins from earlier emerging land plants including ferns, hornworts, liverworts and mosses, and A5 is CGA-specific. Although it seems unnecessary to have such a subclade classification given the small number of proteins in each subclade, the clustering appears to be more significant when more land plant proteins are included, as clade A becomes much larger and well separated than B (Fig 3). In addition A2 actually consists of proteins from both grasses and dicots when more species are included in the phylogeny (Fig 3).

The three land plant subclades have distinct sequence features and expression profiles

According to Figs 2 and 3, the separation of AtIRX14/AtIRX14L (clade B) and AtIRX9L/9 (clade A) predated the emergence of land plants 430~470 million years ago [29], as both clades contain CGA proteins; however, the divergence of AtIRX9L (subclade A1) and AtIRX9 (subclade A3) occurred much later, most likely after the emergence of angiosperms 167~199 million years ago [30], because all three subclades only contain angiosperm proteins.

Among the three subclades, A3 (AtIRX9) has the longest branches (Figs 2 and 3) signifying the most rapid sequence divergence. This is in agreement with findings made in more detailed sequence and expression analyses. For example, AtIRX9 has a smaller number of conserved substrate binding residues than AtIRX9L and AtIRX14 (S2 Table). A3 proteins also have more variable intron-exon gene structures (Fig 3) and more variable sequence lengths (S5 Table). In addition S5 Table shows that grasses have significantly more GT43 genes than other plants especially in subclade A3, consistent with the fact that the cell walls of grasses have higher xylan content [31]. Clade B proteins (AtIRX14/14L) are always the longest and in many cases A3 (AtIRX9) proteins are the shortest. As for GC content, in dicots subclade A3 seems to be the lowest while in grasses subclade A2 is the lowest. These observations tend to be true even when more flowering plants were examined.

In Arabidopsis and rice, A3 genes (yellow) also have the most extreme change in expression (Fig 4): with highest expression in stem tissues but lowest expression in other tissues. A1 and A2 (red), although phylogenetically close to A3, tend to be expressed at lower levels but almost evenly and widely expressed across all tissues. The more distantly related clade B (blue) has a similar profile as A3 but the differential expression change is less drastic than A3. These observations are also made in poplar [32] and switchgrass (Fig 4 and S3 Fig) with more tissues included. Altogether these data suggest rapid sequence and expression evolution in clade A particularly in subclade A3.

Materials and Methods

Sequence and microarray data sources

GT43 protein sequences of fully sequenced land plant genomes were downloaded from PlantCAZyme database [24], which we recently developed. Briefly, protein sequences in PlantCAZyme

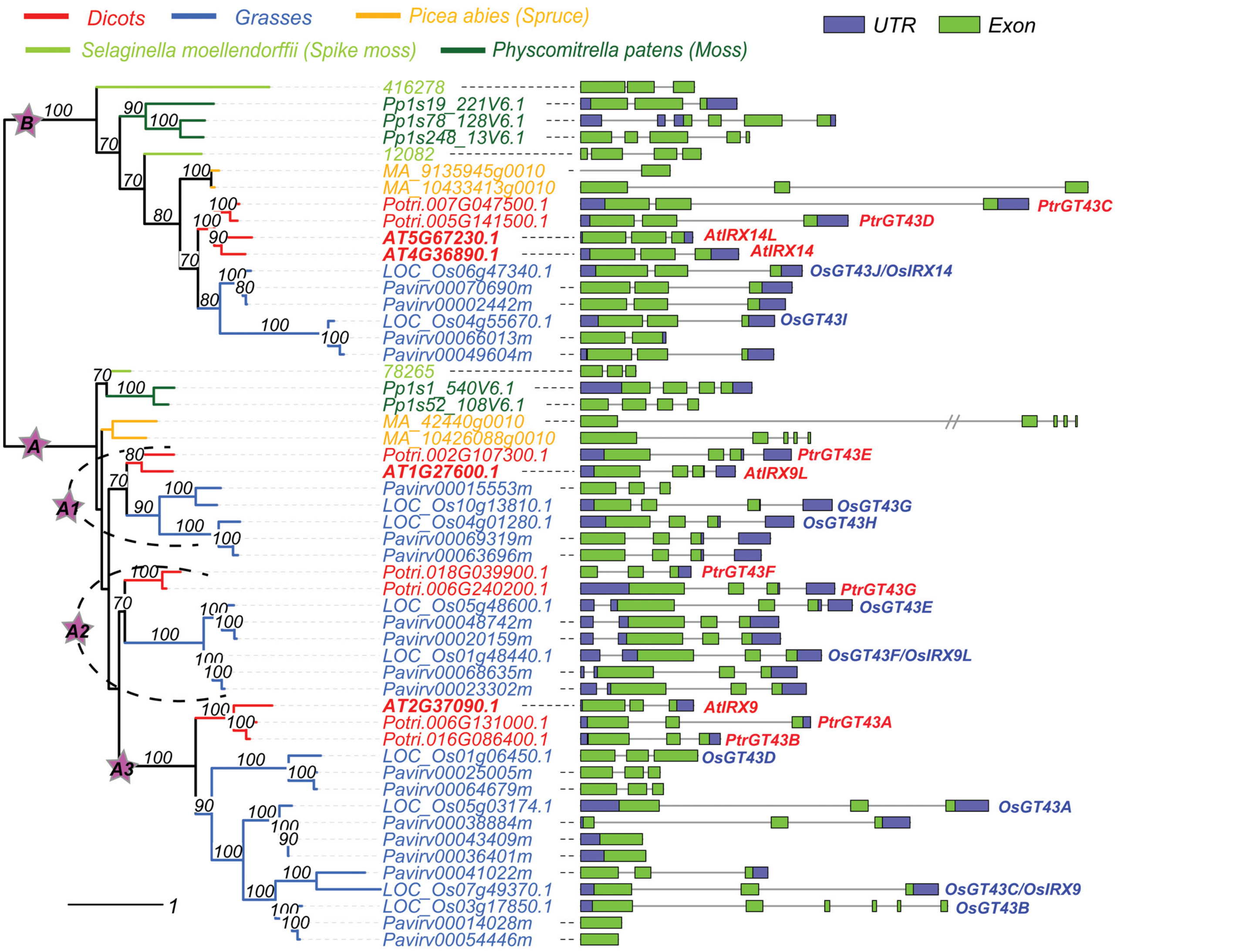

Fig 3. Intron-exon structures of plant GT43 genes. 52 proteins were used to build the phylogeny using RAxML program (see Methods). Gene structures were plotted on the right using the GSDS server [46]. All the protein IDs can be searched at Phytozome [34]. Poplar gene names (e.g. PtrGT43A) were adopted from [47] and rice gene names (e.g. OsGT43A) were adopted from [48].

doi:10.1371/journal.pone.0128409.g003

were collected by running HMMER 3.0 [33] using the GT43 signature domain hidden Markov model (HMM) as the query [18] to search against fully sequenced plant genomes from the Phytozome database [34].

For Arabidopsis, GEO microarray data sets GSE5629-GSE5634 from the GEO database were used (http://arabidopsis.info/). For poplar, GSE6422 was used [35]; for rice, GSE21494 was used [36]; and for switchgrass, microarray data from [37] was used.

Pipeline for data mining of RNA-Seq data

The computational pipeline shown in S1 Fig was used to retrieve homologous reads to GT43 proteins and to assemble them into longer transcripts for further motif and phylogenetic analyses. Details about the pipeline were explained in [23]. In this new pipeline we used Velvet [38] and Oases [39] to assemble the Illumina reads. VelvetOptimiser (http://www.vicbioinformatics.

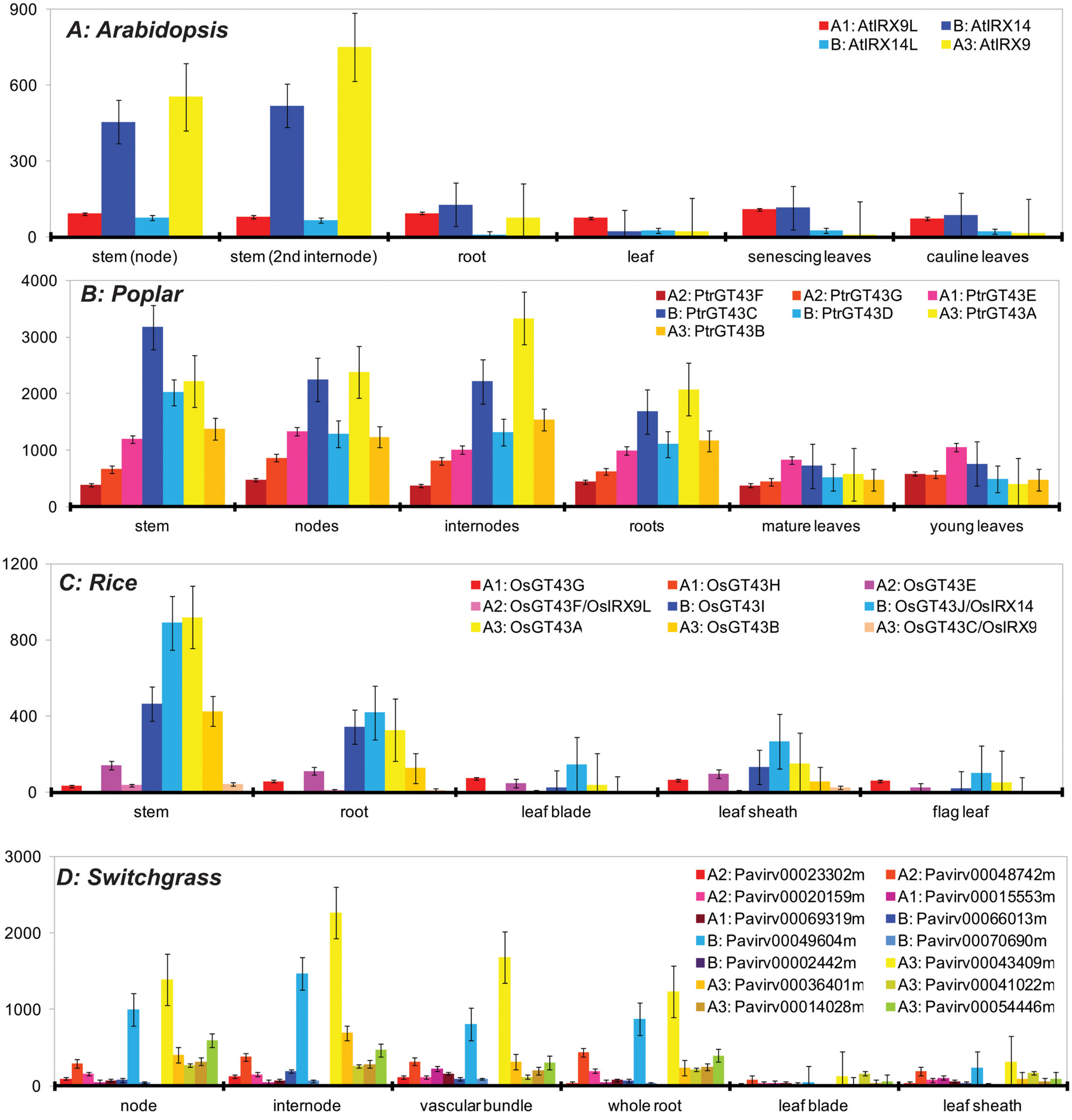

Fig 4. Expression profiles of GT43 genes. In Arabidopsis (A), poplar (B), rice (C) and switchgrass (D). Some genes do not have any probes in the microarray data and were excluded. Inset legends, A1, A2, A3 and B are the clade names followed by the gene names according to Fig 2 and Fig 3. For each tissue, genes are ordered and colored based on their clades: A in red, B in blue and C in yellow. The y-axis shows the expression values from microarray data. A complete version of this figure is available as S3 Fig.

doi:10.1371/journal.pone.0128409.g004

com/software.velvetoptimiser.shtml) was used to optimize the assembly parameters as suggested by [40]. Only assembled transcripts that led to peptide sequences longer than 100 amino acids and matching GT43 proteins of PlantCAZyme database (E-value < 1e-10) and containing GT43 domains (E-value < 1e-10) were kept for further analyses.

Phylogenetic analysis

All the multiple sequence alignments (MSAs) were generated using MAFFT v7.158b with the L-INS-i method [41], which is among the most accurate sequence alignment algorithms. Unless specifically indicated (e.g. S2 Fig), all MSAs were not manually edited, because manually editing sequences are rather subjective and impossible for others to reproduce. All phylogenies were built using RAxML 8.0 [42] with the following parameters: 100 times rapid bootstrap analysis and search for best-scoring maximum likelihood tree (-# 100-f a); JTT substitution model, GAMMA model of rate heterogeneity with estimate of proportion of invariable sites (-m PROTGAMMAIJTT). This parameter setting with the maximum likelihood algorithm implemented by RAxML is considered to be one of the most sophisticated and accurate protein phylogeny reconstruction methods.

Motif analysis

To scan GT43 homologous proteins and peptides, we have derived 8 PROSITE-style motif patterns (Table 1) according to two previous reports [11, 43]. We then used ScanProsite tool [44] to scan the 8 motif patterns in the 21 NCBI-nr protein sequences and 54 peptide sequences from plant transcriptomes. In addition, we also put another constraint on the relative location of the motifs: (i) motifs 1 and 2 must appear in the first half (N-terminal) of the full length sequence; (ii) motifs 7 and 8 must appear in the second half (C-terminal) of the full length sequence and (iii) the rest motifs must appear in the middle part of the sequence.

Supporting Information

S1 Fig. Computational pipeline for transcriptome data mining for GT43 proteins.
(EPS)

S2 Fig. Phylogeny based on edited sequence alignment of 99 GT43 proteins (21 from plant genomes + 12 from metazoa in CAZy + 12 from NCBI-nr early eukaryotes + 54 early plant transcriptomes) by removing long gaps and ambiguously alignment regions.
(EPS)

S3 Fig. Expression profiles of GT43 genes in Arabidopsis (A), poplar (B), rice (C) and switchgrass (D). This is a more complete version of Fig 4 with more tissues.
(EPS)

S1 Table. 21 NCBI-nr proteins that are non-plant and non-animal proteins and hit GT43 domain with E-value < 1e-10; proteins having at least 3 motifs are shown in yellow background.
(XLSX)

S2 Table. Key residues in human GlcAT-I protein reported in [11, 21] and the corresponding residues in other protein included in Fig 1.
(XLSX)

S3 Table. 25 Illumina RNA-Seq datasets that are used in data mining of GT43 genes.
(XLSX)

S4 Table. 54 assembled transcripts that hit their best PlantCAZyme match with E-value < 1e-10; Clade column is based on Fig 2; peptides having at least 3 motifs are shown in yellow background.
(XLSX)

S5 Table. GT43 genes in seven land plants.
(XLSX)

Acknowledgments

We acknowledge the Department of Computer Science of NIU for providing free access to the Linux computing cluster Gaea and the Yin lab members for helpful discussions.

Author Contributions

Conceived and designed the experiments: YY. Performed the experiments: RT. Analyzed the data: RT YY. Contributed reagents/materials/analysis tools: RT YY. Wrote the paper: YY RT.

References

1. Rennie EA, Scheller HV. Xylan biosynthesis. Current opinion in biotechnology. 2014; 26:100–7. doi: 10.1016/j.copbio.2013.11.013 PMID: 24679265.
2. Brown DM, Goubet F, Wong VW, Goodacre R, Stephens E, Dupree P, et al. Comparison of five xylan synthesis mutants reveals new insight into the mechanisms of xylan synthesis. The Plant journal: for cell and molecular biology. 2007; 52(6):1154–68. doi: 10.1111/j.1365-313X.2007.03307.x PMID: 17944810.
3. Wu AM, Hornblad E, Voxeur A, Gerber L, Rihouey C, Lerouge P, et al. 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 physiology. 2010; 153(2):542–54. doi: 10.1104/pp.110.154971 PMID: 20424005; PubMed Central PMCID: PMC2879767.
4. Lee C, O'Neill MA, Tsumuraya Y, Darvill AG, Ye ZH. The irregular xylem9 mutant is deficient in xylan xylosyltransferase activity. Plant & cell physiology. 2007; 48(11):1624–34. doi: 10.1093/pcp/pcm135 PMID: 17938130.
5. Brown DM, Zhang Z, Stephens E, Dupree P, Turner SR. Characterization of IRX10 and IRX10-like reveals an essential role in glucuronoxylan biosynthesis in Arabidopsis. The Plant journal: for cell and molecular biology. 2009; 57(4):732–46. doi: 10.1111/j.1365-313X.2008.03729.x PMID: 18980662.
6. Wu AM, Rihouey C, Seveno M, Hornblad E, Singh SK, Matsunaga T, et al. The Arabidopsis IRX10 and IRX10-LIKE glycosyltransferases are critical for glucuronoxylan biosynthesis during secondary cell wall formation. The Plant journal: for cell and molecular biology. 2009; 57(4):718–31. doi: 10.1111/j.1365-313X.2008.03724.x PMID: 18980649.
7. Zeng W, Jiang N, Nadella R, Killen TL, Nadella V, Faik A. A glucurono(arabino)xylan synthase complex from wheat contains members of the GT43, GT47, and GT75 families and functions cooperatively. Plant physiology. 2010; 154(1):78–97. doi: 10.1104/pp.110.159749 PMID: 20631319; PubMed Central PMCID: PMC2938142.
8. Pena MJ, Zhong R, Zhou GK, Richardson EA, O'Neill MA, Darvill AG, et al. Arabidopsis irregular xylem8 and irregular xylem9: implications for the complexity of glucuronoxylan biosynthesis. The Plant cell. 2007; 19(2):549–63. doi: 10.1105/tpc.106.049320 PMID: 17322407; PubMed Central PMCID: PMC1867335.
9. Lee C, Zhong R, Richardson EA, Himmelsbach DS, McPhail BT, Ye ZH. The PARVUS gene is expressed in cells undergoing secondary wall thickening and is essential for glucuronoxylan biosynthesis. Plant & cell physiology. 2007; 48(12):1659–72. doi: 10.1093/pcp/pcm155 PMID: 17991630.
10. Anders N, Dupree P. Glycosyltransferases of the GT43 Family. Annual Plant Reviews: Wiley-Blackwell; 2010. p. 251–63.
11. Fondeur-Gelinotte M, Lattard V, Oriol R, Mollicone R, Jacquinet JC, Mulliert G, et al. Phylogenetic and mutational analyses reveal key residues for UDP-glucuronic acid binding and activity of beta1,3-glucuronosyltransferase I (GlcAT-I). Protein science: a publication of the Protein Society. 2006; 15(7):1667–78. doi: 10.1110/ps.062089106 PMID: 16815917; PubMed Central PMCID: PMC2242556.

12. Harholt J, Sorensen I, Fangel J, Roberts A, Willats WG, Scheller HV, et al. The glycosyltransferase repertoire of the spikemoss Selaginella moellendorffii and a comparative study of its cell wall. PloS one. 2012; 7(5):e35846. doi: 10.1371/journal.pone.0035846 PMID: 22567114; PubMed Central PMCID: PMC3342304.

13. Mikkelsen MD, Harholt J, Ulvskov P, Johansen IE, Fangel JU, Doblin MS, et al. Evidence for land plant cell wall biosynthetic mechanisms in charophyte green algae. Annals of botany. 2014; 114(6):1217–36. doi: 10.1093/aob/mcu171 PMID: 25204387; PubMed Central PMCID: PMC4195564.

14. Kulkarni AR, Pena MJ, Avci U, Mazumder K, Urbanowicz BR, Pattathil S, et al. The ability of land plants to synthesize glucuronoxylans predates the evolution of tracheophytes. Glycobiology. 2012; 22(3):439–51. doi: 10.1093/glycob/cwr117 PMID: 22048859.

15. Popper Z, Michel G, Herve C, Domozych DS, Willats WG, Tuohy MG, et al. Evolution and Diversity of Plant Cell Walls: From Algae to Flowering Plants. Annu Rev Plant Biol. 2011; 62:567–90. Epub 2011/03/01. doi: 10.1146/annurev-arplant-042110-103809 PMID: 21351878.

16. Popper ZA. Evolution and diversity of green plant cell walls. Current opinion in plant biology. 2008; 11(3):286–92. Epub 2008/04/15. doi: S1369-5266(08)00040-X [pii] doi: 10.1016/j.pbi.2008.02.012 PMID: 18406657.

17. Fangel JU, Ulvskov P, Knox JP, Mikkelsen MD, Harholt J, Popper ZA, et al. Cell wall evolution and diversity. Frontiers in plant science. 2012; 3:152. PMID: 22783271. doi: 10.3389/fpls.2012.00152

18. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic acids research. 2012; 40(Web Server issue):W445–51. doi: 10.1093/nar/gks479 PMID: 22645317; PubMed Central PMCID: PMC3394287.

19. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic acids research. 2014; 42(Database issue):D490–5. doi: 10.1093/nar/gkt1178 PMID: 24270786; PubMed Central PMCID: PMC3965031.

20. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, et al. PROSITE: a documented database using patterns and profiles as motif descriptors. Briefings in bioinformatics. 2002; 3(3):265–74. PMID: 12230035.

21. Pedersen LC, Tsuchida K, Kitagawa H, Sugahara K, Darden TA, Negishi M. Heparan/chondroitin sulfate biosynthesis. Structure and mechanism of human glucuronyltransferase I. The Journal of biological chemistry. 2000; 275(44):34580–5. doi: 10.1074/jbc.M007399200 PMID: 10946001.

22. Jia S, Zhang X, Zhang G, Yin A, Zhang S, Li F, et al. Seasonally variable intestinal metagenomes of the red palm weevil (Rhynchophorus ferrugineus). Environmental microbiology. 2013. doi: 10.1111/1462-2920.12262 PMID: 24102776.

23. Yin Y, Johns MA, Cao H, Rupani M. A survey of plant and algal genomes and transcriptomes reveals new insights into the evolution and function of the cellulose synthase superfamily. BMC genomics. 2014; 15:260. doi: 10.1186/1471-2164-15-260 PMID: 24708035; PubMed Central PMCID: PMC4023592.

24. Ekstrom A, Taujale R, McGinn N, Yin Y. PlantCAZyme: a database for plant carbohydrate-active enzymes. Database: the journal of biological databases and curation. 2014; 2014:bau079. doi: 10.1093/database/bau079 PMID: 25125445; PubMed Central PMCID: PMC4132414.

25. Fondeur-Gelinotte M, Lattard V, Gulberti S, Oriol R, Mulliert G, Coughtrie MW, et al. Molecular basis for acceptor substrate specificity of the human beta1,3-glucuronosyltransferases GlcAT-I and GlcAT-P involved in glycosaminoglycan and HNK-1 carbohydrate epitope biosynthesis, respectively. Glycobiology. 2007; 17(8):857–67. doi: 10.1093/glycob/cwm055 PMID: 17567734.

26. Hori K, Maruyama F, Fujisawa T, Togashi T, Yamamoto N, Seo M, et al. Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. Nature communications. 2014; 5:3978. doi: 10.1038/ncomms4978 PMID: 24865297; PubMed Central PMCID: PMC4052687.

27. Sorensen I, Pettolino FA, Bacic A, Ralph J, Lu FC, O'Neill MA, et al. The charophycean green algae provide insights into the early origins of plant cell walls. Plant Journal. 2011; 68(2):201–11. doi: 10.1111/J.1365-313x.2011.04686.X WOS:000295836500001. PMID: 21707800

28. Popper ZA, Tuohy MG. Beyond the green: understanding the evolutionary puzzle of plant and algal cell walls. Plant physiology. 2010; 153(2):373–83. doi: 10.1104/pp.110.158055 PMID: 20421458; PubMed Central PMCID: PMC2879814.

29. Wodniok S, Brinkmann H, Glockner G, Heidel AJ, Philippe H, Melkonian M, et al. Origin of land plants: Do conjugating green algae hold the key? Bmc Evol Biol. 2011;11. doi: Artn 104 doi: 10.1186/1471-2148-11-104 WOS:000290303000001. PMID: 21226948

30. Bell CD, Soltis DE, Soltis PS. The age and diversification of the angiosperms re-revisited. American journal of botany. 2010; 97(8):1296–303. doi: 10.3732/ajb.0900346 PMID: 21616882.

31. Scheller HV, Ulvskov P. Hemicelluloses. Annual review of plant biology. 2010; 61:263–89. doi: 10.1146/annurev-arplant-042809-112315 PMID: 20192742.

32. Ratke C, Pawar PM, Balasubramanian VK, Naumann M, Duncranz ML, Derba-Maceluch M, et al. Populus GT43 family members group into distinct sets required for primary and secondary wall xylan biosynthesis and include useful promoters for wood modification. Plant biotechnology journal. 2014. doi: 10.1111/pbi.12232 PMID: 25100045.

33. Eddy SR. Accelerated Profile HMM Searches. PLoS computational biology. 2011; 7(10):e1002195. doi: 10.1371/journal.pcbi.1002195 PMID: 22039361; PubMed Central PMCID: PMC3197634.

34. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic acids research. 2012; 40(Database issue):D1178–86. doi: 10.1093/nar/gkr944 PMID: 22110026; PubMed Central PMCID: PMC3245001.

35. Yang X, Kalluri UC, Jawdy S, Gunter LE, Yin T, Tschaplinski TJ, et al. The F-box gene family is expanded in herbaceous annual plants relative to woody perennial plants. Plant physiology. 2008; 148(3):1189–200. doi: 10.1104/pp.108.121921 PMID: 18775973; PubMed Central PMCID: PMC2577272.

36. Sato Y, Antonio B, Namiki N, Motoyama R, Sugimoto K, Takehisa H, et al. Field transcriptome revealed critical developmental and physiological transitions involved in the expression of growth potential in japonica rice. BMC plant biology. 2011; 11:10. doi: 10.1186/1471-2229-11-10 PMID: 21226959; PubMed Central PMCID: PMC3031230.

37. Zhang JY, Lee YC, Torres-Jerez I, Wang M, Yin Y, Chou WC, et al. Development of an integrated transcript sequence database and a gene expression atlas for gene discovery and analysis in switchgrass (Panicum virgatum L.). The Plant journal: for cell and molecular biology. 2013; 74(1):160–73. doi: 10.1111/tpj.12104 PMID: 23289674.

38. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research. 2008; 18(5):821–9. doi: 10.1101/gr.074492.107 PMID: 18349386; PubMed Central PMCID: PMC2336801.

39. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012; 28(8):1086–92. doi: 10.1093/bioinformatics/bts094 PMID: 22368243; PubMed Central PMCID: PMC3324515.

40. Zerbino DR. Using the Velvet de novo assembler for short-read sequencing technologies. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]. 2010; Chapter 11:Unit 11 5. doi: 10.1002/0471250953.bi1105s31 PMID: 20836074; PubMed Central PMCID: PMC2952100.

41. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013; 30(4):772–80. doi: 10.1093/molbev/mst010 PMID: 23329690; PubMed Central PMCID: PMC3603318.

42. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30(9):1312–3. doi: 10.1093/bioinformatics/btu033 PMID: 24451623; PubMed Central PMCID: PMC3998144.

43. Ren Y, Hansen SF, Ebert B, Lau J, Scheller HV. 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. 2014; 9(8):e105014. doi: 10.1371/journal.pone.0105014 PMID: 25118690; PubMed Central PMCID: PMC4132061.

44. de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic acids research. 2006; 34(Web Server issue):W362–5. doi: 10.1093/nar/gkl124 PMID: 16845026; PubMed Central PMCID: PMC1538847.

45. Gouet P, Robert X, Courcelle E. ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins. Nucleic acids research. 2003; 31(13):3320–3. PMID: 12824317; PubMed Central PMCID: PMC168963.

46. Guo AY, Zhu QH, Chen X, Luo JC. [GSDS: a gene structure display server]. Yi chuan = Hereditas / Zhongguo yi chuan xue hui bian ji. 2007; 29(8):1023–6. PMID: 17681935.

47. Lee C, Teng Q, Zhong R, Ye ZH. Molecular dissection of xylan biosynthesis during wood formation in poplar. Molecular plant. 2011; 4(4):730–47. doi: 10.1093/mp/ssr035 PMID: 21596688.

48. Lee C, Teng Q, Zhong R, Yuan Y, Ye ZH. Functional roles of rice glycosyltransferase family GT43 in xylan biosynthesis. Plant signaling & behavior. 2014; 9:e27809. PMID: 24525904; PubMed Central PMCID: PMC4091335.