Genomics 92 (2008) 246–253

Contents lists available at ScienceDirect

Genomics

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

Genomewide comparative phylogenetic and molecular evolutionary analysis of tubby-like protein family in *Arabidopsis*, rice, and poplar

Zefeng Yang [1], Yong Zhou [1], Xuefeng Wang, Shiliang Gu, Jianmin Yu, Guohua Liang, Changjie Yan, Chenwu Xu *

Jiangsu Provincial Key Laboratory of Crop Genetics and Physiology, Key Laboratory of Plant Functional Genomics of Ministry of Education, Yangzhou University, Yangzhou, 225009, China

ARTICLE INFO

Article history:
Received 7 December 2007
Accepted 9 June 2008
Available online 23 July 2008

Keywords:
Tubby-like protein gene family
Species-specific expansion
Segmental duplication
Co-evolution

ABSTRACT

Tubby-like proteins, which are characterized by a highly conserved tubby domain, play an important role in the maintenance and function of neuronal cells during postdifferentiation and development in mammals. In additional to the tubby domain, most tubby-like proteins in plants also possess an F-box domain. Plants also appear to harbor a large number of *TLP* genes. To gain insight into how *TLP* genes evolved in plants, we conducted a comparative phylogenetic and molecular evolutionary analysis of the tubby-like protein gene family in *Arabidopsis*, rice, and poplar. Genomewide screening identified 11 *TLP* genes in *Arabidopsis*, 14 in rice, and 11 in poplar. Phylogenetic trees, domain organizations, and intron/exon structures classified this family into three subfamilies and indicated that species-specific expansion contributed to the evolution of this family in plants. We determined that in rice and poplar, the tubby-like protein family had expanded mainly through segmental duplication events. Tissue-specific expression analysis indicated that functional diversification of the duplicated *TLP* genes was a major feature of long-term evolution. Our results also demonstrated that the tubby and F-box domains had co-evolved during the evolution of proteins containing both domains.

In mammals, *TLP* (tubby-like protein) genes play important roles in maintenance and function of neuronal cells during postdifferentiation and development [1]. Tubby, the base for tubby-like proteins, was first identified from obese mice through positional cloning [2,3]. In addition to *TUBBY*, three other members (*TULP1, TULP2,* and *TULP3*) of this gene family that encoded tubby-like proteins have also been identified in humans and mice [4]. Tubby-like proteins in animals are characterized by a highly conserved domain of about 270 amino acids, the tubby (TUB) domain, which is located at the C terminus, but their N-terminal sequences are quite divergent [4]. Until now, tubby-like proteins had been found in many multicellular organisms from both the plant and animal kingdoms. In addition to TUB domains at the C terminus, most tubby-like proteins in plants also contain highly conserved F-box domains [5]. The C terminus of F-box proteins generally contains one or several highly variable protein-protein interaction domains, such as the Leu-rich repeat (LRR), kelch repeat, tetratricopeptide repeat (TPR), and WD40 repeat [6]. The characteristics of the highly conserved TUB domain in different species demonstrated that these proteins must have fundamental biological functions in multicellular organisms.

Comparatively little research, however, has been done in plants. In the model plant *Arabidopsis*, 11 members of this family had been identified, named *AtTLP1* to *AtLP11* [5]. Among them, *AtTLP1, AtTLP2, AtTL3, AtTLP6, AtTLP7, AtTLP9, AtTLP10,* and *AtTLP11* were expressed ubiquitously in all the organs tested, but expression of *AtTLP5* and *AtTLP8* exhibited dramatic organ specificity. Interaction between AtTLP9 and ASK1 has also been confirmed. And transgenic plants overexpressing *AtTLP9* were shown to be hypersensitive to ABA, suggesting that *AtTLP9* may participate in the ABA signaling pathway. In addition to *Arabidopsis*, *TLP* genes have been demonstrated by homology searches to be present in other plants, such as *Lemna paucicostata* [5], *Oryza sativa* [5], *Cicer arietinum* [5], and *Zea mays* [5,7].

It is well known that gene duplication events are important to gene family evolution, which can occur via three major mechanisms: segmental duplication, tandem duplication, and transposition events such as retroposition and replicative transposition [8]. Among these, tandem and segmental duplication events contribute mostly to the generation of new members in nuclear gene families. Cannon et al. [9] analyzed 50 gene families in *Arabidopsis* and reported that tandem duplications were most prominent in some gene families, whereas segmental duplications occurred more frequently in others. Plants, in particular, appear to harbor a large number of *TLP* genes [1]. Establishment of the complete genomic sequences of *Arabidopsis* [10], rice [11–13],

Abbreviations: ABA, abscisic acid; EST, expressed sequence tag; ME, minimal evolution; MP, maximum parsimony; NJ, neighbor-joining; OC, orthologous cluster; RT-PCR, reverse transcription polymerase chain reaction; TLP, tubby-like protein; TUB domain, tubby domain.

* Corresponding author. Fax: +86 514 87996817.
E-mail address: qtls@yzu.edu.cn (C. Xu).

[1] These authors contributed equally to this work.

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.06.001

and poplar [14] gave us the opportunity to learn all the *TLP* genes in these three plant species and to investigate the expansion patterns of this family. We identified 11, 14, and 11 *TLP* genes in *Arabidopsis*, rice, and poplar, respectively. A phylogenetic tree was constructed to evaluate the evolutionary relationships of *TLP* genes in the three plant species. We also examined the chromosomal distribution of *TLP* genes to explore potential mechanisms leading to their species-specific expansion in plants. Co-evolutionary analysis of F-box and TUB domains was performed. RT-PCR and in silico data analysis were used to examine tissue-specific expression patterns and functional diversification of paralogous *TLP* genes. To examine the driving force for duplicated genes, we performed nonsynonymous and synonymous rate (K_a and K_s) analyses of the paralogous genes. Our systematic analysis provides a solid foundation for further functional dissection of *TLP* genes in plants.

Results

Collection of TLP genes in Arabidopsis, rice, and poplar

After careful survey of three plant genomes, 11 members of the TLP family in *Arabidopsis*, 14 in rice, and 11 in poplar were identified. The nucleotides, CDS, and protein sequences of *TLP* genes in *Arabidopsis* and rice were downloaded from the TIGR database (Supplementary Table 1), whereas those in poplar were downloaded from the JGI *Populus trichocarpa* database (Supplementary Table 2). Domain detection showed that plant TLP proteins can be classified into three major classes according their domain organization. Most proteins (30 of 36) contain both a highly conserved F-box domain at the N terminus and a TUB domain at the C terminus. Five proteins (AtSPL8 in *Arabidopsis*, OsTLP4 in rice, and PtTLP1, PtTLP7, and PtTLP10 in poplar) contain only the TUB domain at the C terminus. AtTLP4 was the only protein that contained two short TUB domains in the middle.

Arabidopsis TLP genes were dispersed on all the chromosomes except chromosome 4. Seven *TLP* genes were found on chromosome 1, two on chromosome 2, and one each on chromosomes 3 and 5, respectively. The 14 rice *TLP* genes were present on 9 of 12 rice chromosomes. One rice *TLP* gene each was located on chromosomes 3, 4, 7, 8, 11, and 12, two on chromosome 2, and three each on chromosomes 1 and 5, respectively. There are 19 chromosomes in the poplar genome. In our analysis, two poplar *TLP* genes were localized to a scaffold that had not been mapped on chromosomes. Among the other poplar genes that were mapped, one gene each was located on chromosomes 1, 7, 8, 10, and 11, and two each on chromosomes 2 and 5.

Phylogenetic relationships of TLP family in Arabidopsis, rice, and poplar

To investigate the molecular evolution and phylogenetic relationships among *TLPs* in *Arabidopsis*, rice, and poplar, three combined phylogenetic trees were constructed with the neighbor-joining (NJ), minimum evolution (ME), and maximum parsimony (MP) methods, respectively. They exhibited the same topology. We selected only the NJ tree for further analysis, as it was supported by the highest bootstrap values. The NJ phylogenetic tree divided the plant *TLP* genes into three distinct subfamilies: A, B, and C (Fig. 1). Subfamily B contained only one member, AtTLP4 in *Arabidopsis*; its domain structure was far different from those of other members of this family, suggesting that *AtTLP4* independently evolved in the *Arabidopsis* genome. The alternative explanation is that members of this subfamily in rice and poplar were lost during the long evolutionary period. Subfamily C contained three proteins: AtTLP8 in *Arabidopsis*, OsTLP4 in rice, and PtTLP1 in poplar. One apparent feature of all three proteins was a TUB domain at the C terminus. These three genes, therefore, may have originated from one gene in an ancestral species, and did not expand after the split between dicot and monocot. All proteins in subfamily A contained both highly conserved TUB and F-box domains except for PtTLP7 and PtTLP10. The proteins in subfamily A were further divided into four distinct orthologous clusters (OCs): A1–A4. These four OCs all contained *Arabidopsis*, rice, and poplar TLP proteins, indicating that the main characteristics of this subfamily in plants were established before the dicot-monocot split.

Fourteen pairs of paralogous genes (three for *Arabidopsis*, six for rice, and five for poplar) were identified at the terminus of the phylogenetic tree. All paralogous genes belonged to subfamily A. This result indicates that most of the *TLP* genes belonging to subfamily A in *Arabidopsis*, rice, and poplar had expanded in a species-specific manner, and probably only a few members originated from the common ancestral genes that existed before the divergence of monocot and dicot. The species differed with respect to expansion of the four OCs; for example, the genes in rice and poplar had expanded in OC A4, but the *Arabidopsis* gene in OC A4 had not expanded. It is also interesting that all of the poplar *TLP* genes in subfamily A were followed in paralogous pairs.

Sequence alignment of the TLP proteins in Arabidopsis, rice, and poplar

All subfamily A proteins contained conserved TUB domains, and most of them also contained highly conserved F-box domains except for PtTLP7 and PtTLP10. Genes *PtTLP7* and *PtTLP10* were demonstrated to be paralogous to each other, and orthologous to *AtTLP7* in *Arabidopsis*, whereas the latter possessed an F-box domain in the middle of its sequence. So it may be deduced that the F-box domain was lost during the long evolutionary period of proteins PtTLP7 and PtTLP10. We aligned all the amino acid sequences of proteins in subfamily A (Supplementary Fig. 1), and found that there were four conserved blocks for the TUB domain in all the protein sequences. PROSITESCAN (http://www.expasy.ch/tools/scanprosite/) was used to search the PROSITE database for functional motifs. Most proteins in subfamily A contained two signature patterns called the TUB1 and TUB2 motifs (Fig. 1). These two motifs were all located at the C termini of TUB domains and contained 14 and 16 amino acid residues, respectively. Although the similarity of the TUB domains was not very high, we found some highly conserved blocks in these domains. For instance, a motif in the middle of the domains with amino acid residues PGPRRM was highly conserved in all proteins, which could be another feature of TUB domains.

The sequences connecting the F-box and TUB domains were also found to be conserved. The sequence length for this segment in all two-domain-containing proteins of this subfamily was 10 amino acid residues, except in OsTLP10, where it was 56 residues, suggesting that the F-box and TUB domains and the sequence connecting them should be evolutionarily conserved among plants. No significant conserved sequences were detected at the N termini of proteins in this family.

Analysis of the intron distributions of the TLP genes in Arabidopsis, rice, and poplar

The intron distribution can also provide important evidence to support phylogenetic relationships in a gene family [15]. In the TLP family, genes in subfamily C were shown to possess more introns than genes in subfamilies A and B (Fig. 1). There were seven introns in the rice gene *OsTLP4*, whereas eight introns were found in *Arabidopsis* gene *AtTLP8* and poplar gene *PtTLP1*. Comparison of the sequences revealed that a redundant intron located at the N terminus had truncated the first exon into two small exons in *Arabidopsis* and poplar. This intron was found only in dicots and might have been gained after the split between dicot and monocot. The majority of the genes in OCs A1, A2, and A4 had three introns (22/25) and their positions were highly conserved. All genes in OC A3 contained four introns and their positions were also conserved. The main difference between the genes in A3 and those in the other OCs was that a redundant intron located at the C terminus had truncated the last

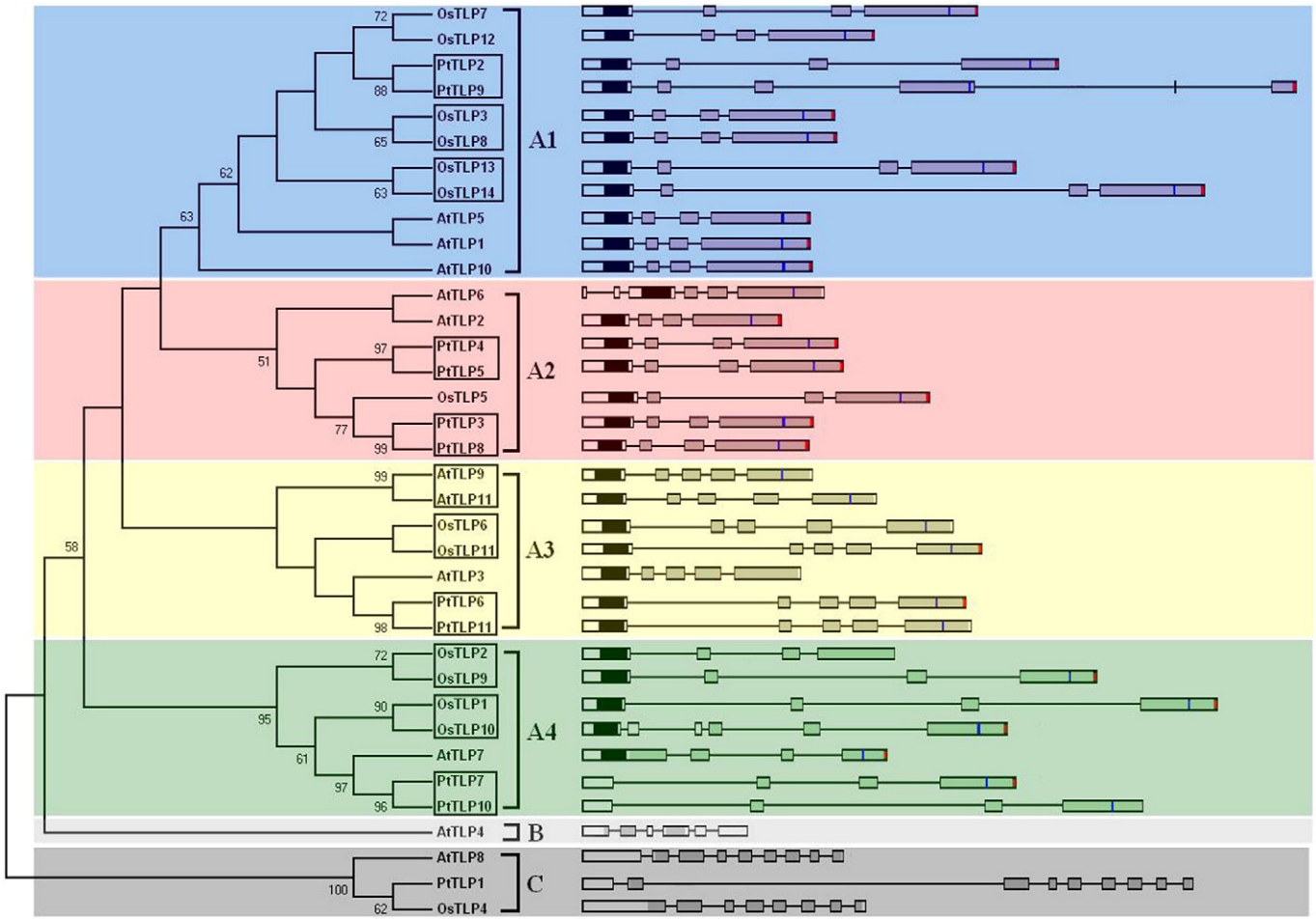

Fig. 1. Phylogenetic analysis and schematic diagram for intron/exon structures and conserved domains of TUB-like proteins in *Arabidopsis*, rice, and poplar. The tree was constructed from a complete alignment of 36 TLP proteins by neighbor-joining methods with bootstrapping analysis (1000 reiterations). The boxes and lines represented exons and introns, respectively. Positions of conserved F-box and TUB domains are also displayed. F-box domains are represented by black boxes, and TUB domains, by gray boxes. Two PROSITE signature patterns (TUB1 and TUB2 motifs) in the TUB domain are represented by the blue- and red-shaded boxes, respectively. Duplicated pairs that arose from segmental duplication events are also represented by boxes on the phylogenetic tree.

exon into two small exons. The gene structure in subfamily A also indicated that the F-box domain and sequences connecting the F-box and TUB domains were located in exon 1, whereas the TUB domain sequences were distributed in exons interrupted by either two or three introns in most genes of this subfamily. These results suggest that the exon/intron structure of this family developed before the monocot—dicot split. Although the intron positions were conserved in most genes of subfamily A, individual introns in different genes varied in length. Intron II, for examples, was 2851 bp long in rice gene *OsTLP14* but only 83 bp long in *Arabidopsis* gene *AtTLP10*. It is worth mentioning that the TUB domain mostly started at the second exon, which was 88 bp long in most genes in subfamily A.

Evolutionary patterns of TLP genes in Arabidopsis, rice, and poplar

Apparent tandem duplications among the TLP gene family in *Arabidopsis*, rice, and poplar were identified. We searched for contiguous *TLP* genes in both the sharing region and neighboring regions. But no genes in this family were found to be located in tandem repeats, indicating that tandem duplication did not contribute to the expansion of this family in these three organisms.

We also tested the hypothesis that large-scale duplication events played a leading role in the evolution of the TLP gene family in *Arabidopsis*, rice, and poplar. For each *TLP* gene, we tallied the number of flanking protein-coding genes with a best non-self match to a protein-coding gene neighboring its paralog (Table 1). For *Arabidopsis*, three pairs of paralogous genes were identified on the phylogenetic tree. There were six pairs of genes flanking *AtTLP9* on chromosome 3 and *AtTLP11* on chromosome 5 that showed high conservation, indicating that these two *TLP* genes were formed through large-scale duplication in *Arabidopsis*. We did not find evidence that other pairs of paralogous genes in *Arabidopsis* originated from duplicated blocks. These results indicate that the *Arabidopsis* TLP family arose mainly through random insertion events rather than tandem duplications and segmental duplications. In rice, there were six pairs of paralogous genes at the terminus of the phylogenetic tree. There were highly conserved genes

Table 1
Duplicated *TLP* genes and the number of conserved protein-coding genes flanking them in *Arabidopsis*, rice, and poplar

Duplicated TLP gene 1	Duplicated TLP gene 2	Number of conserved flanking protein-coding genes
AtTLP9	AtTLP11	6
OsTLP1	OsTLP10	6
OsTLP2	OsTLP9	2
OsTLP3	OsTLP8	10
OsTLP6	OsTLP11	3
OsTLP13	OsTLP14	2
PtTLP2	PtTLP9	13
PtTLP3	PtTLP8	15
PtTLP4	PtTLP5	8
PtTLP6	PtTLP11	3
PtTLP7	PtTLP10	8

Table 2
Estimates of absolute dates for large-scale duplication events in *Arabidopsis*, rice, and poplar

Duplicated pair	*n*	Mean K_s	SD K_s	Minimum K_s	Maximum K_s	Date (million years)
AtTLP9–AtTLP11	7	0.6670	0.2045	0.3278	0.9022	22.23
OsTLP1–OsTLP10	7	0.8763	0.1493	0.7442	1.1395	67.41
OsTLP2–OsTLP9	3	0.8045	0.0446	0.7720	0.8361	61.88
OsTLP3–OsTLP8	11	0.7691	0.3714	0.1363	1.6732	59.16
OsTLP6–OsTLP11	4	0.6184	0.0841	0.5361	0.7354	47.57
OsTLP13–OsTLP14	3	0.9487	0.2650	0.7102	1.2339	72.98
PtTLP2–PtTLP9	14	0.2220	0.0533	0.1397	0.3188	16.09
PtTLP3–PtTLP8	16	0.2592	0.0906	0.1243	0.4788	18.78
PtTLP4–PtTLP5	9	0.2082	0.0201	0.1845	0.2390	15.08
PtTLP6–PtTLP11	4	0.2072	0.0457	0.1485	0.2515	15.01
PtTLP7–PtTLP10	9	0.3789	0.3191	0.1145	0.9606	27.45

among the flanking regions for all pairs of paralogous genes except *OsTLP7*/ *OsTLP12*, indicating that this pair of paralogs formed through random translocations and insertion events, whereas other paralogous genes arose from segmental duplication events. These results illustrate that segmental duplication events were the dominant pattern in the evolution of *TLP* genes in rice. There were five pairs of paralogous *TLP* genes in poplar located at the terminus of the phylogenetic tree. Only *PtTLP1* had no paralogous gene in poplar. There were highly conserved genes among the flanking regions for all pairs of paralogous genes, suggesting that all of the paralogous *TLP* genes in poplar arose from segmental duplication events. Thus, although TLP family in *Arabidopsis*, rice, and poplar had expanded in a species-specific manner, their evolutionary patterns were quite different. The main evolutionary pattern for *TLP* genes in *Arabidopsis* was random translocation and insertion events, whereas in rice and poplar it was segmental duplication events.

We also used K_s as the proxy for time and the conserved flanking protein-coding genes to estimate the dates of the segmental duplication events. The mean K_s values and the estimated dates for all segmental duplication events corresponding to *TLP* genes are listed in Table 2. The segmental duplicated events in rice may have occurred earlier than those in *Arabidopsis* and poplar, which occurred within the last 47.57 to 72.98 million years. Only one segmental duplication event for *TLP* genes in *Arabidopsis* was identified in this analysis. Its mean K_s value was 0.9022, dating the duplication event to 22.23 million years ago. For poplar, the segmental duplication event for the pair *PtTLP7—PtTLP11* occurred approximately 27.45 million years ago, whereas other segmental duplication events had occurred within the last 15 to 19 million years.

Expression patterns of paralogous TLP genes

Paralogous genes in the same genome were created by gene duplication events and usually had different functions. To investigate whether paralogous *TLP* genes in the three species had different functions, we analyzed the tissue-specific expression patterns of paralogous *TLP* genes in *Arabidopsis*, rice, and poplar using three methods.

First, we tested tissue-specific expression of rice *TLP* genes using RT-PCR. Total RNA from stems, leaves, flower clusters, roots, and seeds of soil-grown rice were isolated for tissue-specific expression analysis of *OsTLP* genes. The results revealed that *OsTLP1* and *OsTLP5–OsTLP14* were expressed in all organs tested, although many appeared to contain quantitatively different levels of mRNA levels in certain tissues (Fig. 2). In contrast, *OsTLP2* was expressed primarily in flower clusters and seeds. *OsTLP3* was not expressed only in stems among the organs tested. The tissue-specific expression of *OsTLP2* and *OsTLP3* may reflect their specific roles in particular organs. It is worth noting that *OsTLP9* and *OsTLP8*, the paralogous genes for *OsTLP2* and *OsTLP3*, respectively, were expressed in all organs tested. Only one rice *TLP* gene, *OsTLP4*, which belonged to subfamily C, was found not to be expressed in all organs tested.

Next, we analyzed the *Arabidopsis* and rice massively parallel signature sequencing (MPSS) database. This database identified 17- or 20-bp sequence tags, each representing the 3′ end of a single mRNA detected in transcript libraries isolated from various tissues and treatments. Here, data for the full set of 17-base signature libraries for *Arabidopsis* and rice were analyzed (data not shown). We evaluated overlapping tissue expression in each of the *Arabidopsis* and rice paralog pairs for which gene expression data were available for both duplicates, and found that all of the paired paralogous *TLP* genes in *Arabidopsis* and rice exhibited different expression patterns. Among the nine pairs of *Arabidopsis* and rice *TLP* genes, four (*AtTLP9/11*, *OsTLP3/OsTLP8*, *OsTLP6/OsTLP11*, and *OsTLP13/OsTLP14*) had relatively high overlapping expression (> 50%); all of these duplicated pairs had arisen from segmental duplication events. Other pairs had low overlap (< 50%). No duplicate pairs exhibited a 100% unique tissue-specific expression pattern.

The information on expression of poplar *TLP* genes based on EST searches is summarized in Supplementary Table 3. All of the *TLP* genes in poplar seemed to be expressed in various tissues. Leaf showed the highest *TLP* gene expression. *PtTLP4* and *PtTLP6* were expressed in all tissues listed; other poplar *TLP* genes had more restricted expression profiles. *PtTLP1* from subfamily C possessed only two ESTs, which were expressed in suspension cells, whereas other poplar *TLP* genes were expressed in many organs. The results also revealed different expression patterns for poplar paralogous *TLP* genes; for example, *PtTLP4* was found to be expressed in all tissues tested, but its paralogous gene *AtTLP5* was present only in leaf and wood. There were

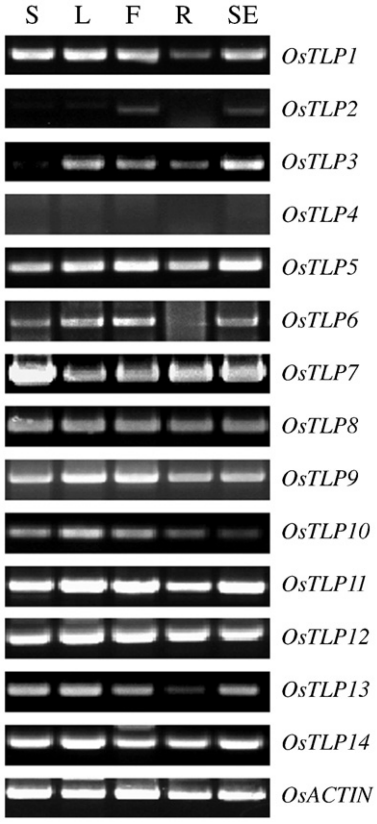

Fig. 2. Expression pattern of *TLP* genes in rice. RT-PCR analysis was performed with stems (S), leaves (L), flower clusters (F), roots (R), and seeds (SE) of rice. The rice *Actin* gene was used as an internal control.

also no duplicate pairs with non-overlapping expression among the paired paralogous poplar *TLP* genes.

Driving forces for genetic divergence

The ratio of K_a (nonsynonymous substitutions per nonsynonymous site) to K_S (synonymous substitutions per synonymous site) is a measure of selective constraint on coding sequences. To explore whether Darwinian positive selection was involved in driving gene divergence after duplication, the coding regions of 14 pairs of paralogs in *Arabidopsis*, rice, and poplar were used to calculate the K_a/K_s ratio with a sliding window (window size: 20AA, movement: 10AA). We also used MEME to identify conserved motifs in 14 pairs of paralogous proteins (Supplementary Fig. 2). A total of 10 conserved motifs were identified in all 28 TLP protein sequences. The corresponding sequences for motif 3, integer of motifs 5 and 9, motif 1, and motif 2 corresponded to conserved blocks 1, 2, 3, and 4, respectively. The integer of motifs 6 and 7 corresponded to F-box domain sequences in the middle of the proteins. The sequences connecting the F-box and TUB domains were also highly conserved, and their corresponding motif was motif 4. For all paralogs, K_a/K_s ratios were always close to zero for motif 2, which contained two signature patterns named theTUB1 and TUB2 motifs, suggesting strong purifying selection on this motif. In contrast, much higher K_a/K_s ratios were also found in the TUB domain regions outside motif 2, especially in the intermotif regions, and some ratios were greater than 1. These results indicate that both purifying and positive selections contributed to the evolution of the TUB domain in plants. The K_a/K_s ratios in F-box regions were generally less than 1, but higher than those in motif 2, suggesting relaxed purifying selection on this domain.

Co-evolution of F-box and TUB domains

Most proteins in the TLP family contained both F-box and TUB domains, and the sequence connecting them, generally 10 amino acid residues long, was also highly conserved. The sequences of F-box and TUB domains in this family were used to construct phylogenetic trees using NJ methods (Fig. 3). The two trees exhibited similar topology, only with minor modification at deep nodes. The two phylogenetic trees exhibited all four distinct clusters, and each cluster contained the same members in the F-box, TUB, and full-length protein trees, suggesting that the F-box and TUB domains had co-evolved in this family.

The method developed by Goh et al. [16] was used to test this hypothesis. Three correlation coefficients were obtained: 0.4037 for the TUB domain and interdomain, 0.4003 for the F-box domain and interdomain, and 0.8348 for the TUB and F-box domains. These correlation coefficients were all significantly greater than zero ($P<0.01$), indicating that they had undergone highly correlated co-evolution during plant evolution. Only the correlation coefficient between TUB and F-box was higher than 0.8, which, as suggested by Goh et al. [16], indicates co-evolution of the two domains. The possible co-evolution suggested interplay of these domains and interaction with a putative partner(s) for TLP protein function.

Discussion

The species-specific expansion of TLP genes in plants

Lineage-specific expansion is defined as the proliferation of a protein family in a particular lineage, relative to the sister lineage, with which it is compared [17,18]. Lineage-specific expansion of gene

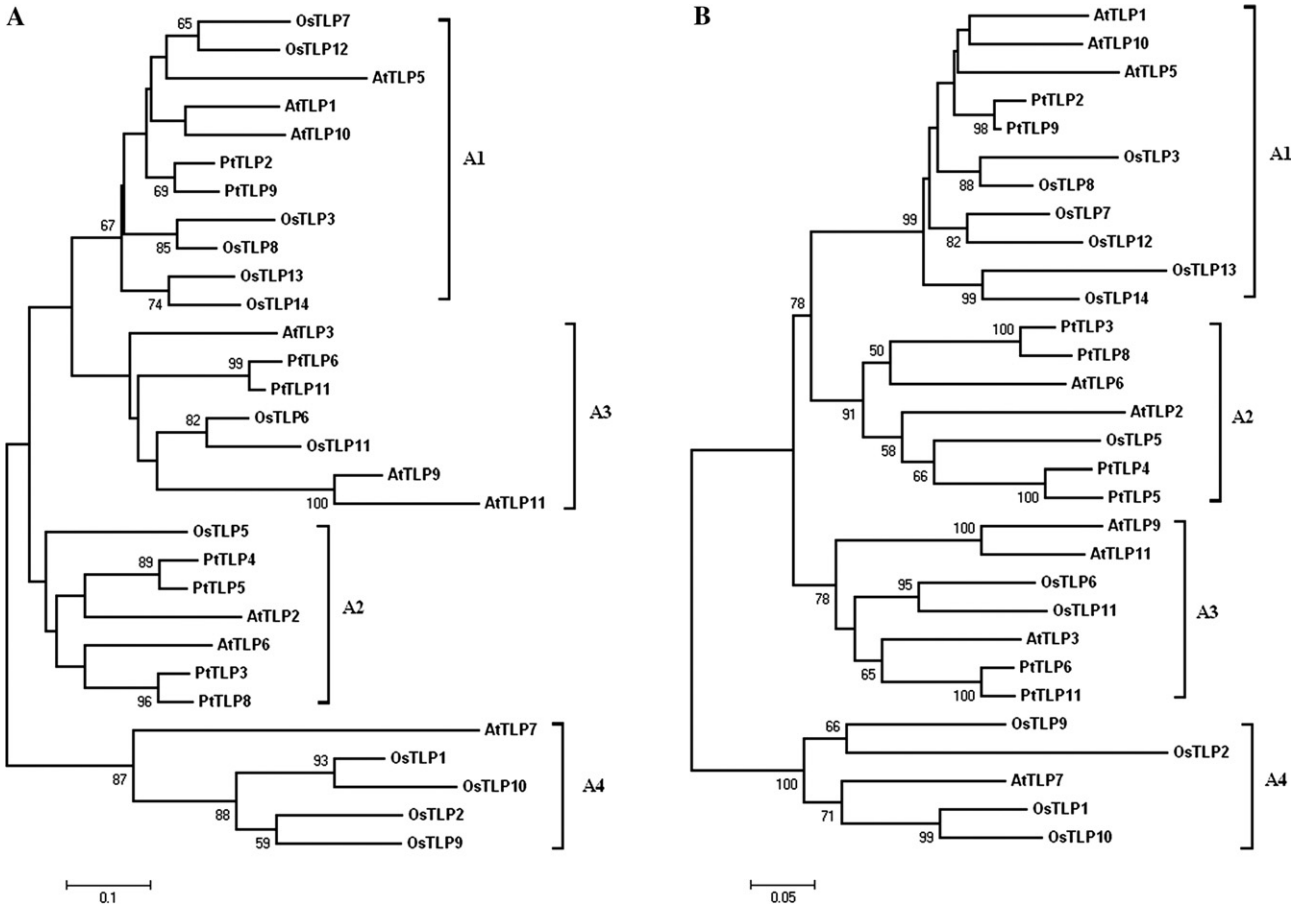

Fig. 3. Phylogenetic trees of the F-box (left) and TUB (right) domain sequences. The trees were inferred by the neighbor-joining method after alignment of the F-box and TUB domain amino acid sequences of the 30 proteins that contained both F-box and TUB domains in *Arabidopsis*, rice, and poplar.

families played an important role in the growth and differentiation of the proteomes of muticellular eukaryotes [18]. It was demonstrated that up to 80% of the genes in the model plant *Arabidopsis* are the result of lineage-specific expansion [18]. In the work described here, more members of TLP gene family were found in plants than in animals. And 14 pairs of paralogous genes were identified from the phylogenetic tree, illustrating that most *TLP* genes in three species had expanded in a species-specific manner, and probably only a few members had originated from common ancestral genes that existed before the divergence of monocot and endogen. This type of divergence between monocot and dicot species has also been observed for other gene families [19–21].

Gene duplication events are important to gene family evolution, because duplicated genes provide the raw materials for the generation of new genes, which, in turn, facilitate the generation of new functions [8]. Three principal evolutionary patterns were attributed to gene duplications: segmental duplication, tandem duplication, and transposition events such as retroposition and replicative transposition [8]. Among these, segmental duplication occurs most frequently in plants because most plants are diploidized polyploids and retain numerous duplicated chromosomal blocks within their genomes [9]. In previous studies, several rounds of whole-genome duplication in both the *Arabidopsis* and rice genomes were reported [22–25]. It was also demonstrated that large-scale gene duplication events had occurred in poplar, and all poplar species shared the same large-scale duplication events as determined by comparison of ESTs [26]. In our analysis, we found that the TLP family in rice and poplar had expanded mainly through segmental duplications, whereas only one of three pairs of paralogous *TLP* genes had expanded through segmental duplication in *Arabidopsis*. Other pairs of paralogous genes were found to be have evolved through random translocations and insertional events. No *TLP* genes in the genomes of these three species were located in tandem repeats. Yu et al. [22] found 18 distinct pairs of duplicated segments that cover 65.7% of the genome; 17 of these pairs date back to a common time before the divergence of the grasses. After examining the rice duplication blocks identified by Yu et al. [22], we noticed that part of the long arm of rice chromosome 1 (*OsTLP1–OsTLP3*) and part of the long arm of chromosome 5 (*OsTLP10–OsTLP8*) constituted a pair of duplicated segments. The other two duplicated segments (*OsTLP6–OsTLP11* and *OsTLP13–OsTLP14*) were consistent with the results of Yu et al. [22]. The paralogous *TLP* genes in poplar were all found to have formed through segmental duplication events in our analysis, though some genes were located in the scaffold and had not been mapped on chromosomes.

Functional diversification of paralogous TLP genes

One of the main goals of comparative phylognetic analysis is to identify putative orthologous and paralogous genes. *Orthologs* are defined as genes in different genomes that were created by the splitting of taxonomic lineages, whereas *paralogs* are genes in the same genome that were created by gene duplication events [27,28]. Orthologs usually retain the same function, whereas paralogs might have different functions [29]. Blanc et al. [30] analyzed the duplicated genes formed by polyploidy during *Arabidopsis* evolution and found that functional diversification of the surviving duplicated genes was a major feature of the long-term evolution of polyploids. Generally, duplicated genes, if not silenced (nonfunctionalization), may either acquire novel functions (neofunctionalization) or perform part of the original function (subfunctionalization) [31]. Recently, a new model called *subneofunctionalization* was proposed, according to which a short period of subfunctionalization, new function, and expression arose in both duplicated genes and lasted over a prolonged period [32]. To test the functional diversification of paralogous *TLP* genes in *Arabidopsis*, rice, and poplar, their tissue-specific expression patterns were analyzed by both RT-PCR analysis and database searches. There was both experimental proof and bioinformatic proof that functional diversification contributed to the evolution of *TLP* paralogous genes in the three organisms. Among all of the duplicate pairs, none exhibited a 100% unique tissue-specific expression pattern, nor did both genes in one duplicate pair share the same expression pattern, indicating that neither subfunctionalization nor nonfunctionalization describes the functional diversification of paralogous *TLP* genes. So we concluded that the neofunctionalization and/or subneofunctionalization may contribute to the maintenance of all duplicated *TLP* genes.

The K_a/K_s ratio provides a sensitive measure of selective pressure on the protein. Most amino acids in a functional protein are under structural and functional constraints, and adaptive evolution probably affects only a few sites at a few time points. So positive selection was thought to be one of the major forces in the emergence of new motifs/functions in protein after gene duplication [33]. Through tissue-specific expression pattern analysis of paralogous *TLP* genes, all of the duplicate pairs showed evidence of functional diversification. The K_a/K_s ratios indicated that most of the duplicated pairs possessed sites or regions that were under positive selection, and all of the other pairs had some sites or regions under neutral selection. Maybe positive selection and/or neutral selection promoted the functional diversification of paralogous *TLP* genes in *Arabidopsis*, rice, and poplar.

Co-evolution of TUB and F-box domains

The co-evolution of domains within a single protein is more easily understood than the co-evolution of proteins produced by different genes, because the domains within a single protein are covalently linked to one another by the polypeptide chain, and the relationship between any two domains that interact with each other is one by one. The method of Goh et al. [16] could be used to deduce the co-evolution of domains in a single protein. This method was based on the assumption that changes in the amino acid sequence within one of the domains would result either in counterselection or in compensation of changes in the amino acid sequence of the other domain if the two domains in a single protein acted cooperatively for proper function [34]. It was also demonstrated that about 70% true interactions could be detected when the empiric cutoff value (0.8) was used [35]. The PGK [16], MLO [34], and LSD1-like [36] gene families all possess multiple domains in their protein products, and it has been illustrated that the domains in a single protein co-evolved during the evolutionary periods of these families using this method. Most TLP proteins in our analysis contained both F-box and TUB domains. The high correlation coefficient between TUB and F-box domains provided the evidence for their co-evolution and interplay. The phylogenetic trees for the F-box and TUB domain sequences showed similar topology, another indicator of co-evolution.

Materials and methods

Sequence database searches

Multiple database searches were performed to collect all members of the *TLP* family in *Arabidopsis*, rice, and poplar. First, BLASTP searches against the TIGR *Arabidopsis thaliana* database (http://www.tigr.org/tdb/e2k1/ath1/), the TIGR rice annotation database (http://www.tigr.org/tdb/e2k1/osa1/), and JGI *Populus trichocarpa* v1.1 Home (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html) were performed using consensus TUB domain sequences as queries. The programs TBLSATN and BLASTP were also used to search the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/blast/). If a protein sequence satisfied $E \leq 10^{-10}$, it was selected as a candidate protein. Second, the tool Pfam [37] was used to predict

the TUB domain (PF01167) of all candidate proteins. The deduced nucleotide and protein sequences of *TLP* genes in *Arabidopsis* and rice were downloaded from the TIGR database, and the sequences of *TLP* genes in poplar were downloaded from the JGI *Populus trichocarpa* database.

Multiple sequence alignment, phylogenetic tree construction, and motif identification

Multiple sequence alignment of TLP proteins was performed using the Clustal X 1.83 program [38] with default parameters and displayed using Genedoc software (http://www.nrbsc.org/gfx/genedoc/index.html). To construct the combined phylogenetic tree of TLP proteins in *Arabidopsis*, rice, and poplar, multiple sequence alignments performed using Clustal X were saved and executed by MEGA Version 4.0 [39] to generate NJ, ME, and MP trees with bootstrapping analysis. Phylogenetic trees were also viewed with the help of MEGA. Motifs of the paralogous TLP proteins were identified statistically using MEME [40] with default settings, except that the maximum number of motifs to find was set at 10, and motif length, at 6–200.

Analysis of TLP gene expansion patterns

For this analysis, we focused on the processes of segmental and tandem duplication. To categorize apparent expansions of the TLP gene family, we looked at the physical locations of all members of this family in *Arabidopsis*, rice, and poplar. Tandem duplications were characterized as multiple members of this family occurring within the same intergenic region or neighboring intergenic regions. To identify segmental duplications, a method similar to that of Maher et al. [41] was used. We first identified the paralogous *TLP* genes at the terminal nodes of the phylogenetic tree. Second, 10 protein-coding genes upstream and downstream of each pair of psralogs were obtained from the Gramene database [42] for *Arabidopsis* and rice; the same searches were also performed in the JGI *Populus trichocarpa* database for poplar. As the last step, the genes flanking one *TLP* gene were used to match the genes flanking the other *TLP* gene in one pair of paralogs. Therefore, we considered paralogous *TLP* genes to originate from a duplication event if they resided within a region of conserved protein-coding genes.

Calculating K_a and K_s and dating the duplication events

The paralogs for *TLP* genes in *Arabidopsis*, rice, and poplar were inferred from the phylogenetic tree. Pairwise alignments of the paralogous nucleotide sequences were performed using Clustal X 1.83, with the corresponding protein sequences as alignment guides. Gaps in the alignments were removed manually. The program K-Estimator 6.1 [43] was used to estimate K_a and K_s for the paralogous genes.

To better explain these patterns of macroevolution, estimates of the evolutionary rates would be extremely useful. Assuming a molecular clock, the synonymous substitution rates (K_s) of duplication genes are expected to be similar over time [44], so we used K_s as the proxy for time and the conserved flanking protein-coding genes to estimate the dates of the segmental duplication events. The mean K_s value was calculated for each pair of protein-coding genes within a duplicated block and then used to date the duplication events. K_s values greater than 2.0 were discarded because of the risk of saturation. The approximate date of the duplication event was then calculated using the mean K_s values ($T=K_s/2\lambda$), assuming clocklike rates (λ) of synonymous substitution of 1.5×10^{-8} substitutions/synonymous site/year for *Arabidopsis* [45], 6.5×10^{-9} for rice [22], and 9.1×10^{-9} for poplar [46].

Tissue-specific expression pattern analysis

Tissue-specific expression patterns of *OsTLP* genes were studied using a RT-PCR-based method. Total RNA was isolated from roots, stems, leaves, flower clusters, and seeds of rice variety Nipponbare using the TRLzol method according to the manufacturer's instructions, followed by DNase I treatment to remove any genomic DNA contamination. Reactions were performed using Super Scirpt One-Step RT-PCR with Platinum Taq and 100 ng RNA from each sample. The thermal cycling conditions were 30 min at 50°C, 2 min at 94°C, 35 cycles of 30 s at 94°C, 45 s at 54°C, and 1 min at 72°C, and a final extension of 10 min at 72°C. Amplification products were fractionated on 10% agarose gel. The gene-specific primers used are listed in Supplementary Table 4.

In addition to the RT-PCR analysis of *TLP* genes in rice, we evaluated tissue samples from the 17-bp signatures in the *Arabidopsis* and rice MPSS database [47]. Overlapping tissue expression for *Arabidopsis* and rice *TLP* paralog pairs was determined by calculating the ratio of the number of tissues in which both duplicates were expressed to the number of tissues in which at least one duplicate was expressed.

ESTs provide a useful means of studying mRNA expression profiles (digital Northerns) [48]. To gain insight into poplar *TLP* gene expression patterns, we carried out an in silico expression study using *Populus* EST sequences. Comparison of protein sequences of different *Populus* species revealed that their sequences were highly similar or nearly identical [49]. As sequences from various *Populus* species were highly similar, we also used the EST sequences from all eight closely related *Populus* species or hybrids that were available at the NCBI EST database: *P. trichocarpa*, *P. tremula* × *P. alba*, *P. tremula* × *P. tremuloides*, *P. deltoids*, *P. tremuloides*, *P. euphratica*, *P. trichocarpa* × *P. deltodides*, and *P. trichocarpa* × *P. nigra*. Only one EST with >95% identity, *E* value $<10^{-10}$, and length >160 bp was counted as a match to query *TLP* genes. To get a better picture of the tissue-specific expression of poplar *TLP* genes, we divided the ESTs into eight tissue categories: flower, bark, leaf, root, shoot, wood, cambium, and other.

Co-evolution analysis

There were two highly conserved domains in most plant TLP proteins; one was the F-box domain, the other was the TUB domain. The sequences connecting them were also highly conserved. So they might have co-evolved during the evolutionary period. To test this hypothesis, we employed Goh and colleagues' method [16] to perform the correlation analysis on every possible domain-domain pair for the TLP family. Full-length proteins containing F-box and TUB domains were dissected into four single regions (N terminus, F-box domain, interdomain, and TUB domain). As there were significant differences among the sequences of N-terminal regions, only the sequences of the F-box domain, interdomain, and TUB domain were independently aligned using Clustal X with the default parameters; pairwise evolutionary distances for all multiple alignments were calculated using MEGA 4. Then, linear Pearson correlation coefficients (r) between the distance matrices of all possible interacting regions were calculated. Positive values of r indicate a positive correlation, and r values around zero indicate no correlation. Additionally, negative values of r indicate anticorrelation [16].

Acknowledgments

The authors are grateful to the editor and two anonymous reviewers for their helpful comments and criticisms. This work was supported by the National Basic Research Program of China (Grant 2006CB101700), the National High-Tech R and D Program (Grant 2006AA10Z165), the Program for New Century Excellent Talents in University (no. NCET2005-05-0502), and Program for Innovative Research of Graduate Students in Jiangsu Province (Grant CX07B-186z).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2008.06.001.

References

[1] A. Ikeda, P.M. Nishina, J.K. Naggert, The tubby-like proteins, a family with roles in neuronal development and function, J. Cell Sci. 115 (2002) 9–14.
[2] P.W. Kleyn, et al., Identification and characterization of the mouse obesity gene tubby: a member of a novel gene family, Cell 85 (1996) 281–290.
[3] K. Noben-Trauth, J.K. Naggert, M.A. North, P.M. Nishina, A candidate gene for the mouse mutation tubby, Nature 380 (1996) 534–538.
[4] M.A. North, J.K. Naggert, Y. Yan, K. Noben-Trauth, P.M. Nishina, Molecular characterization of TUB, TULP1, and TULP2, members of the novel tubby gene family and their possible relation to ocular diseases, Proc. Natl. Acad. Sci. USA 94 (1997) 3128–3133.
[5] C.P. Lai, et al., Molecular analyses of the *Arabidopsis* TUBBY-like protein gene family, Plant Physiol. 134 (2004) 1586–1597.
[6] M. Jain, et al., F-box proteins in rice, Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress, Plant Physiol 143 (2007) 1467–1483.
[7] K. Carroll, C. Gomez, L. Shapiro, Tubby proteins: the plot thickens, Nat. Rev. Mol. Cell Biol. 5 (2004) 55–63.
[8] H. Kong, et al., Patterns of gene duplication in the plant SKP1 gene family in angiosperms: evidence for multiple mechanisms of rapid gene birth, Plant J. 50 (2007) 873–885.
[9] S.B. Cannon, A. Mitra, A. Baumgarten, N.D. Young, G. May, The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*, BMC Plant Biol. 4 (2004) 10.
[10] The Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana, Nature 408 (2000) 796–815.
[11] International Rice Genome Sequencing Project, The map-based sequence of the rice genome, Nature 436 (2005) 793–800.
[12] J. Yu, et al., A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*), Science 296 (2002) 79–92.
[13] S.A. Goff, et al., A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*), Science 296 (2002) 92–100.
[14] G.A. Tuskan, et al., The genome of black cottonwood, *Populus trichocarpa* (Torr. and Gray), Science 313 (2006) 1596–1604.
[15] X. Li, et al., Genome-wide analysis of basic/helix–loop–helix transcription factor family in rice and *Arabidopsis*, Plant Physiol 141 (2006) 1167–1184.
[16] C.S. Goh, A.A. Bogan, M. Joachimiak, D. Walther, F.E. Cohen, Co-evolution of proteins with their interaction partners, J. Mol. Biol. 299 (2000) 283–293.
[17] I.K. Jordan, K.S. Makarova, J.L. Spouge, Y.I. Wolf, E.V. Koonin, Lineage-specific gene expansions in bacterial and archaeal genomes, Genome Res. 11 (2001) 555–565.
[18] O. Lespinet, Y.I. Wolf, E.V. Koonin, L. Aravind, The role of lineage-specific gene family expansion in the evolution of eukaryotes, Genome Res. 12 (2002) 1048–1059.
[19] Z. Yang, et al., Comparative study of SBP-box gene family in *Arabidopsis* and rice, Gene 407 (2008) 1–11.
[20] S. Zhang, et al., Evolutionary expansion, gene structure, and expression of the rice wall-associated kinase gene family, Plant Physiol. 139 (2005) 1107–1124.
[21] M. Jain, A.K. Tyagi, J.P. Khurana, Genome-wide analysis, evolutionary expansion, and expression of early auxin-responsive SAUR gene family in rice (*Oryza sativa*), Genomics 88 (2006) 360–371.
[22] J. Yu, et al., The genomes of *Oryza sativa*: a history of duplications, PLoS Biol. 3 (2005) e38.
[23] X. Wang, X. Shi, B. Hao, S. Ge, J. Luo, Duplication and DNA segmental loss in the rice genome: implications for diploidization, New Phytol. 165 (2005) 937–946.
[24] J. Raes, K. Vandepoele, C. Simillion, Y. Saeys, Y. Van de Peer, Investigating ancient duplication events in the *Arabidopsis* genome, J. Struct. Funct. Genomics 3 (2003) 117–129.
[25] C. Simillion, K. Vandepoele, M.C. Van Montagu, M. Zabeau, Y. Van de Peer, The hidden duplication past of *Arabidopsis thaliana*, Proc. Natl. Acad. Sci. USA 99 (2002) 13627–13632.
[26] L. Sterck, et al., EST data suggest that poplar is an ancient polyploidy, New Phytol. 167 (2005) 165–170.
[27] J.W. Thornton, R. DeSalle, Gene family evolution and homology: genomics meets phylogenetics, Annu. Rev. Genomics Hum. Genet. 1 (2000) 41–73.
[28] D. Lijavetzky, P. Carbonero, J. Vicente-Carbajosa, Genome-wide comparative phylogenetic analysis of the rice and *Arabidopsis* Dof gene families, BMC Evol. Biol. 3 (2003) 17.
[29] R.L. Tatusov, E.V. Koonin, D.J. Lipman, A genomic perspective on protein families, Science 278 (1997) 631–637.
[30] G. Blanc, K.H. Wolfe, Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution, Plant Cell 16 (2004) 1679–1691.
[31] H. Shan, et al., Patterns of gene duplication and functional diversification during the evolution of the AP1/SQUA subfamily of plant MADS-box genes, Mol. Phylogenet. Evol. 44 (2007) 26–41.
[32] X. He, J. Zhang, Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution, Genetics 169 (2005) 1157–1164.
[33] X. Yang, G.A. Tuskan, M.Z. 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.
[34] A. Devoto, et al., Molecular phylogeny and evolution of the plant-specific seven-transmembrane MLO family, J. Mol. Evol. 56 (2003) 77–88.
[35] F. Pazos, A. Valencia, Similarity of phylogenetic trees as indicator of protein–protein interaction, Protein Eng. 14 (2001) 609–614.
[36] Q. Liu, Q. Xue, Molecular phylogeny, evolution, and functional divergence of the LSD1-like gene family: inference from the rice genome, J. Mol. Evol. 64 (2007) 354–363.
[37] E.L. Sonnhammer, S.R. Eddy, R. Durbin, Pfam: a comprehensive database of protein domain families based on seed alignments, Proteins 28 (1997) 405–420.
[38] A. Aiyar, The use of CLUSTAL W and CLUSTAL X for multiple sequence alignment, Methods Mol. Biol. 132 (2000) 221–241.
[39] K. Tamura, J. Dudley, M. Nei, S. Kumar, MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0, Mol. Biol. Evol. 24 (2007) 1596–1599.
[40] T.L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, Proc. Int. Conf. Intell. Syst. Mol. Biol. 2 (1994) 28–36.
[41] C. Maher, L. Stein, D. Ware, Evolution of Arabidopsis microRNA families through duplication events, Genome Res 16 (2006) 510–519.
[42] P. Jaiswal, et al., Gramene: a bird's eye view of cereal genomes, Nucleic Acids Res. 34 (2006) D717–D723.
[43] J.M. Comeron, K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals, Bioinformatics 15 (1999) 763–764.
[44] S.H. Shiu, W.M. Karlowski, R. Pan, Y.H. Tzeng, K.F. Mayer, W.H. Li, Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice, Plant Cell 16 (2004) 1220–1234.
[45] G. Blanc, K.H. Wolfe, Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes, Plant Cell 16 (2004) 1667–1678.
[46] M. Lynch, J.S. Conery, The evolutionary fate and consequences of duplicate genes, Science 290 (2000) 1151–1155.
[47] J. Reinartz, et al., Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms, Brief Funct. Genomics Proteomics 1 (2002) 95–104.
[48] J. Ohlrogge, C. Benning, Unraveling plant metabolism by EST analysis, Curr. Opin. Plant Biol. 3 (2000) 224–228.
[49] C.H. Leseberg, A. Li, H. Kang, M. Duvall, L. Mao, Genome-wide analysis of the MADS-box gene family in *Populus trichocarpa*, Gene 378 (2006) 84–94.