Plant Science xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Plant Science

journal homepage: www.elsevier.com/locate/plantsci

Comparative analysis of GT14/GT14-like gene family in *Arabidopsis*, *Oryza*, *Populus*, *Sorghum* and *Vitis*

Chu-Yu Ye, Ting Li, Gerald A. Tuskan, Timothy J. Tschaplinski, Xiaohan Yang*

Biosciences Division and BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

ARTICLE INFO

Article history:
Received 28 August 2010
Received in revised form 26 January 2011
Accepted 27 January 2011
Available online xxx

Keywords:
Glycosyltransferase
GT14 family
Branch domain
DUF266
Cell wall
Poplar

ABSTRACT

Glycosyltransferase family14 (GT14) belongs to the glycosyltransferase (GT) superfamily that plays important roles in the biosynthesis of cell walls, the most abundant source of cellulosic biomass for bioethanol production. It has been hypothesized that DUF266 proteins are a new class of GTs related to GT14. In this study, we identified 62 GT14 and 106 DUF266 genes (named GT14-like herein) in *Arabidopsis*, *Oryza*, *Populus*, *Sorghum* and *Vitis*. Our phylogenetic analysis separated GT14 and GT14-like genes into two distinct clades, which were further divided into eight and five groups, respectively. Similarities in protein domain, 3D structure and gene expression were uncovered between the two phylogenetic clades, supporting the hypothesis that GT14 and GT14-like genes belong to one family. Therefore, we proposed a new family name, GT14/GT14-like family that combines both subfamilies. Variation in gene expression and protein subcellular localization within the GT14-like subfamily were greater than those within the GT14 subfamily. One-half of the *Arabidopsis* and *Populus* GT14/GT14-like genes were found to be preferentially expressed in stem/xylem, indicating that they are likely involved in cell wall biosynthesis. This study provided new insights into the evolution and functional diversification of the GT14/GT14-like family genes.

1. Introduction

Plant cell walls are the major source of renewable biomass for alternative biofuels (e.g., bioethanol) production [1]. One of the key steps of cell wall biosynthesis involves glycosyltransferases (GTs), which catalyze the transfer of sugar moieties from donor molecules to specific acceptors, creating glycosidic bonds [2]. GTs are found in most living organisms, but are particularly abundant in plants [3]. It was reported that the DNA sequences encoding GTs occupied ~1.6% of the genomic sequence of *Arabidopsis thaliana*, much higher compared to the ratio of ~0.7% in humans [4]. GTs have been classified into more than 90 families in the CAZy database [5]. Previous studies revealed important roles of GTs in biological processes toward cell wall formation. For example, cellulose synthases (GT2 family) are involved in cellulose synthesis [6]. *A. thaliana* fucosyltransferase 1 (GT37 family) is involved in xyloglucan biosynthesis [7]. Galacturonosyltransferases (GT8 family) are involved in pectin biosynthesis [8]. Fragile fiber 8 (GT47 family) may function as a xylan biosynthetic enzyme [9]. GT14 enzymes identified in animals catalyze the transfer of β-(1-6)-linked N-acetylglucosaminyl and β-linked xylosyl residues to proteins [10]. To our knowledge, no GT14 genes have been functionally characterized in plants. Two GT14 genes were identified to be preferentially expressed in the xylem tissue of *Populus tremula* × *tremuloides* [10], indicating that GT14 genes may play an important role in secondary cell wall biosynthesis.

Recently, a new class of proteins, Domain of Unknown Function 266 (DUF266), was identified in *Oryza sativa* [11]. It was reported that the *OsBC10* gene in *O. sativa* encoded a DUF266 domain-containing protein with GT activity [11]. Bioinformatic studies suggested that DUF266 genes were distantly related to GT14 genes [12]. DUF266 and GT14 genes were merged into one category based on one shared protein domain, i.e., the Branch domain (PF02485) in the Pfam database [13]. Therefore, it is hypothesized that GT14 and DUF266 (herein named GT14-like) proteins belong to one gene family, which we tentatively named as the GT14/GT14-like family. To test this hypothesis, this study investigated the phylogenetic relationship of GT14 and GT14-like genes in five representative plant species. We also studied protein domains, three dimensional (3D) structure, subcellular localization and gene expression patterns. Our results provide new insights into the evolution and functional diversification of GT14 and GT14-like genes.

* Corresponding author at: Biosciences Division, Oak Ridge National Laboratory, P.O. Box 2008 MS 6422, Oak Ridge, TN 37831-6422, USA. Tel.: +1 865 241 6895; fax: +1 865 576 9939.
E-mail address: yangx@ornl.gov (X. Yang).

0168-9452/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.plantsci.2011.01.021

2. Materials and methods

2.1. Gene identification

The annotated protein sequences of five plant genomes were obtained from http://www.arabidopsis.org/ (*A. thaliana*; version 9), http://www.phytozome.net/ (*Populus trichocarpa*; version 2.0), http://rice.plantbiology.msu.edu/ (*O. sativa*; version 6.1), http://www.phytozome.net/ (*Sorghum bicolor*; version 1.4), and http://www.genoscope.cns.fr/ (*Vitis vinifera*; version 1.0). To test the hypothesis that GT14 and DUF266 (or GT14-like) proteins belong to one family with the signature protein domain Branch, the HMM profile for Branch domain (PF02485) was obtained from Pfam (http://pfam.janelia.org/). HMMER [14] was used to search the customized database containing the protein sequences of the five plant genomes for proteins containing the Branch domain with the threshold set at 1/100 of the Pfam GA gathering cutoff [15], resulting in 184 HMMER-selected proteins, which were then scanned for the presence of Branch domain by InterProScan [16]. Two sequences without the Branch domain were excluded from further analysis. The remaining 182 protein sequences were divided into three categories: category I (150 proteins) in which each protein corresponds to a unique gene locus with full-length transcripts (containing both start and stop codons); category II (10 proteins) in which each protein corresponds to a unique gene locus with truncated transcripts (missing start or stop codons); and category III (22 proteins) in which multiple protein sequences corresponds to one gene locus with full-length transcripts. For proteins in category II, the JGI *Populus* and *Sorghum* genome browsers (http://genome.jgi-psf.org/Poptr1_1/, http://genome.jgi-psf.org/Sorbi1/) were used to search for better alternative gene models that have full-length transcripts with higher protein sequence homology to *Arabidopsis* proteins. As such, seven incomplete protein sequences were changed to full-length alternatives. For the proteins in category III, only one (i.e., the longest) protein sequence was retained for each of the 11 gene loci. Altogether, 168 non-redundant full-length protein sequences (150 category I + 7 category II + 11 category III) were selected for further analysis (Supplemental Table 1).

All of the *Arabidopsis* glycosyltransferase (GT) protein sequences were obtained from ftp://ftp.arabidopsis.org/home/tair/Genes/Gene_families/. The protein sequences (RefSeq) of *Homo sapiens* were obtained from NCBI (http://www.ncbi.nlm.nih.gov/).

2.2. Phylogenetic tree construction

The protein sequences were aligned using MAFFT [17]. The phylogenetic tree was constructed using PhyML [18] with aLRT SH-like branch support and protein evolution model JTT, which was selected by ModelGenerator [19]. The gene tree was reconciled with a species tree ([[*Vitis*, [*Arabidopsis*, *Populus*]], [*Oryza*, *Sorghum*]]) using Notung [20] to estimate upper and lower bounds of the time of duplication. The phylogenetic tree of all *Arabidopsis* GT sequences were constructed using http://mafft.cbrc.jp/alignment/server/clustering.html with accurate distance measure and the UPMGA clustering method. The trees were displayed using MEGA version 4.1 [21] with 50% threshold of branch value.

2.3. Gene expression analysis

The *Arabidopsis* and *Populus* microarray gene expression data were obtained from AtGenExpress [22] and poplar eFP [23], respectively. Paralog and ortholog pairs were identified according to the phylogenetic trees. Heatmaps of gene expression were generated using R (http://www.r-project.org/). Clustering of gene expression pattern was performed with SOTA in GEPAS [24]. In addition, the gene expression pattern of *Populus* wood series was obtained from PopGenIE [25–27].

Table 1
The number of GT14/GT14-like genes in *Arabidopsis*, *Oryza*, *Populus*, *Sorghum* and *Vitis*.

Subfamily	*Arabidopsis*	*Populus*	*Oryza*	*Sorghum*	*Vitis*	Total
GT14	11	17	12	12	10	62
GT14-like	22	27	19	22	16	106
Total	33	44	31	34	26	168

2.4. Prediction of three-dimensional (3D) protein structure and subcellular localization

3D protein structures were predicted by I-TASSER [28]. The structure alignment was performed with TM-align [29]. The comparison of the 3D protein structures was assessed using ANOVA with Waller–Duncan mean separation method. The protein subcellular localization was predicted by YLoc [30].

2.5. Identification of protein motifs

Protein sequences were scanned for domains using BlastProDom, FPrintScan, HMMPIR, HMMPfam, HMMsmart, HMMTigr, ProfileScan, Scan-RegExp, and SuperFamily implemented in InterProScan [16].

3. Results

3.1. Identification of GT14 and DUF266 (or GT14-like) proteins

To test the hypothesis that GT14 and DUF266 proteins belong to one family with the signature protein domain Branch, a HMMER-InterProScan approach [15] was used to obtain Branch domain-containing protein sequences in five sequenced plant species, including *A. thaliana* (annual eudicot), *O. sativa* (monocot), *P. trichocarpa* (perennial eudicot), *S. bicolor* (monocot), and *V. vinifera* (perennial eudicot). A total of 168 non-redundant full-length Branch domain-containing protein sequences were identified, with 33 in *A. thaliana*, 31 in *O. sativa*, 44 in *P. trichocarpa*, 34 in *S. bicolor*, and 26 in *V. vinifera* (Table 1; Supplemental Table 1). The length of the protein sequences ranged from 75 to 651 amino acids (aa), with an average of 376 aa. The protein set generated by the HMMER-InterProScan approach contains all of the 11 *A. thaliana* GT14 proteins in CAZy database [5] and the 12 *O. sativa* GT14 proteins in the rice GT database [1], indicating that our method for gene selection is valid.

3.2. Phylogenetic relationship among the GT14 and GT14-like genes

A phylogenetic tree was created using full-length Branch domain-containing protein sequences in the five plant species (Fig. 1A). According to the phylogenetic tree, the 168 Branch domain-containing proteins were separated into two phylogenetic clades, designated as GT14 and GT14-like. The GT14-like proteins, previously known as DUF266 proteins, were manually classified into 5 groups (A1–A5). The GT14 proteins were divided into 8 groups (B1–B8). We used InterProScan [16] to investigate the 168 GT14 and GT14-like protein sequences and identified three protein domains: Core-2/1-Branching enzymatic function domain (Branch; IPR021141), glycosyltranferase 14 (GT14) domain (IPR003406), and a calcium-binding site domain (IPR018247). These three domains

Please cite this article in press as: C.-Y. Ye, et al., Comparative analysis of GT14/GT14-like gene family in *Arabidopsis*, *Oryza*, *Populus*, *Sorghum* and *Vitis*, Plant Sci. (2011), doi:10.1016/j.plantsci.2011.01.021

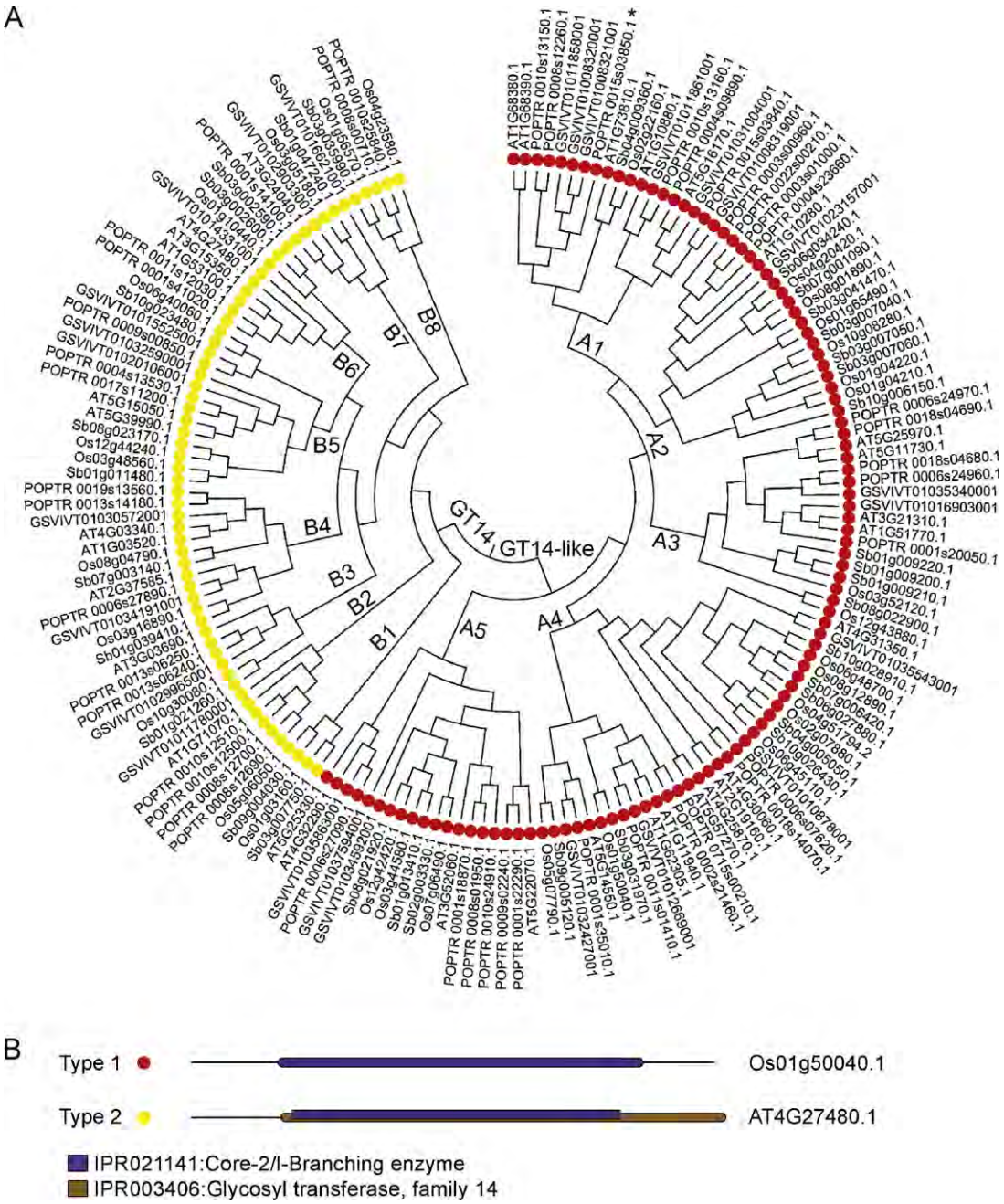

Fig. 1. Phylogenetic relationship (A) and protein domain structure (B) among the GT14 and GT14-like genes in *Arabidopsis*, *Oryza*, *Populus*, *Sorghum* and *Vitis*. The phylogenetic tree was constructed using protein sequences. **Note*: There is an additional calcium-binding site (IPR018247) in one protein sequence (POPTR_0015s03850.1) in clade GT14-like.

organized into two types of domain structures: Type 1 having Branch domain only and Type 2 having both Branch domain and GT14 domain (Fig. 1B). The Type 2 domain structure is exclusively contained in the phylogenetic clade GT14, and the Type 1 domain structure is exclusively contained in the phylogenetic clade GT14-like (Fig. 1A). The consistency between the domain structure classification and the phylogenetic classification validated our phylogenetic analysis. We also constructed a phylogenetic tree using the protein-coding sequences of the 168 Branch domain genes (Supplemental Fig. 1). In this DNA sequence-based tree, the clade GT14 is still distinguishable from the clade GT14-like, consistent with the phylogenetic tree (Fig. 1) based on protein sequences.

3.3. Evolution of the GT14 and GT14-like genes

To study the evolutionary dynamics of the GT14 and GT14-like genes, a reconciled phylogenetic tree (Supplemental Fig. 2) was constructed by combining the gene tree (Fig. 1A) and the species tree ([[*Vitis*, [*Arabidopsis*, *Populus*]], [*Oryza*, *Sorghum*]]). It was suggested that there were two rounds of genome duplication in *Arabidopsis* (α and β) and *Populus* ('salicoid' and 'eurosid' duplication), whereas there was no genome-wide duplication detected in *Vitis* after a shared ancient γ triplication [31–33]. The phylogenetic relationship among the genes in Group B6 (Supplemental Fig. 2, Fig. 1A) was consistent with the genome duplication history in *Arabidopsis*, *Populus* and *Vitis*. In the phylogenetic tree (Fig. 1A), the

Please cite this article in press as: C.-Y. Ye, et al., Comparative analysis of GT14/GT14-like gene family in *Arabidopsis*, *Oryza*, *Populus*, *Sorghum* and *Vitis*, Plant Sci. (2011), doi:10.1016/j.plantsci.2011.01.021

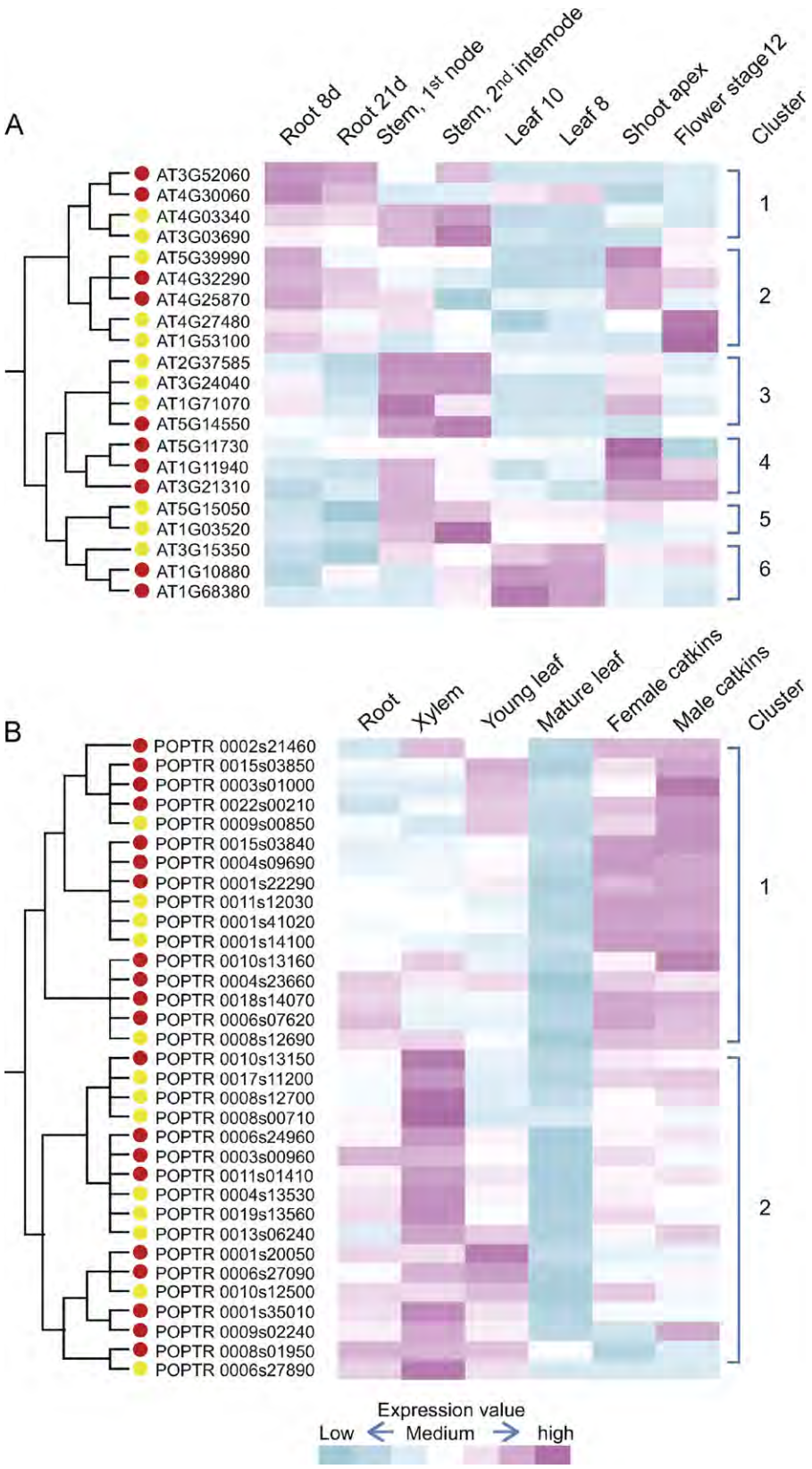

Fig. 2. Clustering of GT14 (marked by yellow-filled circle) and GT14-like (marked by red-filled circle) gene in *Arabidopsis* (A) and *Populus* (B).

number of genes from dicot plants is ~1.5 times that from monocot plants in both clade GT14 (38 dicot vs. 24 monocot) and clade GT14-like (65 dicot vs. 41 monocot). However, this dicot/monocot gene number ratio was distorted in several phylogenetic groups, such as Group A1 (17 dicot vs. 2 monocot), Group B2 (only dicot), Group A2 (6 dicot vs. 13 monocot), and B1 (only monocot), indicating that both GT14 and GT14-like subfamilies experienced lineage-specific expansion in dicots (Groups A1 and B2) and monocots

Please cite this article in press as: C.-Y. Ye, et al., Comparative analysis of GT14/GT14-like gene family in *Arabidopsis*, *Oryza*, *Populus*, *Sorghum* and *Vitis*, Plant Sci. (2011), doi:10.1016/j.plantsci.2011.01.021

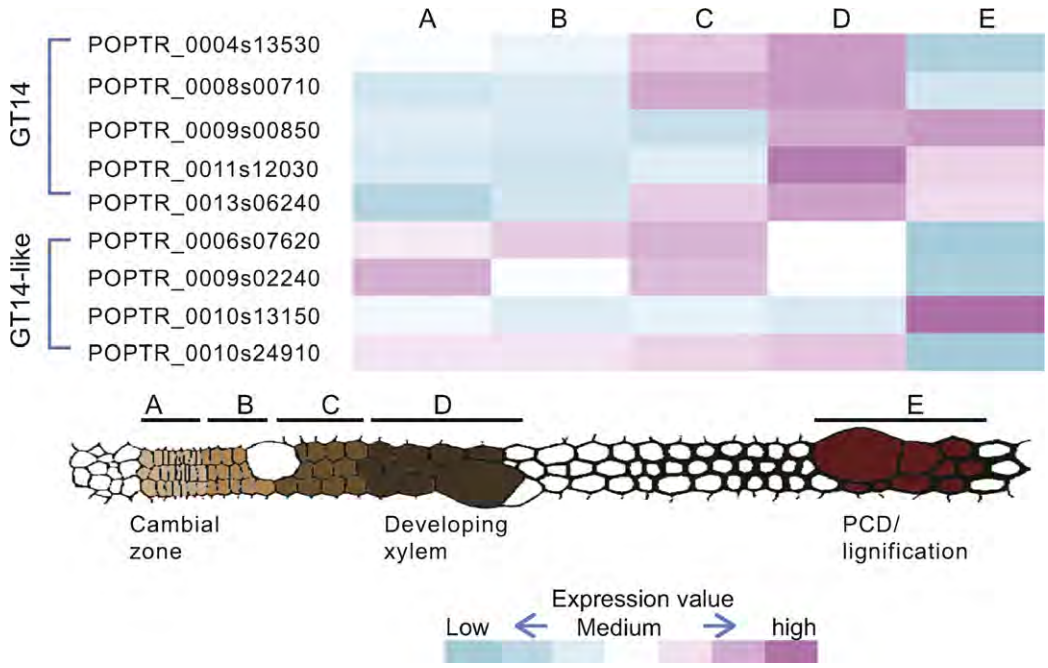

Fig. 3. *Populus* gene expression during wood formation. A, B, C, D and E zones indicate the positions of samples adopted from PopGenIE [27].

(Groups A2 and B1). On the other hand, Group B8 contains *Oryza* and *Sorghum* genes in monocots, and only *Populus* genes in dicots, possibly indicating lineage-specific gene loss in the other two dicot plant genomes (*Arabidosis* and *Vitis*).

3.4. Gene expression pattern of the GT14 and GT14-like genes

To investigate the functional diversity of the GT14 and GT14-like genes in *Arabidopsis* and *Populus*, we studied the tissue-specific expression patterns using the public microarray data. Based on the similarity of expression pattern, the GT14 and GT14-like genes in *Arabidopsis* were divided into six co-expression clusters (Fig. 2A). For example, Cluster 1 showed preferential expression in root and stem. Cluster 2 showed low level of gene expression in leaf (Fig. 2A). The *Populus* GT14 and GT14-like genes were divided into two co-expression clusters: one showing preferential expression in flower (i.e., the female and male catkins) and the other showing preferential expression in xylem (Fig. 2B). Most of the expression clusters included both GT14 and GT14-like genes (Figs. 2A, B). This indicates that the GT14 and GT14-like genes are involved in the biological processes associated with root, stem/xylem, and flower development. Wood formation in *Populus* can be delineated into five developing zones (A–E) from the outer to the inner side along the cross-section of the stem [25,26]. Limited data from the *Populus* microarray database PopGenIE [25–27] revealed subtle differences in the expression pattern during wood formation between the two subfamilies, with the GT14 subfamily preferentially expressed in zones D–E and GT14-like subfamily preferentially expressed in zones A–C (Fig. 3).

To understand the functional diversification among the phylogenetic lineages of GT14 and GT14-like subfamilies, we examined tissue-specific expression pattern of the *Arabidopsis* and *Populus* genes within the phylogenetic context. Some duplicated gene pairs in *Arabidopsis* displayed alternate expression profiles. For example, AT1G53100 was preferentially expressed in flower and root, whereas its paralog (AT3G15350) was preferentially expressed in leaf; AT5G39990 was preferentially expressed in root and shoot apex whereas its paralog AT5G15050 was preferentially expressed in stem (Fig. 4A). Compared with *Arabidopsis* genes (Fig. 4A), the *Populus* GT14 and GT14-like genes showed less variation in the tissue-specific expression patterns, with most of these genes expressed preferentially in xylem and catkins (Fig. 4B). One example of differential expression between duplicated genes in *Populus* was observed for POPTR_0008s12690, which was preferentially expressed in root, xylem and catkins, whereas its paralog (POPTR_0008s12700) was preferentially expressed in xylem and male catkins (Fig. 4B). These differential expression data suggest subfunctionalization of duplicated genes [34–36].

We also compared the expression of the *Arabidopsis*–*Populus* orthologous gene pairs in GT14 and GT14-like subfamilies. The GT14-like genes showed less conserved tissue-specific expression pattern between *Arabidopsis* and *Populus* than did the GT14 genes (Fig. 5).

3.5. 3D protein structure and subcellular localization

The 3D protein structures of representative *Arabidopsis* genes selected from each phylogenetic group in the GT14/GT14-like phylogeny (Fig. 1A) was predicted by I-TASSER [28], with the exception

Table 2
The TM-scores (reflecting similarity of topologies of two protein structures) of pairwise alignment between 3D structures of representative proteins in the phylogenetic groups (A1–5 and B2–7) defined in Fig. 1A.

	A2	A3	A4	A5	B2	B3	B4	B5	B6	B7
A1[a]	0.67	0.76	0.79	0.83	0.68	0.73	0.73	0.68	0.71	0.69
A2		0.72	0.72	0.81	0.66	0.70	0.69	0.64	0.66	0.67
A3			0.76	0.84	0.69	0.73	0.73	0.68	0.69	0.69
A4				0.84	0.72	0.76	0.79	0.71	0.73	0.74
A5					0.73	0.77	0.77	0.69	0.72	0.72
B2						0.88	0.84	0.75	0.78	0.80
B3							0.88	0.75	0.80	0.82
B4								0.77	0.81	0.83
B5									0.84	0.84
B6										0.86

[a] *Note*: A1 is represented by AT1G68380.1; A2 by AT1G10280.1; A3 by AT5G25970.1; A4 by AT5G14550.1; A5 by AT3G52060.1; B2 by AT1G71070.1; B3 by AT3G03690.1; B4 by AT2G37585.1; B5 by AT5G15050.1; B6 by AT4G27480.1; B7 by AT3G24040.1.

Please cite this article in press as: C.-Y. Ye, et al., Comparative analysis of GT14/GT14-like gene family in *Arabidopsis*, *Oryza*, *Populus*, *Sorghum* and *Vitis*, Plant Sci. (2011), doi:10.1016/j.plantsci.2011.01.021

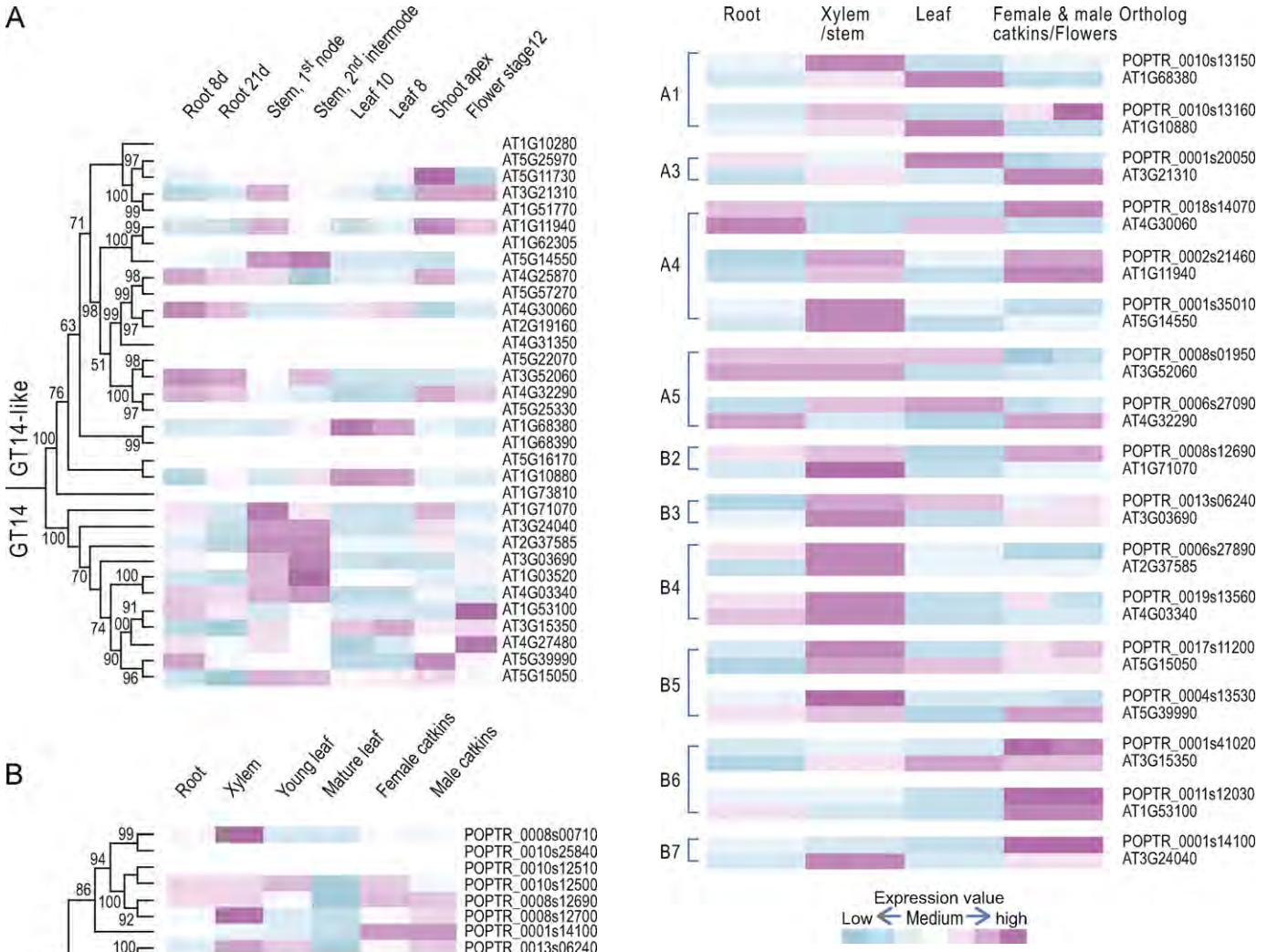

Fig. 5. Comparative analysis of gene expression pattern of orthologs between *Arabidopsis* and *Populus*. A1–A5 and B2–B7 are phylogenetic groups defined in Fig. 1A. The names of *Populus* gene and *Arabidopsis* gene are started with "POPTR" and "AT", respectively.

that Goups B1 and B8 were excluded due to the absence of *Arabidopsis* genes. The pairwise alignments between the 3D protein structures were performed using the TM-align program [29]. All of the pairwise TM-scores (reflecting similarity of topologies of two protein structures) between the representative proteins were greater than 0.65 (Table 2), indicating that these proteins are structurally similar with each other because TM-score of >0.5 means the same folding shared by two structures [37]. Still, the average TM score (0.71) between phylogenetic clades GT14 and GT14-like were significantly ($P < 0.0001$) lower than that between the proteins within clade GT14 (0.77) or GT14-like (0.82).

Subcellular localization of all the 33 *Arabidopsis* GT14/GT14-like proteins was predicted using YLoc [30]. Almost all the GT14 proteins in *Arabidopsis* were predicted to be located at the Golgi apparatus (Table 3), which is consistent with the previous studies

Fig. 4. Expression pattern of GT14 and GT14-like genes in *Arabidopsis* (A) and *Populus* (B). The *Arabidopsis* gene expression data were obtained from AtGenExpress (http://www.weigelworld.org/). The *Populus* gene expression data were obtained from poplar eFP (http://www.bar.utoronto.ca). Phylogenetic trees were showed on left. Blank rows indicate that gene expression data are not available in AtGenExpress and poplar eFP.

Please cite this article in press as: C.-Y. Ye, et al., Comparative analysis of GT14/GT14-like gene family in *Arabidopsis*, *Oryza*, *Populus*, *Sorghum* and *Vitis*, Plant Sci. (2011), doi:10.1016/j.plantsci.2011.01.021

Table 3
Predicted subcellular localization of the GT14/GT14-like proteins in *Arabidopsis*.

Subfamily	Gene name	Predicted location
GT14-like	AT1G10280.1	Golgi apparatus
	AT1G10880.1	Cytoplasm and nucleus
	AT1G11940.1	Golgi apparatus
	AT1G51770.1	Plasma membrane and extracelluar space
	AT1G62305.1	Golgi apparatus
	AT1G68380.1	Plasma membrane
	AT1G68390.1	Peroxisome and plasma membrane
	AT1G73810.1	Cytoplasm and plasma membrane
	AT2G19160.1	Golgi apparatus
	AT3G21310.1	Plasma membrane, perxisome, cytoplasm, and extracellular space
	AT3G52060.1	Golgi apparatus
	AT4G25870.1	Peroxisome, cytoplasm, and nucleus
	AT4G30060.1	Golgi apparatus
	AT4G31350.1	Golgi apparatus and plasma membrane
	AT4G32290.1	Plasma membrane
	AT5G11730.1	Plasma membrane and Golgi apparatus
	AT5G14550.1	Golgi apparatus
	AT5G16170.1	Peroxisome
	AT5G22070.1	Plasma membrane and Golgi apparatus
	AT5G25330.1	Plasma membrane and Golgi apparatus
	AT5G25970.1	Plasma membrane and extracelluar space
	AT5G57270.1	Golgi apparatus
GT14	AT1G03520.1	Golgi apparatus
	AT1G53100.1	Cytoplasm
	AT1G71070.1	Golgi apparatus
	AT2G37585.1	Golgi apparatus
	AT3G03690.1	Golgi apparatus
	AT3G15350.1	Golgi apparatus
	AT3G24040.1	Cytoplasm and nucleus
	AT4G03340.1	Golgi apparatus
	AT4G27480.1	Golgi apparatus
	AT5G15050.1	Golgi apparatus, cytoplasm, peroxisome, and mitochondrion
	AT5G39990.1	Golgi apparatus

showing that GTs are usually localized at the Golgi apparatus [38]. In contrast, there was more variation in the localization of GT14-like family proteins, with the majority of the localizations occurring at the Golgi apparatus and plasma membrane (Table 3).

4. Discussion

4.1. DUF266 and GT14 proteins belong to a single family

In this study, we tested the hypothesis that DUF266 and GT14 proteins belong to a single family. We obtained several lines of evidence supporting this hypothesis. Firstly, DUF266 and GT14 proteins share one signature Branch domain (Fig. 1B). Secondly, DUF266 and GT14 genes have a similar tissue-specific expression profile (Fig. 2). Finally, the DUF266 and GT14 proteins have a similar 3D folding pattern, as revealed by the alignment of 3D protein structure (Table 2). Previous studies also suggest the functional similarity and close relationship between DUF266 and GT14 proteins [11,12]. Clustering of all the *Arabidopsis* GT sequences also showed that DUF266 proteins are more closely related to the GT14 family than to other GT families (Supplemtenal Fig. 3). Therefore, we suggest that DUF266 and GT14 proteins be merged into one family, named GT14/GT14-like family, with the original GT14 proteins defined as a GT14 subfamily and the DUF266 proteins as a GT14-like subfamily. This new gene family name is consistent with the current GT superfamily classification system, which contains more than 90 families (e.g., GT1, GT2, GT3, etc.) [5].

4.2. Evolutionary relationship between GT14 and GT14-like genes

GT14 subfamily members have been identified in both plants and animals [5]. It has been suggested that the GT14-like subfamily genes exist exclusively in plants [11]. We also constructed a phylogenetic tree (Supplemental Fig. 4) based on Core-2/I-Branching domain proteins in *Arabidopsis*, *Oryza*, *Populus*, *Sorghum*, *Vitis* and human. The phylogeny shows that GT14 genes exist in both human and plants whereas GT14-like genes are plant-specific, not being found in human. This suggests that GT14 genes have essential functions shared by animals and plants, and GT14-like genes are involved in biological processes specific to plants. However, by searching the NCBI nr protein database with the same HMMER-InterProScan approach as that used to identify GT14/GT14-like proteins in *Arabidopsis*, *Oryza*, *Populus*, *Sorghum* and *Vitis*, we found both GT14 and GT14-like subfamily genes in bacteria (data not shown), suggesting that GT14-like subfamily genes do not exist exclusively in plants. Moreover, the GT14-like genes may have evolved more diverse functions than the GT14 genes. Firstly, the expression of the *Arabidopsis–Populus* orthologous gene pairs in GT14-like subfamily is less conserved between the two species than that in the GT14 subfamily (Fig. 4). Secondly, the average of the pairwise TM-scores (reflecting similarity of topologies of two protein structures) between the representative proteins within the GT14-like subfamily (0.77) is significantly ($P < 0.0001$) lower than that between the representative proteins within the GT14 subfamily (0.82), indicating that there is more protein structural diversity within the GT14-like subfamily than within the GT14 subfamily. Finally, there is more diversity in protein subcellular localization within the GT14-like subfamily than within the GT14 subfamily (Table 3).

4.3. GT14/GT14-like family may play important roles in cell wall biosynthesis

Two members of the GT14 subfamily were identified as xylem-specific genes in *P. tremula* × *tremuloides*, indicating their potential role in secondary cell wall biosynthesis [10]. A recent genetic study showed that *OsBC10*, a GT14-like gene in *Oryza*, was involved in cellulose biosynthesis and formation of arabinogalactan protein [11], suggesting that GT14-like genes are related to cell wall biosynthesis. In this study, our gene expression analysis revealed that approximately one-half of the GT14/GT14-like family members in both *Arabidopsis* and *Populus* were preferentially expressed in stem/xylem (Fig. 2), suggesting that the GT14/GT14-like gene family may play important roles in the development of stem/xylem. *Populus* xylem tissue is enriched with cell walls, the most abundant source of cellulosic biomass for bioethanol production. *Populus* GT14/GT14-like genes (i.e., POPTR0010s13150, POPTR0017s11200, POPTR0008s12700, POPTR0008s00710) showing preferential expression in xylem (Fig. 2B) could be high-likelihood candidates for future studies on functional genomics of cell wall biosynthesis, toward overcoming biomass recalcitrance for biofuels production in *Populus*, a major bioenergy crop.

Acknowledgments

We thank S.D. Wullschleger and J. Chen for insightful comments on the manuscript. This research was supported by the U.S. DOE BioEnergy Science Center. The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract Number DE–AC05–00OR22725.

Please cite this article in press as: C.-Y. Ye, et al., Comparative analysis of GT14/GT14-like gene family in *Arabidopsis*, *Oryza*, *Populus*, *Sorghum* and *Vitis*, Plant Sci. (2011), doi:10.1016/j.plantsci.2011.01.021

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.plantsci.2011.01.021.

References

[1] P.J. Cao, L.E. Bartley, K.H. Jung, P.C. Ronald, Construction of a rice glycosyltransferase phylogenomic database and identification of rice-diverged glycosyltransferases, Mol. Plant 1 (2008) 858–877.
[2] W.R. Scheible, M. Pauly, Glycosyltransferases and cell wall biosynthesis: novel players and insights, Curr. Opin. Plant Biol. 7 (2004) 285–295.
[3] K. Keegstra, N. Raikhel, Plant glycosyltransferases, Curr. Opin. Plant Biol. 4 (2001) 219–224.
[4] E.K. Lim, Plant glycosyltransferases: their potential as novel biocatalysts, Chem. Eur. J. 11 (2005) 5486–5494.
[5] B.L. Cantarel, P.M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, B. Henrissat, The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics, Nucleic Acids Res. 37 (2009) D233–D238.
[6] T. Richmond, Higher plant cellulose synthases, Genome Biol. 1 (2000), reviews3001.1-3001.6.
[7] G.F. Vanzin, M. Madson, N.C. Carpita, N.V. Raikhel, K. Keegstra, W.D. Reiter, The *mur2* mutant of *Arabidopsis thaliana* lacks fucosylated xyloglucan because of a lesion in fucosyltransferase AtFUT1, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 3340–3345.
[8] J.D. Sterling, M.A. Atmodjo, S.E. Inwood, V.S. Kumar Kolli, H.F. Quigley, M.G. Hahn, D. Mohnen, Functional identification of an *Arabidopsis* pectin biosynthetic homogalacturonan galacturonosyltransferase, Proc. Natl. Acad. Sci. U.S.A. 103 (2006) 5236–5241.
[9] R.Q. Zhong, M.J. Pena, G.K. Zhou, C.J. Nairn, A. Wood-Jones, E.A. Richardson, W.H. Morrison, A.G. Darvill, W.S. York, Z.H. Ye, *Arabidopsis fragile fiber8*, which encodes a putative glucuronyltransferase, is essential for normal secondary wall synthesis, Plant Cell 17 (2005) 3390–3408.
[10] H. Aspeborg, J. Schrader, P.M. Coutinho, M. Stam, A. Kallas, S. Djerbi, P. Nilsson, S. Denman, B. Amini, F. Sterky, et al., Carbohydrate-active enzymes involved in the secondary cell wall biogenesis in hybrid aspen, Plant Physiol. 137 (2005) 983–997.
[11] Y. Zhou, S. Li, Q. Qian, D. Zeng, M. Zhang, L. Guo, X. Liu, B. Zhang, L. Deng, G. Luo, et al., BC10, a DUF266-containing and golgi-located type II membrane protein, is required for cell-wall biosynthesis in rice (*Oryza sativa* L.), Plant J. 57 (2009) 446–462.
[12] S.F. Hansen, E. Bettler, M. Wimmerova, A. Imberty, O. Lerouxel, C. Breton, Combination of several bioinformatics approaches for the identification of new putative glycosyltransferases in *Arabidopsis*, J. Proteome Res. 8 (2009) 743–753.
[13] R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, et al., The Pfam protein families database, Nucleic Acids Res. 38 (2010) D211–D222.
[14] S.R. Eddy, Profile hidden Markov models, Bioinformatics 14 (1998) 755–763.
[15] X.H. Yang, U.C. Kalluri, S. Jawdy, L.E. Gunter, T.M. Yin, T.J. Tschaplinski, D.J. Weston, P. Ranjan, G.A. Tuskan, The F-Box gene family is expanded in herbaceous annual plants relative to woody perennial plants, Plant Physiol. 148 (2008) 1189–1200.
[16] S. Hunter, R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, et al., InterPro: the integrative protein signature database, Nucleic Acids Res. 37 (2009) D211–D215.
[17] K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Res. 30 (2002) 3059–3066.
[18] A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J.F. Dufayard, S. Guindon, V. Lefort, M. Lescot, et al., Phylogeny.fr: robust phylogenetic analysis for the non-specialist, Nucleic Acids Res. 36 (2008) W465–W469.
[19] T.M. Keane, C.J. Creevey, M.M. Pentony, T.J. Naughton, J.O. Mclnerney, Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified, BMC Evol. Biol. 6 (2006) 29.
[20] K. Chen, D. Durand, M. Farach-Colton, NOTUNG: a program for dating gene duplications and optimizing gene family trees, J. Comput. Biol. 7 (2000) 429–447.
[21] S. Kumar, M. Nei, J. Dudley, K. Tamura, MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences, Brief Bioinform. 9 (2008) 299–306.
[22] M. Schmid, T.S. Davison, S.R. Henz, U.J. Pape, M. Demar, M. Vingron, B. Scholkopf, D. Weigel, J.U. Lohmann, A gene expression map of *Arabidopsis thaliana* development, Nat. Genet. 37 (2005) 501–506.
[23] O. Wilkins, H. Nahal, J. Foong, N.J. Provart, M.M. Campbell, Expansion and diversification of the *Populus* R2R3-MYB family of transcription factors, Plant Physiol. 149 (2009) 981–993.
[24] J. Herrero, A. Valencia, J. Dopazo, A hierarchical unsupervised growing neural network for clustering gene expression patterns, Bioinformatics 17 (2001) 126–136.
[25] M. Hertzberg, H. Aspeborg, J. Schrader, A. Andersson, R. Erlandsson, K. Blomqvist, R. Bhalerao, M. Uhlen, T.T. Teeri, J. Lundeberg, et al., A transcriptional roadmap to wood formation, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 14732–14737.
[26] J. Schrader, J. Nilsson, E. Mellerowicz, A. Berglund, P. Nilsson, M. Hertzberg, G. Sandberg, A high-resolution transcript profile across the wood-forming meristem of poplar identifies potential regulators of cambial stem cell identity, Plant Cell 16 (2004) 2278–2292.
[27] A. Sjodin, N.R. Street, G. Sandberg, P. Gustafsson, S. Jansson, The *Populus* Genome Integrative Explorer (PopGenIE): a new resource for exploring the *Populus* genome, New Phytol. 182 (2009) 1013–1025.
[28] A. Roy, A. Kucukural, Y. Zhang, I-TASSER: a unified platform for automated protein structure and function prediction, Nat. Protoc. 5 (2010) 725–738.
[29] Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score, Nucleic Acids Res. 33 (2005) 2302–2309.
[30] S. Briesemeister, J. Rahnenfuhrer, O. Kohlbacher, YLoc—an interpretable web server for predicting subcellular localization, Nucleic Acids Res. 38 (2010) W497–W502.
[31] H. Tang, X. Wang, J.E. Bowers, R. Ming, M. Alam, A.H. Paterson, Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps, Genome Res. 18 (2008) 1944–1954.
[32] G.A. Tuskan, S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, et al., The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray), Science 313 (2006) 1596–1604.
[33] X.H. Yang, U.C. Kalluri, S.P. DiFazio, S.D. Wullschleger, T.J. Tschaplinski, Z.M. Cheng, G.A. Tuskan, Poplar genomics: state of the science, Crit. Rev. Plant Sci. 28 (2009) 285–308.
[34] A. Force, M. Lynch, F.B. Pickett, A. Amores, Y.L. Yan, J. Postlethwait, Preservation of duplicate genes by complementary, degenerative mutations, Genetics 151 (1999) 1531–1545.
[35] R. Hovav, J.A. Udall, B. Chaudhary, R. Rapp, L. Flagel, J.F. Wendel, Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant, Proc. Natl. Acad. Sci. U.S.A. 105 (2008) 6191–6195.
[36] X. Yang, G.A. Tuskan, Z.M. Cheng, Divergence of the Dof gene families in poplar, *Arabidopsis*, and rice suggests multiple modes of gene evolution after duplication, Plant Physiol. 142 (2006) 820–830.
[37] Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality, Proteins 57 (2004) 702–710.
[38] R. Perrin, C. Wilkerson, K. Keegstra, Golgi enzymes that synthesize plant cell wall polysaccharides: finding and evaluating candidates in the genomic era, Plant Mol. Biol. 47 (2001) 115–130.

Please cite this article in press as: C.-Y. Ye, et al., Comparative analysis of GT14/GT14-like gene family in *Arabidopsis*, *Oryza*, *Populus*, *Sorghum* and *Vitis*, Plant Sci. (2011), doi:10.1016/j.plantsci.2011.01.021