BMC Genomics

RESEARCH

Open Access

Whole-genome and dispersed duplication, including transposed duplication, jointly advance the evolution of *TLP* genes in seven representative Poaceae lineages

Huilong Chen[1,2], Yingchao Zhang[1*] and Shuyan Feng[1]

Abstract

Background In the evolutionary study of gene families, exploring the duplication mechanisms of gene families helps researchers understand their evolutionary history. The tubby-like protein (TLP) family is essential for growth and development in plants and animals. Much research has been done on its function; however, limited information is available with regard to the evolution of the *TLP* gene family. Herein, we systematically investigated the evolution of *TLP* genes in seven representative Poaceae lineages.

Results Our research showed that the evolution of *TLP* genes was influenced not only by whole-genome duplication (WGD) and dispersed duplication (DSD) but also by transposed duplication (TRD), which has been neglected in previous research. For *TLP* family size, we found an evolutionary pattern of progressive shrinking in the grass family. Furthermore, the evolution of the *TLP* gene family was at least affected by evolutionary driving forces such as duplication, purifying selection, and base mutations.

Conclusions This study presents the first comprehensive evolutionary analysis of the *TLP* gene family in grasses. We demonstrated that the *TLP* gene family is also influenced by a transposed duplication mechanism. Several new insights into the evolution of the *TLP* gene family are presented. This work provides a good reference for studying gene evolution and the origin of duplication.

Keywords Transposed duplication, Codon bias, Selection pressure, Synteny network, Tubby-like protein, Evolution

Introduction

Genes can be duplicated through a variety of mechanisms, including whole-genome duplication (WGD), tandem duplication (TD), proximal duplication (PD), dispersed duplication (DSD), and transposed duplication (TRD). Exploring the different duplication mechanisms of gene families helps us understand the origin and evolution of the genes and provides unique insights. By identifying the types of duplication origins of cold resistance genes in plants, Song et al. found that cold resistance genes can originate from singletons, DSD, PD, TD, and/or WGD and that WGD and DSD were the major contributors to gene duplication, thus proposing the hypothesis that cold resistance genes were preferentially retained after polyploidization events [1]. Liu et al. found that both WGD and TD played an important role in the expansion of intronless genes in intron-poor subfamilies by identifying the type

*Correspondence:
Yingchao Zhang
17704716113@163.com
[1] College of Life Sciences, North China University of Science and Technology, Tangshan 063210, Hebei, China
[2] College of Grassland Science and Technology, China Agricultural University, Beijing 100193, China

of origin of duplication of intron-poor and intronless family genes in plants [2]. Nezamivand-Chegin et al. proposed that WGD or DSD types were the most frequent contributors to SPX expansion by identifying the duplication types of SPX family genes in plants [3]. Therefore, it is useful to perform an analysis of the type of duplication origin of genes in evolutionary studies of gene families.

These duplication types are recognized by the duplicate_gene_classifier program in the MCScanX package and are divided into five types: singleton, WGD, TD, PD, and DSD [4, 5]. MCScanX enforces the determination that the origin of duplication of a gene can belong to only one of the five duplication types and does not allow a gene to originate through different duplication mechanisms. However, the true meaning of a singleton, as mentioned above, is that it does not undergo duplication; i.e., it should not be called a type of duplication. There are five types of gene duplication mechanisms: WGD, TD, PD, DSD, and TRD [5, 6]. Therefore, an important duplication mode (TRD) was missing in the above studies. TRD often leads to the formation of pseudogenes, while other types of duplication lead to rapid expansion of the plant genome [7], resulting in severe functional redundancy and increased functional differentiation within plant gene families [7]. TRD can occur through DNA-based or RNA-based mechanisms. For example, in graminaceous plants, DNA transposons such as packmules (rice (*Oryza sativa* L.)) [8], helitrons (maize (*Zea mays* L.)) [9] and CACTA elements (sorghum (*Sorghum bicolor* L.)) [10] can relocate duplicated genes or gene fragments to new chromosomal locations. Although there are many methods and tools available to perform gene duplication origin analysis [4, 11–13], they do not identify all types of duplication origin due to algorithmic limitations. This may be the reason for the phenomenon of incomplete conclusions. Qiao et al. developed a pipeline named DupGen_finder that can identify all duplication types by optimizing the MCScanX algorithm [5]. This makes it possible to study the complete duplication origin of genes.

The *TLP* family gene was first discovered in the mouse genome [14] and widely exists in animal and plant genomes [15]. TLP protein has a tubby domain (PF01167) of approximately 270 aa at the *C*-terminus, with other possible distinct domains at the *N*-terminus [16–18]. In 1999, Boggon et al. published the crystal structure of the tubby domain, which consists of a hydrophobic α-helix surrounded by a β-barrel structure containing 12 sheets of inverted β-folds. The hydrophobic alpha helix is positioned at the *C*-terminus of the TLP protein [19]. In contrast to the diversity of *N*-terminal structures in animals, the *N*-terminal end of TLP proteins in plants often contains a conserved F-box domain [20].

TLP proteins are essential for growth and development in plants and animals, and deletion of TLP family proteins often leads to changes in the phenotypic characteristics of animals and even to serious diseases. For example, mice deficient in the tubby gene show symptoms such as obesity, blindness, and deafness [14, 21–23]. Mutations in the *TULP1* gene in humans cause an autosomal recessive form of retinitis pigmentosa [24–26]. In a TLP study in *Arabidopsis* (*Arabidopsis thaliana* (L.) Heynh.), Lai et al. found that overexpression of the *AtTLP9* gene in transgenic plants enhanced their sensitivity to ABA [16]. In rice TLP studies, the OsTLP2 protein was found to bind to the PRE4 *cis*-element of the promoter region of the *OsWRKY13* transcription factor, which is induced by pathogens and regulates resistance to bacteria and fungi in rice [27]. Furthermore, many studies on plant *TLP* genes have shown that these genes are involved in plant responses to biotic and abiotic stresses, enhancing plant resistance to a variety of stresses [17, 28, 29].

Previous research has shown that TLPs evolved from an ancestor of the scramblase-like protein family [30]. The sequences corresponding to the tubby domains are conserved in both unicellular and multicellular organisms, but the *N* termini of TLPs are distinct [31]. Intriguingly, plant TLPs have conserved F-box domains in the *N*-terminal sequences, whereas they are highly divergent in animals [27, 28]. In plants, there are more members of the TLP family than in animals [20]. For example, five TLP family members have been identified in vertebrates [22], while 11 family members have been identified in the genome of *Arabidopsis* [16]. Moreover, it has been observed that up to 80% of genes in *Arabidopsis* are the result of lineage-specific expansion [32]. The sequence and architectural similarity of the *TLP* gene family in rice suggests that the rice TLP family may have originated from the same ancestral gene [27]. In cotton, 28 of the 29 paralogous gene pairs have undergone purifying selection [33]. In a comprehensive study of the *TLP* gene family in *Arabidopsis*, rice, and poplar (*Populus trichocarpa* (Torr. and Gray)), *Ka*/*Ks* ratios indicated that most of the duplicated gene pairs experienced positive selection, while some of the remaining gene pairs experienced neutral selection [34]. By using the *Arabidopsis TLP* gene family as a reference, it was found in Brassica that *TLP* genes in *Brassica napus* are not directly amplified compared to those in the diploid parents. Instead, indirect amplification of the *TLP* gene family occurs in the two diploid parents [35].

In research on the Brassica *TLP* gene family, Wang et al. identified five origins of *TLP* duplication using the MCScanX program and found that *TLP* genes in Brassica

originated only from WGD and DSD [35]. Based on the above analysis, an important duplication mode (TRD) was missing in this study. Therefore, we hypothesize that the origin of duplication of *TLP* genes may include TRD. Because the *TLP* gene family in the grass family, belonging to the monocots, has not been studied, we selected seven representative grass species for duplication type detection of the *TLP* gene family. Moreover, since the evolutionary trajectory of the grass *TLP* gene family is not yet known, we aimed to fill this gap and provide unique insights into the evolution of the Gramineae *TLP* gene family. We also performed evolutionary analysis of *TLP* gene families in the seven examined grass species using our previously established gene family analysis pipeline [36, 37]. The purpose of this study was to determine whether TRD facilitated the evolution of the *TLP* gene family and to explore the evolutionary footprint of the *TLP* gene family in grass species.

Results

The duplication origin modes of *TLP* genes are WGD, DSD, and TRD

To investigate whether the duplication origin types of *TLP* genes include TRD, we used the DupGen_finder pipeline to identify all possible duplication types: WGD, TD, PD, DSD, and TRD. WGD and DSD were detected, and no TD and PD were detected (Fig. 1A, Additional file 1: Table S1), which is consistent with the findings of Wang et al.. However, our results showed that the evolution of *TLP* genes was also affected by TRD, as we hypothesized (Fig. 1A, Additional file 1: Table S1). Furthermore, we found that the number of gene pairs with DSD was the highest for the *TLP* gene family in all grass species studied. The number of TRD pairs was greater than the number of WGD pairs in maize and barley (*Hordeum vulgare* L.), equal to the number of WGD pairs in *Brachypodium* (*Brachypodium distachyon* L.), and less than the number of WGD pairs in the other species (sorghum, foxtail millet (*Setaria italica* L.), green foxtail (*Setaria viridis* L.), and rice) (Fig. 1, Additional file 1: Table S1).

Further calculations of *ka*/*ks* for the three duplicated gene pairs showed that the three duplicates of the seven species generally underwent purifying selection ($ka/ks < 1$) (Fig. 1B, Additional file 1: Table S1). In particular, some TRD gene pairs in foxtail millet, green foxtail, and rice were subject to positive selection ($ka/ks > 1$). This may be related to the active nature of transposons. The distribution of *Ks* values for the three duplicated gene pairs showed no obvious regularity in the timing of these three duplications (Fig. 2, Additional file 1: Table S1). In general, WGD occurred later than TRD and DSD, with DSD occurring somewhat earlier. Of course, the small amount of current data limits the generality of this conclusion.

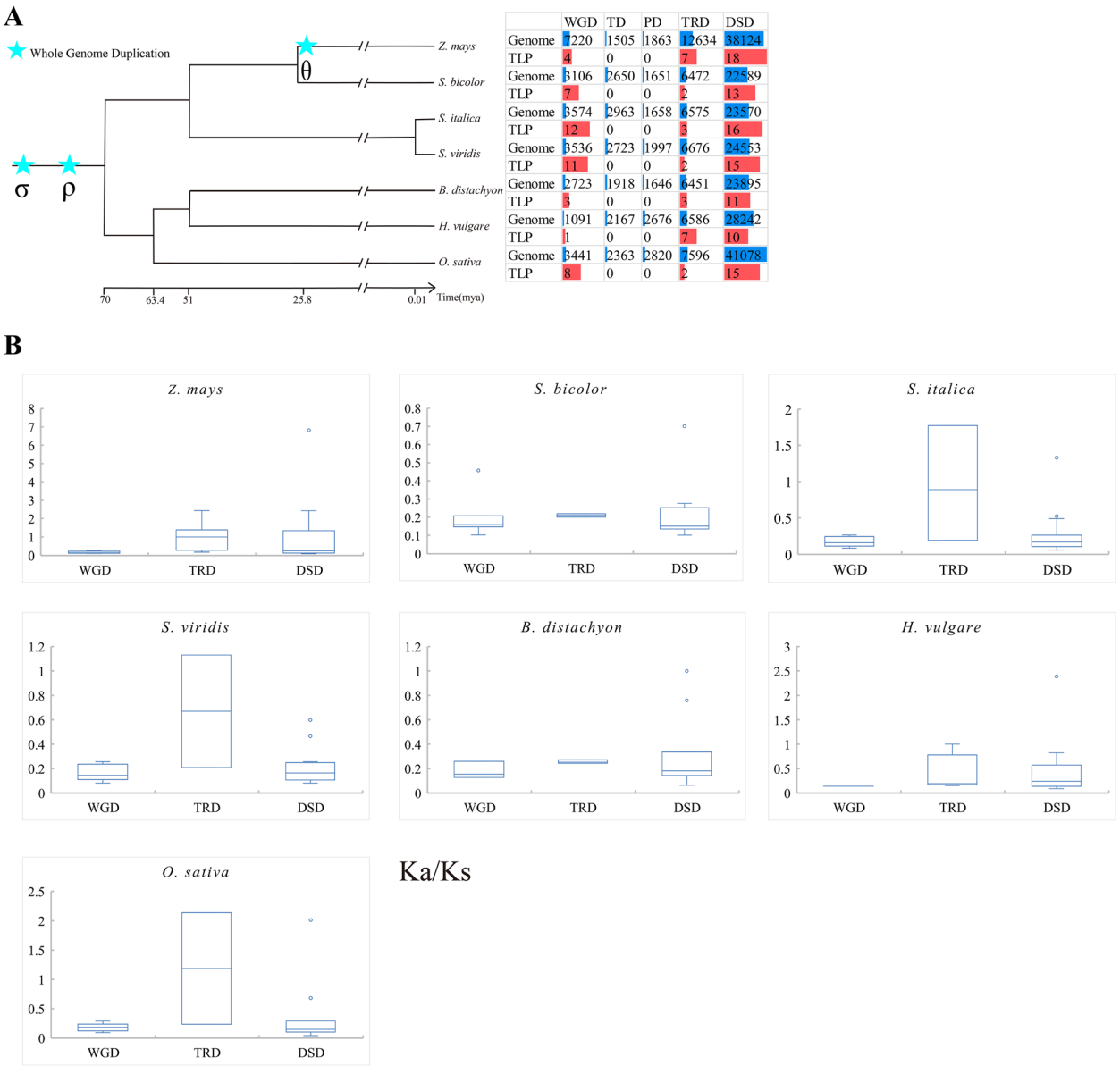
Phylogenetic analysis suggests the phenomenon of duplication and loss

A total of 97 *TLP* genes were identified, including 15 in maize, 13 in sorghum, 16 in foxtail millet, 15 in green foxtail, 12 in *Brachypodium*, 11 in barley, and 15 in rice (Fig. 3, Additional file 2: Table S2). Based on Wang et al.'s research and the topology of the tree. The phylogenetic tree of grass TLPs could be divided into six main groups, Group A, Group B, Group C, Group D, Group E, and Group F. Furthermore, Group A was classified as an intron-poor clade, and Groups B-F were classified as intron-rich clades (Additional file 3: Fig. S1, Additional file 4: Table S3).

We observed significant differences between these two clades, as the *TLP* genes in the intron-rich clade maintained a quantitative distribution consistent with the species tree of life, mostly accompanied by the loss of a duplicated maize gene after polyploidy. For example, a copy of *Zm2G472945_T01* was deleted in Group F. However, a copy of *Zm2G129288_T01*, *Zm5G871407_T01*, was retained in Group E. In addition, *Os03g22800.1* showed a duplicate gene, *Os03g22655.1*, but a barley *TLP* orthologue was lost in Group B. Group A in the intron-poor clade was the youngest branch with more divisions. The above analyses suggest that the *TLP* gene family has undergone different degrees of duplication and loss in different grass species.

Convergence in family size by shrinkage

Based on the phylogenetic tree, it could be inferred that many independent gene gains and losses occurred during different stages of grass evolution (Fig. 4A). We found that the *TLP* gene family shrank after the WGD event common to grasses ~ 100 million years ago (Fig. 4B). The analysis of gene increases and decreases based on the gene phylogeny showed that after the grass-wide WGD event, there should have been at least 28 *TLP* genes in the grass common ancestor (Fig. 4B). Starting with 28 ancestral genes, the family size of each sublineage then gradually decreased over time to result in the family size of the extant species. For instance, the common ancestor of *Brachypodium*, barley, and rice had 22 *TLP* genes after having gained one gene and lost seven genes. After that, rice had 15 *TLP* genes, gaining zero genes and losing seven genes (Fig. 4B, C). Similar decreases in family size were found in the other studied grasses (Fig. 4B, C). Thus, the grass species that we studied exhibit a "consistent shrinking" evolutionary pattern.



	WGD	TD	PD	TRD	DSD
Genome	7220	1505	1863	12634	38124
TLP	4	0	0	7	18
Genome	3106	2650	1651	6472	22589
TLP	7	0	0	2	13
Genome	3574	2963	1658	6575	23570
TLP	12	0	0	3	16
Genome	3536	2723	1997	6676	24553
TLP	11	0	0	2	15
Genome	2723	1918	1646	6451	23895
TLP	3	0	0	3	11
Genome	1091	2167	2676	6586	28242
TLP	1	0	0	7	10
Genome	3441	2363	2820	7596	41078
TLP	8	0	0	2	15

Fig. 1 The number and *Ka*/*Ks* ratio distributions of gene pairs derived from different modes of duplication in seven grass species. **A** The number of gene pairs derived from different modes of duplication in seven grass species. Whole-genome duplication (WGD), tandem duplication (TD), proximal duplication (PD), transposed duplication (TRD), and dispersed duplication (DSD). The WGDs that occurred on different branches are labelled. The tree of life and polyploidy information of grass species came from previous reports. **B** The nonsynonymous substitution rate (*ka*)/synonymous substitution rate (*ks*) (*Ka*/*Ks*) ratio distributions of gene pairs derived from different modes of duplication in seven grass species. WGD: whole-genome duplication, TRD: transposed duplication, DSD: dispersed duplication

Synteny network and phylogenomic analyses of TLPs reveal duplication events

Phylogenomic synteny network analyses can identify the different types of duplicates from WGDs or TDs on the basis of phylogenies [38, 39]. To further understand the duplication history of TLPs, we performed synteny network and phylogenomic analyses of TLPs. We found that the synteny network of grass TLPs contained five clusters, named Cluster 1, Cluster 2, Cluster 3, Cluster 4, and Cluster 5, with sizes of 41, 7, 8, 12, and 22, respectively (Fig. 5A, B, Additional file 5: Table S4). Duplications and losses of several genes were found; for example, one maize *TLP* gene from Cluster 3 was duplicated during the theta (θ) WGD event, and two barley *TLP* genes from Cluster 4 were lost.

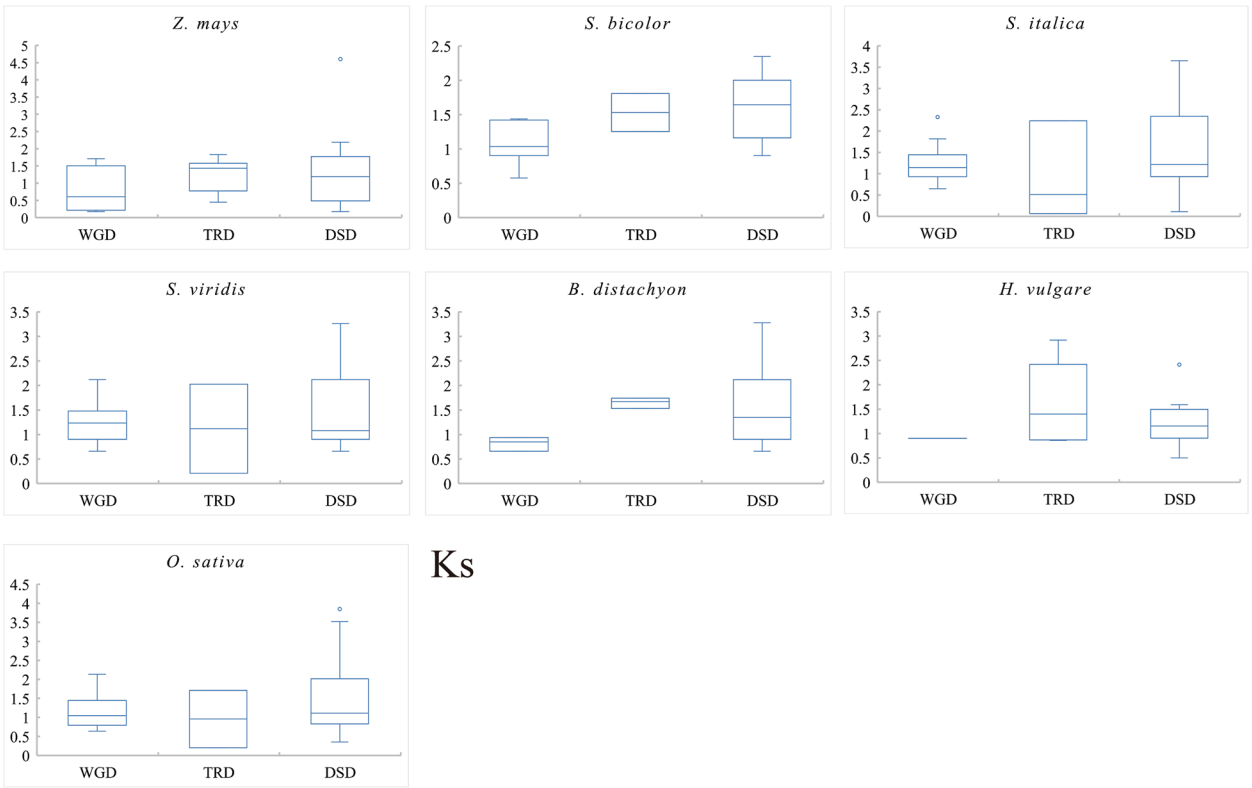


Fig. 2 *Ks* distributions of gene pairs derived from different modes of duplication in seven grass species. Whole-genome duplication (WGD), transposed duplication (TRD), dispersed duplication (DSD)

According to the phylogenetic tree of the *TLP* genes, the genes of Clusters 1, 2, and 4 belong to the intron-rich clade, and the genes of Clusters 3 and 5 belong to the intron-poor clade (Figs. 3 and 5B, C). We found a high degree of consistency between the gene sets in each cluster and the phylogenetic tree. For example, Cluster 2 genes were all contained in Group F, and Cluster 1 genes were contained in adjacent Groups E, D, and C (Fig. 5B, C). The above phenomenon implies that the evolutionary pattern of grass *TLP* genes is one of conservation.

Homology clustering reflects species proximity and strong purifying selection

The identification of homologous genes in grasses showed that foxtail millet had the most orthologous gene pairs with the *TLP* gene family of green foxtail, and maize had the fewest orthologous gene pairs with the *TLP* gene families of barley and rice (Fig. 6A, Additional file 6: Table S5). This reflects the fact that the more closely related to each other the species are, the more *TLP* homologous gene pairs there are. The MCL homology clustering of grass *TLP* genes yielded 16 classes. Among them, there were four clusters containing genes from each species, and they showed single-copy gene patterns (Fig. 6A, Additional file 7: Table S6). Phylogenetic tree reconstruction and selection pressure on these four single-copy gene clusters showed that a total of 41 branches (95.35%, 43 branches in total) were subject to strong purifying selection (the ratio of the nonsynonymous to synonymous distances (ω) ranged from 0.0001 to 0.399558) (Fig. 6B), suggesting that grass *TLP* gene evolution has involved strong purifying selection.

Codon bias is weak and similar

To investigate whether codon usage bias has contributed to *TLP* gene evolution, we performed codon bias analysis on the *TLP* genes of each species using the CodonW program. The value of ENC denotes the number of effective codons and varies from 20 to 61; an ENC value less than 35 indicates strong codon bias [40, 41]. Our results show that the ENC values for the *TLP* gene family in each species are greater than 39 (Fig. 7A, Additional file 8: Tables S7, Additional file 9: S8), suggesting that the codon bias of the *TLP* genes is very weak.

ENC-GC3s plots were used to evaluate the influence of mutation selection on codon bias [40, 42]. Most of the *TLP* genes of the seven species were below and away from the standard curve (Fig. 7B), indicating a large

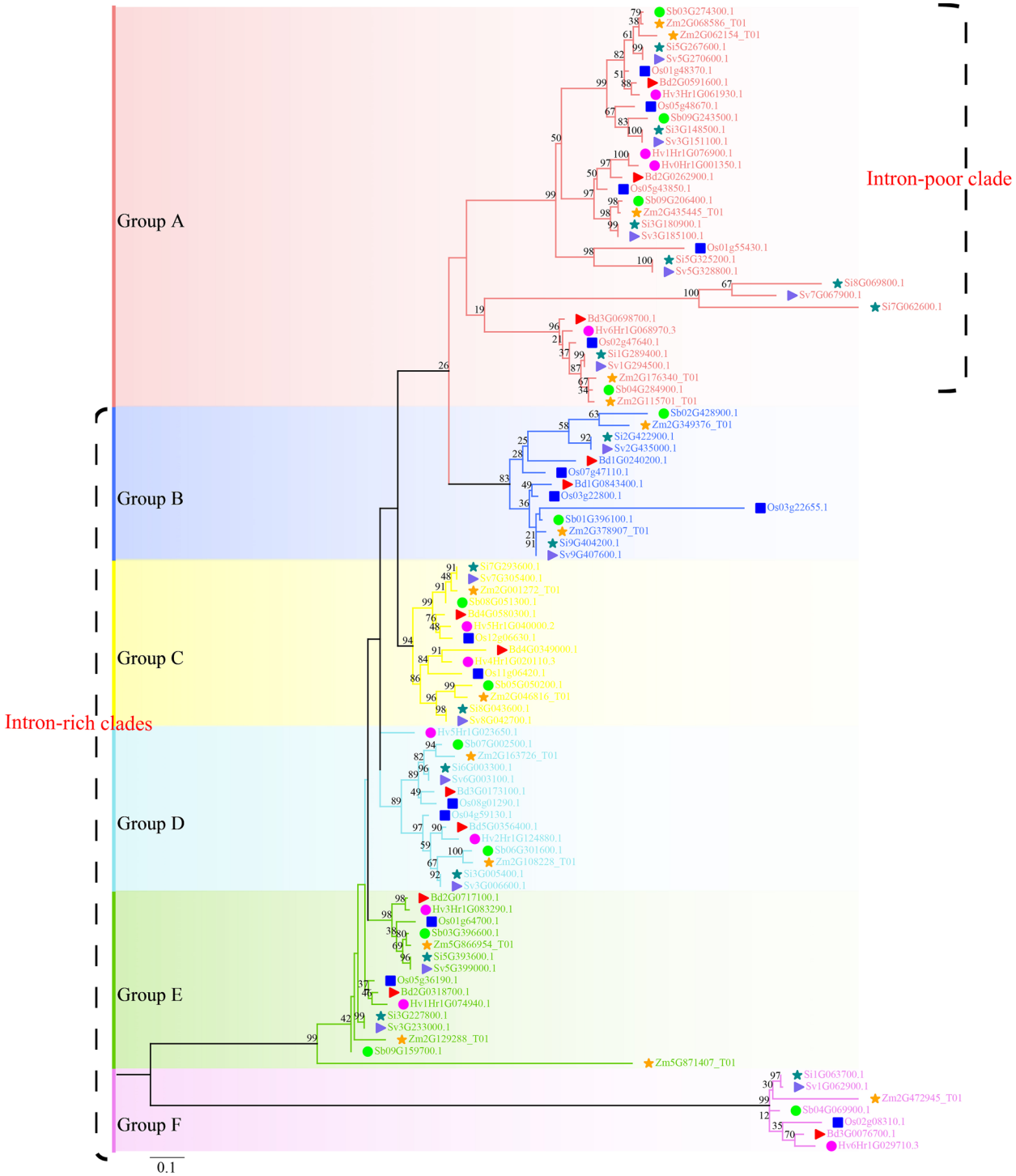


Fig. 3 Phylogenetic tree and classification of the TLPs of seven grass species. Here, gene IDs show their respective origin: Zm for maize, Sb for sorghum, Si for foxtail millet, Sv for green foxtail, Bd for *Brachypodium*, Hv for barley, and Os for rice. We used shapes and colours to distinguish different species, with orange stars, green circles, dark green stars, mauve triangles, red triangles, purple circles, and blue squares representing the tubby-like protein (*TLP*) genes in maize, sorghum, foxtail millet, green foxtail, *Brachypodium*, barley, and rice, respectively

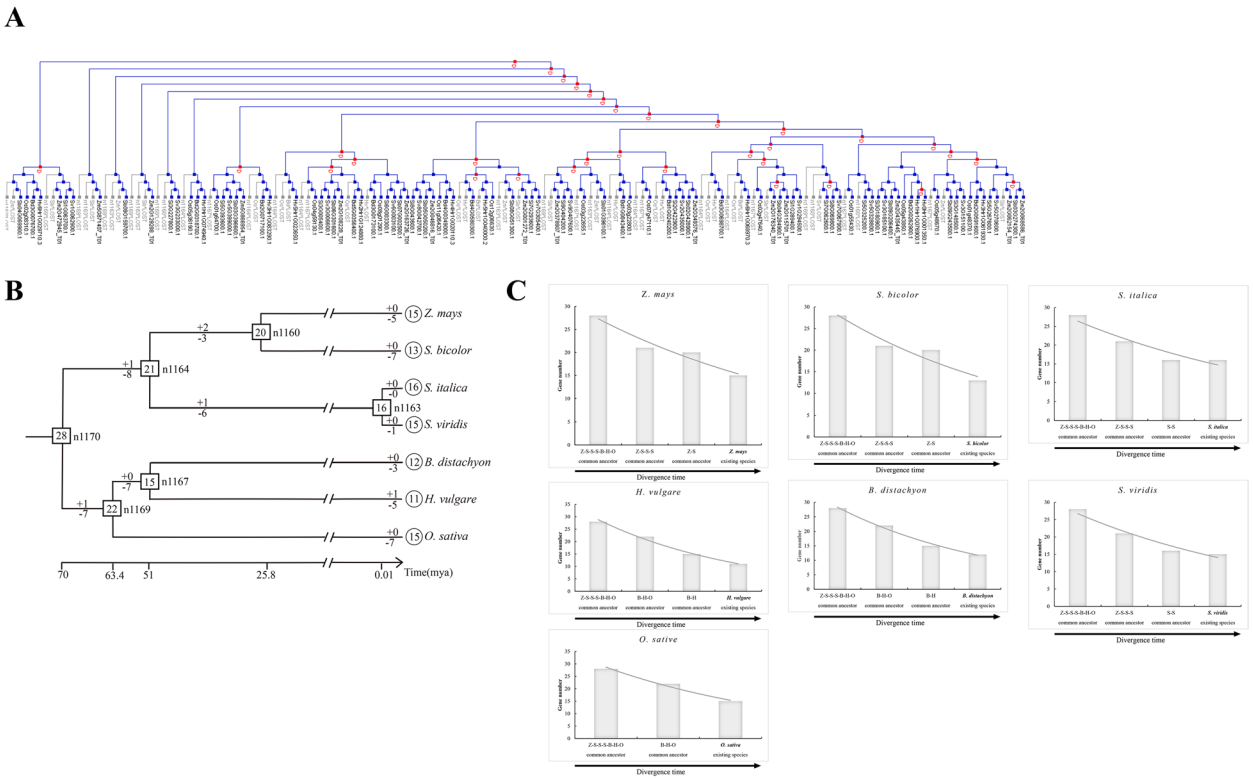


Fig. 4 Gain and loss of *TLP* genes in seven grass species. **A** The reconciliation between the species tree and gene tree along with the confirmation of the gene loss/duplication scenario was performed using Notung. The species tree is shown in Fig. 2B. The gene tree is shown in Fig. 1. Red "D"s at branching points indicate predicted gene duplications. Grey branches indicate gene losses. **B** Schematic diagram of gain and loss of tubby-like protein (*TLP*) genes in seven grass species. The numbers in the rectangles and circles represent the number of *TLP* genes in ancestors and existing species. The + and − signs represent the gain and loss of genes, respectively. **C** Evolutionary patterns of *TLP* genes in seven grass species. Z-S-S-S-B-H-O indicates the common ancestor of all seven grass species; Z-S-–S-S indicates the common ancestor of maize, sorghum, foxtail millet, and green foxtail; Z-S indicates the common ancestor of maize and sorghum; S-S indicates the common ancestor of foxtail millet and green foxtail; B-H-O indicates the common ancestor of *Brachypodium*, barley, and rice; and B-H indicates the common ancestor of *Brachypodium* and barley

difference between their actual and expected ENC values, suggesting that base mutations are not the main factor influencing their codon bias but that they may also be influenced by natural selection and other factors. A small number of genes were located at and near the standard curve, indicating that their actual ENC values were close to the expected values and suggesting that their codon bias is influenced by base mutations. Therefore, in conjunction with the previous selection pressure analysis, these findings indicate that base mutation and selection pressure jointly promoted the codon bias of *TLP* genes in the seven grass species, and purifying selection pressure may have a greater impact.

Regarding optimal codons, maize has 17 optimal codons, 1 of which ends in A and 16 of which end in G/C; sorghum has 13 optimal codons, 5 of which end in A/U and 8 of which end in G/C; foxtail millet has 18 optimal codons, 1 of which ends in A and 8 of which end in G/C; green foxtail has 14 optimal codons, 1 of which ends in A and 13 of which end in G/C; *Brachypodium* has 12 optimal codons, 0 of which end in A/U and 12 of which end in G/C; barley has 13 optimal codons, 1 of which ends in A and 12 of which end in G/C; and rice has 12 optimal codons, 0 of which end in A/U and 12 of which end in G/C (Fig. 8, Additional file 10: Table S9). These results imply that the optimal codon of the *TLP* gene family in all seven species prefers to end in C or G.

Venn network analysis showed that the seven grass *TLP* gene families share three optimal codons (AGG, AAG, and CAG). Maize shares a single optimal codon (GUC), sorghum shares four optimal codons (CAU, CCA, ACA, and GCU), grain shares three optimal codons (GCG, CAC, and ACG), barley shares one optimal codon (UGA) and rice shares one optimal codon (GGG) (Fig. 8, Additional file 11: Table S10). In summary, all results indicate that the codon biases of the seven species are nearly identical.

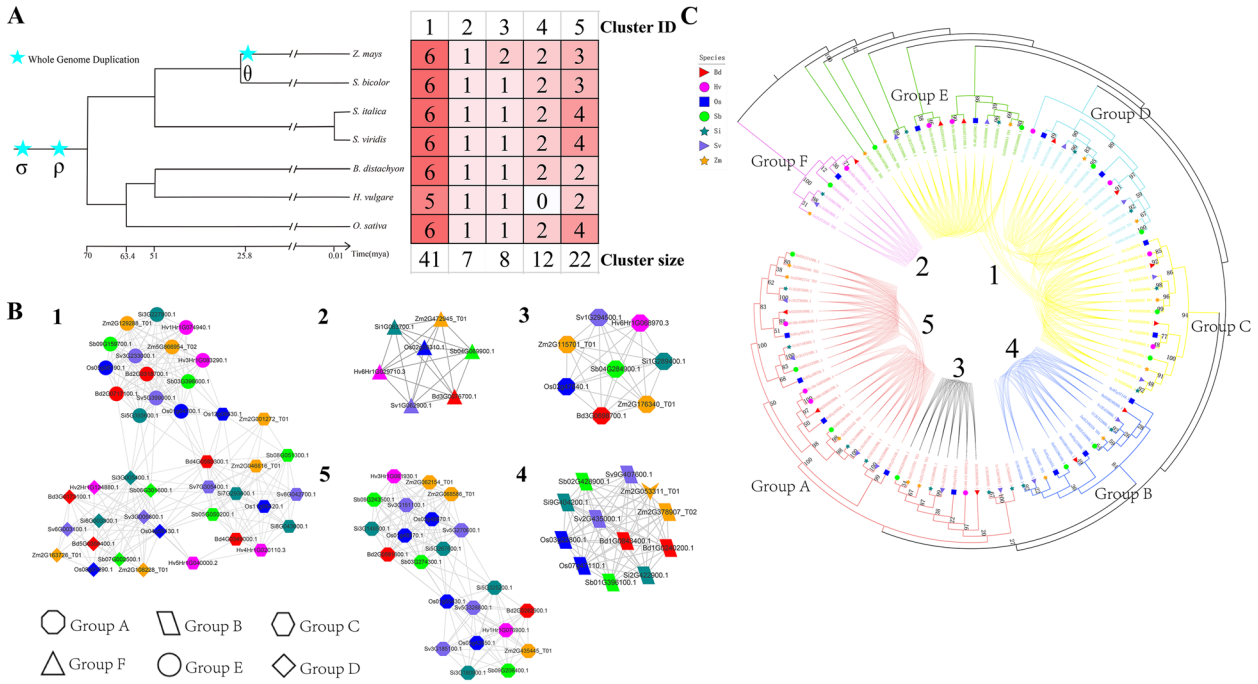


Fig. 5 Synteny network clusters and phylogenetic profiling of the *TLP* gene families in seven grass species. **A** Species composition for each of the five network communities. Red-coloured cells depict the presence of tubby-like protein (*TLP*) syntelogs (syntenic homologous genes) in the different species. The five network communities were identified using CFinder at k = 3. The cluster ID and size are indicated at the top and bottom, respectively. The tree of life and polyploidy information of grass species came from previous reports. **B** Detailed visualization for each of the TLP synteny network communities. Nodes in different colours represent different grass species, and the different node shapes represent the different groupings (Groups A to F) belonging to the phylogenetic tree. The V shape indicates that the gene does not belong to the *TLP* gene family. **C** Maximum-likelihood gene tree for the *TLP* gene family and syntenic relationships between the genes. Each connecting line located inside the inverted circular gene tree (implemented in iTOL) indicates a syntenic relationship between two *TLP* genes (syntelogs). This phylogenetic tree is consistent with Fig. 1, and the numbers 1–5 are cluster IDs, which are consistent with **A**, **B**

Discussion

Five mechanisms for the origin of gene duplication

MCScanX is often used to identify the type of origin of gene duplication in the evolutionary analysis of gene families [1–3, 35, 41, 43–53]. However, the duplication types identified by the duplicate_gene_classifier program of MCScanX lack the TRD type, and it is enforced that a gene can originate from only one duplication type. Therefore, caveats and limitations need to be recognized, presented, and discussed when interpreting the results from MCScanX.

Genes can be duplicated through a variety of mechanisms, except for WGD, which are collectively referred to as small-scale duplication/single gene duplication. TD copies are consecutive in the genome, and copies from PD are close to each other but separated by several genes. These two patterns of gene duplication are thought to arise through unequal crossing over or localized transposon activities. DSD copies are not contiguous within genomes and homologous chromosome segments. Distant single-gene transposition can explain the widespread dispersed duplication within and among genomes. TRD may occur through DNA-based or RNA-based mechanisms. DNA-based mechanisms occur by relocating the copied gene or gene fragment to a new chromosomal locus via DNA transposons. RNA-based transposed duplication works by reverse transcription of spliced messenger RNA to produce a single-exon retrocopy from a multi-exon parental gene. The new retrogene is deposited in a new chromosomal environment with new (i.e., nonancestral) neighbouring genes [5, 6, 54]. Therefore, the identification of transposed duplication requires the assistance of a reference genome. Based on this principle, Qiao et al. developed an analytical pipeline (DupGen_finder) to address the above issues by optimizing the MCScanX algorithm, opening up avenues for researchers to comprehensively analyse patterns of gene duplication origins.

The evolution of *TLP* genes was influenced not only by whole-genome duplication and dispersed duplication but also by transposed duplication

Previous studies have shown that *TLP* family genes in Brassica originated only from WGD and DSD [35].

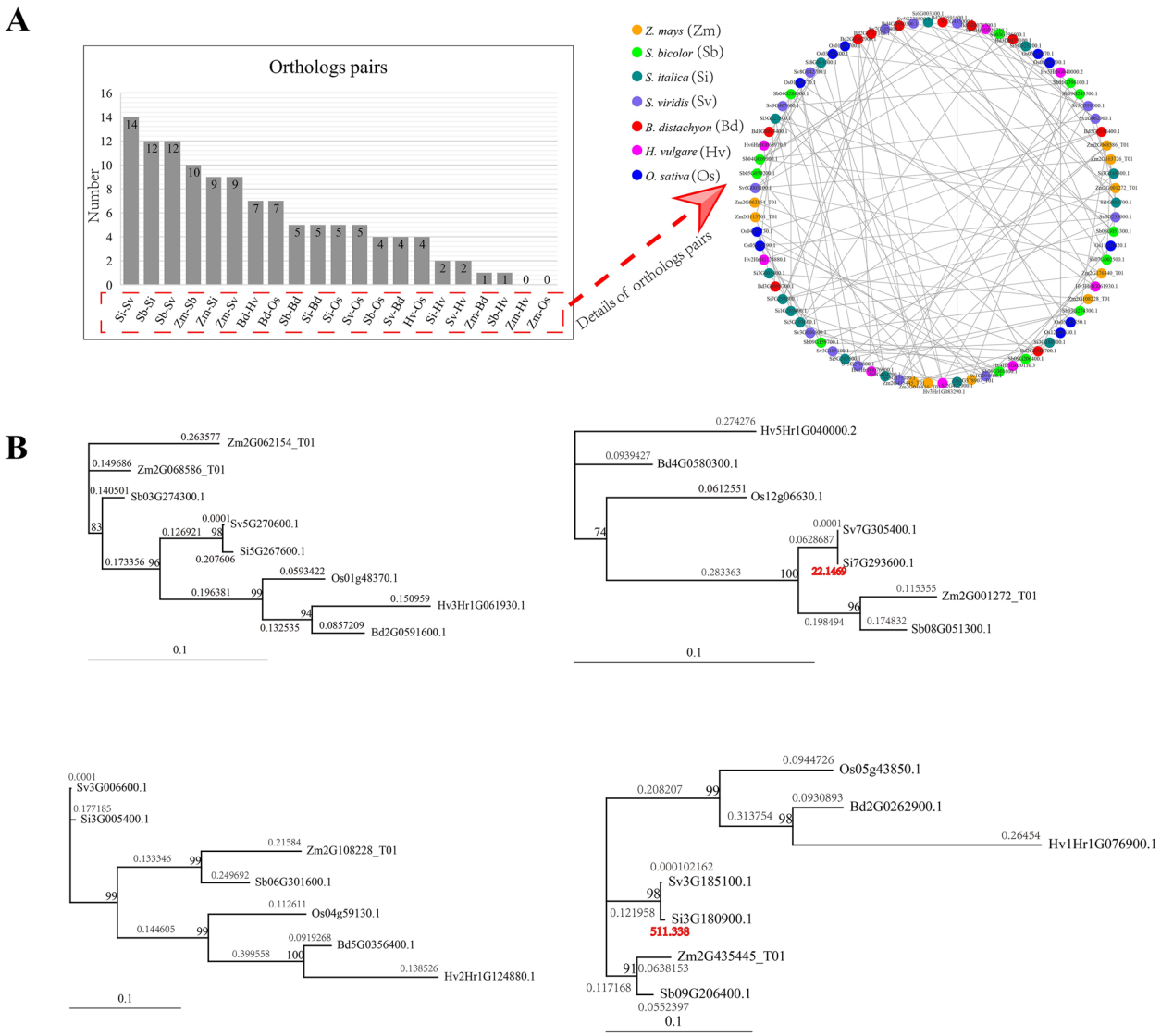


Fig. 6 Orthologous network and selection pressure analysis of *TLP* genes in seven grass species. **A** Quantitative distribution of tubby-like protein (*TLP*) orthologous gene pairs in seven grass species. Nodes in different colours represent different grass species. **B** Maximum likelihood gene tree and selection pressure of single-copy genes clustered by the Markov cluster (MCL) algorithm. A nonsynonymous substitution rate (*ka*)/synonymous substitution rate (*ks*) (*ka*/*ks*) less than 1 indicates purifying selection, and a *ka*/*ks* greater than 1 indicates positive selection. The range for all values of *ka*/*ks* less than 1 is 0.0001 to 0.399558, and values of *ka*/*ks* greater than 1 are highlighted in red

Here, we perform a series of evolutionary analyses on the *TLP* gene family in grasses. The *TLP* genes were found to have originated not only from WGD and DSD but also from TRD, which has been neglected. Previous studies on model plants have shown that WGD and TD contribute most to genetic redundancy, while other duplication modes contribute more to evolutionary novelty. Among them, inferred transposon-mediated gene duplication tends to reduce gene expression levels [6]. We found that *TLP* genes are subject to TRD, and therefore, this may affect *TLP* gene expression. However, this requires complex experiments for verification.

A plant gene family is a group of genes with related functions that arise from a single copy of an ancestral gene source through gene duplication and retain similar sequences and structures. Based on the extent of duplication, the size of the duplication region and the impact of transposons, gene families can be associated with duplication events such as TD, WGD, or TRD. The direct result of gene duplication may be functional differentiation of genes, including subfunctionalization,

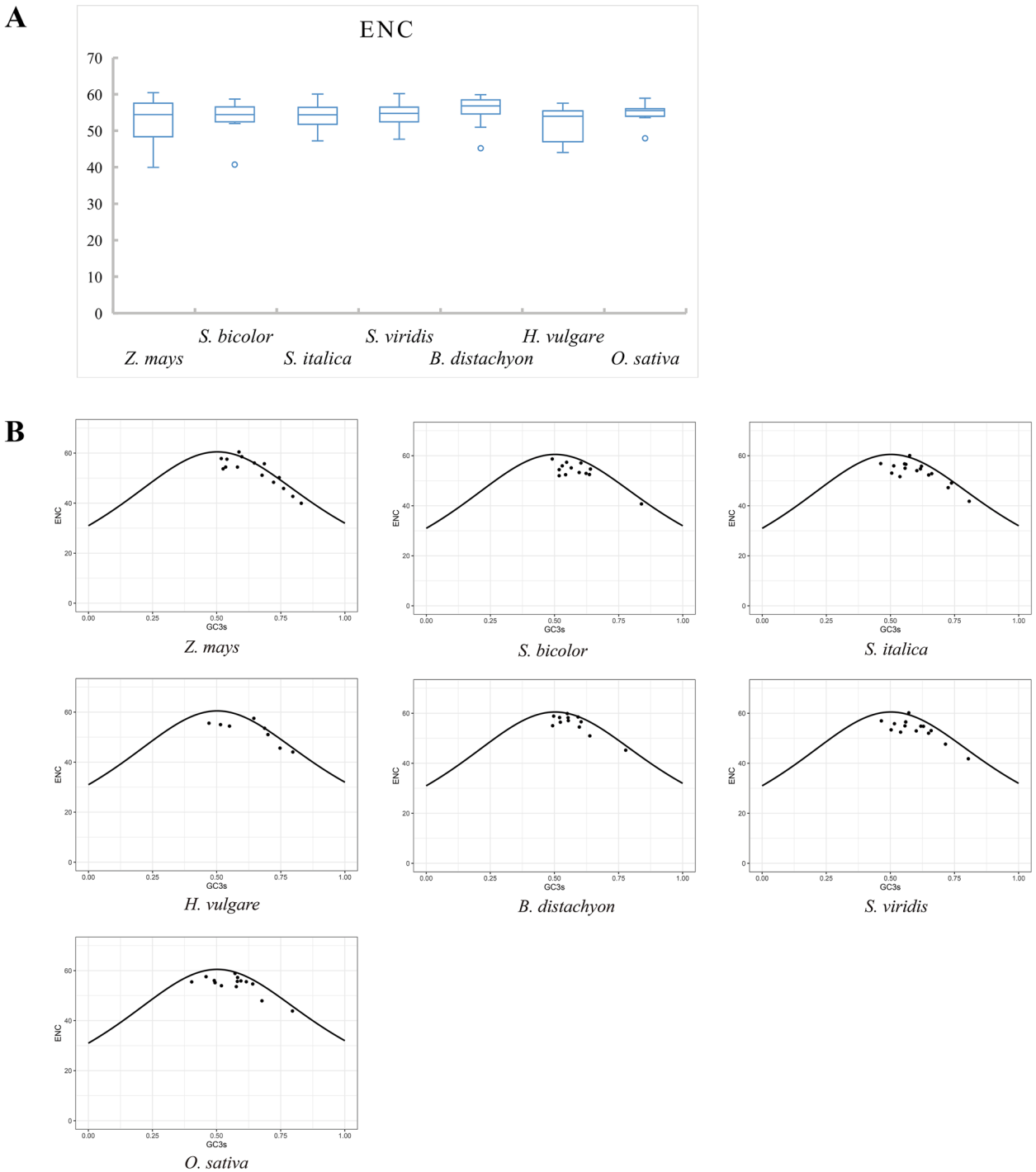


Fig. 7 Codon bias and ENC-GC3s plots of seven grass *TLP* gene families. **A** Box plot of ENC values for seven grass tubby-like protein (*TLP*) gene families. ENC indicates the number of valid codons. The value of ENC ranges from 20 to 61. An ENC value less than 35 indicates strong codon bias. **B** ENC-GC3s plots of seven grass *TLP* gene families. The ENC-GC3s plot was used to evaluate the influence of mutation selection on codon bias. GC3s indicates the GC content at the third codon site. The horizontal axis represents the GC3s value, and the vertical axis represents the ENC value

neofunctionalization, pseudogenization, and concerted evolution [55]. Evolutionary studies of genes can provide important clues to explain the functional differentiation of genes. Consequently, for the evolutionary analysis of a gene family, it is essential to parse the history of duplication experienced by the family. This is the main reason

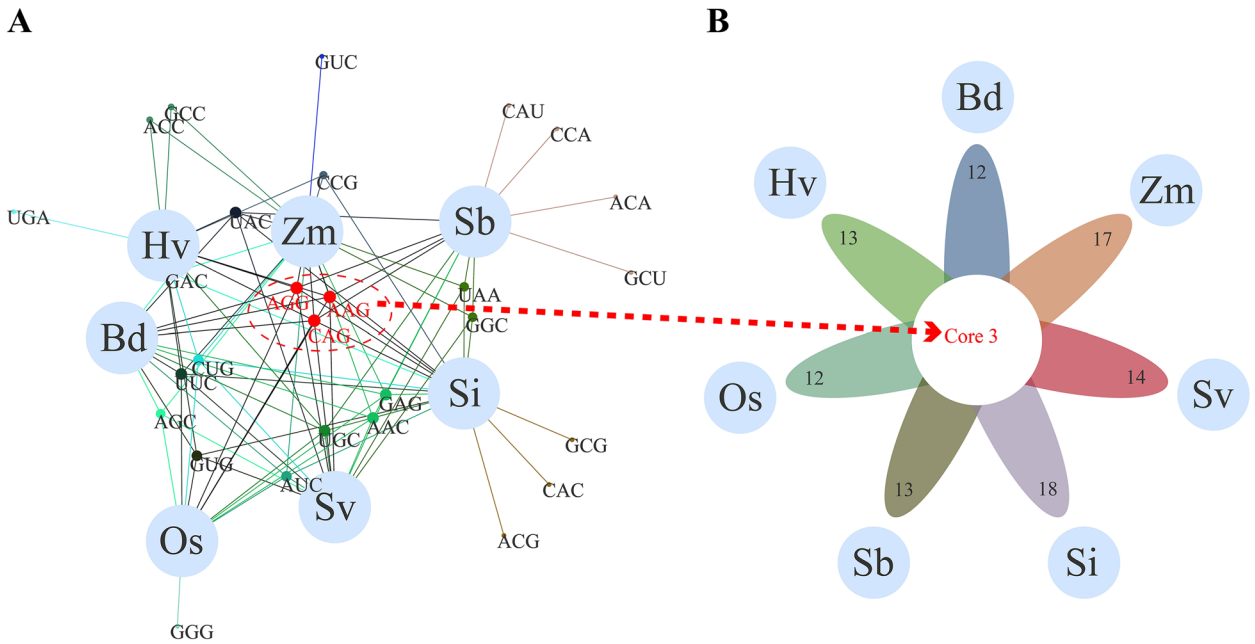


Fig. 8 Optimal codon analysis of seven grass *TLP* gene families. **A** Venn network map of the optimal codons of the seven grass tubby-like protein (*TLP*) gene families drawn by Evenn. The differently coloured lines represent the optimal codons shared by different species combinations. Optimal codons shared by the seven grass *TLP* gene families are highlighted in red. Zm for maize, Sb for sorghum, Si for foxtail millet, Sv for green foxtail, Bd for *Brachypodium*, Hv for barley, and Os for rice. **B** A flower plot of the optimal codons of the seven grass *TLP* gene families. Zm for maize, Sb for sorghum, Si for foxtail millet, Sv for green foxtail, Bd for *Brachypodium*, Hv for barley, and Os for rice

why in recent years, whenever the evolution of a gene family was considered, the type of origin of gene duplication was analysed. Herein, our study successfully demonstrates a complete gene family duplication origin analysis, and our results provide new insights into the evolution of the *TLP* gene family.

The "consistent shrinking" evolutionary pattern for the *TLP* gene family in grasses

A recent large-scale survey of plant gene family evolution showed that gene duplication and gene loss occurred in almost all gene families during plant evolution [7]. The same family evolutionary history was found in our study. Furthermore, we found that the grass *TLP* gene family underwent more loss than duplication, achieving similar gene numbers through constant shrinkage (Fig. 4). This evolutionary pattern of the *TLP* gene family may be beneficial to grasses, but the exact benefits remain to be explored.

Genes generate functional innovation through multiple duplication mechanisms and contribute to the adaptive evolution of species [55–58]. Nevertheless, plants often do not need very many copies of genes, as more copies may complicate the regulatory pathways and more homologues, mutated or intact, can induce conflicts and lead to plant death [37]. Many studies have shown that there is significant loss of new genes as a result of duplication and that the rate of gene loss tends to be inconsistent from species to species [59–61]. In our study, the synteny network (Fig. 5) and the homology network (Fig. 6) showed the loss of gene copies caused by WGD.

Base mutation and selection pressure jointly promoted the codon bias of *TLP* gene families

The drivers of genetic evolution, such as natural selection and base mutations, are diverse. Our study shows that the evolution of *TLP* genes mainly involved purifying selection (Figs. 1B and 6B), which is consistent with the evolutionary pattern of most genes [36, 41–43, 56, 62]. Codon usage bias has been hypothesized by some to have contributed to adaptive gene evolution [63]. Our codon bias analysis of *TLP* showed that base mutation and selection pressure jointly promoted the codon bias of grass *TLP* genes, and purifying selection pressure may have had a greater impact. Moreover, we found that the codon biases of the *TLP* gene family were consistent across the seven grass species. Optimal codons usually have high expression levels and can therefore be used as one of the codon bias parameters, providing a basis for codon modification during later transgenesis [41]. Our study showed that the optimal codon numbers of the seven grass *TLP* gene families ranged from 12 to 18, and these small numbers reflected the effects of purifying selection and mutational

pressure [64]. Moreover, previous research findings have shown that GC content elevation results in codon usage bias [56, 65]. The majority of the optimal codons of the grass *TLP* family contain GC bases, which may be evidence that codon usage bias is related to GC content.

Conclusions

In summary, by selecting the grass *TLP* gene family for duplication origin type analysis, we supported our conjecture that *TLP* genes originated from TRD in addition to the previously reported WGD and DSD. Moreover, our research shows that the evolution of the *TLP* gene family is at least affected by forces such as duplication, natural selection, and base mutations. We hope that our work can provide a reference for complete studies of gene duplication patterns and advance our understanding of the evolution of the *TLP* gene family.

Materials and methods

Collection of data and identification of *TLP* genes

Based on previous studies [37, 58], rice, maize, sorghum, foxtail millet, green foxtail, *Brachypodium*, and barley can be considered representative species for Poaceae lineages. Therefore, we obtained genomic and annotation data for these seven Gramineae species from public databases. First, the latest proteome of rice (version 7.0) was obtained from the dedicated rice database (http://rice.uga.edu/) [66]. The corresponding data for the remaining representative species of grasses were downloaded from the Phytozome database (https://phytozome-next.jgi.doe.gov/) [67] [*Zea mays Ensembl-18*, *Sorghum bicolor Rio v2.1*, *Setaria italica v2.2*, *Setaria viridis v2.1*, *Brachypodium distachyon Bd21-3 v1.1*, and *Hordeum vulgare r1*]. A hidden Markov model for the tubby domain of TLP (PF01167) was downloaded from the Pfam database [18]. HMMER software [68] was then used to identify proteins in the proteome that contained tubby domains (*e* values less than 1*e*-10), and these candidates were further identified by the Pfam, NCBI-CDD [69], and SMART [70] databases. To identify family members as comprehensively as possible, we employed the blastp program (*e* values less than 1*e*-10) to search for all possible TLP family members in the proteomes of these seven grass species, using the amino acid sequence of rice TLPs as a reference. All candidate family members were then confirmed through the domain database above. Finally, the members identified by blastp were identical to those identified by HMMER. In addition, files required for subsequent analyses were downloaded from the corresponding databases, including General Feature Format Version 3 (GFF3) and coding sequence (CDS) files.

Identification of *TLP* gene duplications

The different modes of gene duplication were identified using the DupGen_finder pipeline developed by Qiao et al. [5]. The DupGen_finder pipeline was used for the specific identification of gene pairs corresponding to the five duplication types in a species (see Additional file 12: Fig. S2). First, gene pairs with the five duplication types were identified within the whole genomes of all species, including WGD, TD, PD, TRD, and DSD, using *Spirodela polyrhiza* as the outgroup for monocot plants (the latest proteome and GFF3 annotation files for *S. polyrhiza* were obtained from Phytozome), according to previously described methods and criteria [4, 5, 71, 72]. Then, a custom Python script (Additional file 13: Program S1) was used to extract all duplication pairs of *TLP* for each species.

Calculation of nonsynonymous (*ka*) and synonymous substitutions (*ks*)

A previously published custom Perl program was used to calculate *ka* and *ks* for duplicate gene pairs using the BioPerl module, the ClustalW program, the NG method, and a Poaceae evolution rate of 6.5×10^{-9} [36, 41, 73, 74]. The boxplots of *Ka*/*Ks* and *Ks* for different types of duplicated gene pairs in the seven grass species were drawn by MS Excel.

Construction of a phylogenetic tree

First, full-length amino acid sequences of the TLPs of all grasses were aligned in MUSCLE using default parameters [75]. Then, the Jones–Taylor–Thornton + gamma distributed (JTT + G) model was determined to be the best model via MEGA X [76]. Finally, MEGA X was employed to construct maximum likelihood (ML) trees with the above model and 1000 bootstrap replicates.

According to the number of introns, eukaryotic genes can be divided into three categories: intronless (no introns), intron-poor (three or fewer introns per gene), and intron-rich (more than three introns per gene) [2]. Thus, the number of introns for each *TLP* gene was calculated by CFVisual software [77] based on GFF3 files and characterised into different categories. Finally, based on the distribution of these categories of genes in the phylogenetic tree, Adobe Illustrator (Ai) was used to delabel the intron-poor clade and intron-rich clade of the phylogenetic tree. In addition, the tree of life and polyploidy information of grass species in this study came from previous reports [36, 58, 78].

The constructed phylogenetic tree of the *TLP* gene family was reconciled with the species tree of life using Notung software [79] to infer duplication and loss events of *TLP* genes in grasses. Finally, the number of *TLP* family genes for each ancestral node was back-projected from the family gene size of the extant species [36, 80–82].

Construction and clustering of the synteny network and phylogenetic profiling of clustered communities

We used the synteny network (Synet) method developed by Zhao et al. for syntenic block calculations, network construction, and community detection [38]. First, pairwise all-against-all comparisons were performed using Diamond [83] with default settings [84] for whole-genome proteins of each of the seven grasses. Then, MCScanX was used to compute genomic collinearity between all pairwise genome combinations using default parameters (minimum match size for a collinear block = 5 genes, maximum gaps allowed = 25 genes) [38, 85]. Then, the output files from all the intra- and interspecies comparisons were integrated into a single file named "SynNet-k5s5m25". Finally, a custom Python script (Additional file 14: Program S2) was used to extract the Synet of the grass *TLP* gene family according to Zhao et al.'s criteria; rows containing at least one family gene were retrieved into subnetworks [38]. Clique percolation as implemented in CFinder [86–88] was used to locate all possible k-clique communities for the TLP synteny network to identify communities (clusters of gene nodes) [38]. Then, the clustering results of the TLP synteny network (k = 3) were visualized in Cytoscape [89] to depict an undirected and unweighted network [39]. Finally, gene pairs in different clusters of the TLP synteny network were linked in a phylogenetic tree of the grass *TLP* gene family using iTOL [90] to perform phylogenetic profiling of clustered communities with differently coloured Bezier curves.

Inference of orthologues and selection pressure of single-copy gene clusters

The homology of the grass TLP family was inferred using OrthoMCL software (default parameters) [91]. Then, Cytoscape software was used to draw the lineal orthologous relationship network. Clustering analysis was performed using the Markov cluster (MCL) algorithm (−I > 1.5) [45, 92]. Single-copy clustered genes were used to construct the ML phylogenetic tree via FastTree [93], and the trees were then used to perform further maximum likelihood analysis using the Codeml program in PAML [94]. To detect whether a specific *TLP* gene had been positively selected, we compared two types of competing models, a free ratio model and a ratio-restriction model [95], following our previous steps [36, 56].

Codon bias and determination of optimal codons

To ensure the accuracy of the analysis, we analysed only the given CDSs with ATG (AUG) as the start codon, for which TAG (UAG), TGA (UGA), or TAA (UAA) was the stop codon, whose length was at least 300 bp and in which only A, T, C, and G bases were present [41]. The CodonW program (https://sourceforge.net/projects/codonw/) was used to analyse codon bias (default parameters), and the ENC-plot was created using the R package to detect the effect of base composition on codon bias, with the standard curve calculated as ENC = 2 + GC3s + 29/[GC3s2 + (1—GC3s)2] [40, 96]. Optimal codons for each species were determined as described in our previous studies [41]. Finally, optimal codons shared by different species were mined by Evenn [97].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09389-z.

Additional file 1: Table S1. Patterns of *TLP* gene duplication origins and *Ka*/*Ks* in seven grass species.

Additional file 2: Table S2. List of *TLP* genes in seven studied grass species.

Additional file 3: Figure S1. Exon-intron distribution of all TLP genes in the seven grasses.

Additional file 4: Table S3. Basic statistical details on structural elements of all *TLP* genes in seven grass species.

Additional file 5: Table S4. Synteny network of TLPs in seven grass species.

Additional file 6: Table S5. Orthologous network of TLPs in seven grass species.

Additional file 7: Table S6. The cluster of *TLP* genes in seven grass species obtained using the MCL algorithm.

Additional file 8: Table S7. Analysis of codon bias of *TLP* genes in seven grass species.

Additional file 9: Table S8. Detailed parameter values for codon bias of eligible *TLP* genes in seven grass species.

Additional file 10: Table S9. Optimal codons of *TLP* gene families in seven grass species.

Additional file 11: Table S10. Venn calculation results of optimal codons for seven grass species.

Additional file 12: Figure S2. The flowchart of DupGen_finder pipeline.

Additional file 13: Program S1. Extraction of duplication gene pairs of family genes from all duplication gene pairs of a certain type in a species.

Additional file 14: Program S2. Extraction of synteny networks of family genes from genome-wide synteny networks of all studied species.

Acknowledgements

Not applicable.

Authors' contributions

The study was conceived by H.C.. H.C. and S.F. contributed to data collection and bioinformatics analysis. H.C. and Y.Z. participated in preparing and writing the manuscript. All authors contributed to revising the manuscript. All authors have read and approved the final manuscript.

Funding
This research was supported by a PhD research start-up grant from North China University of Science and Technology (grant number 28424199).

Availability of data and materials
The datasets generated and/or analysed during the current study are available in this published article and the additional files.
Oryza sativa L. genome: the Rice Genome Annotation Project (RGAP) (http://rice.uga.edu/).
Zea mays Ensembl-18, *Sorghum bicolor* Rio v2.1, *Setaria italica* v2.2, *Setaria viridis* v2.1, *Brachypodium distachyon* Bd21-3 v1.1, and *Hordeum vulgare* r1 genome: the Phytozome database (https://phytozome-next.jgi.doe.gov/).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 17 March 2023 Accepted: 18 May 2023
Published online: 30 May 2023

References

1. Song X-M, Wang J-P, Sun P-C, Ma X, Yang Q-H, Hu J-J, Sun S-R, Li Y-X, Yu J-G, Feng S-Y. Preferential gene retention increases the robustness of cold regulation in Brassicaceae and other plants after polyploidization. Horticult Res. 2020;7:202.
2. Liu H, Lyu HM, Zhu K, Van de Peer Y, Cheng ZM. The emergence and evolution of intron-poor and intronless genes in intron-rich plant gene families. Plant J. 2021;105(4):1072–82.
3. Nezamivand-Chegini M, Ebrahimie E, Tahmasebi A, Moghadam A, Eshghi S, Mohammadi-Dehchesmeh M, Kopriva S, Niazi A. New insights into the evolution of SPX gene family from algae to legumes; a focus on soybean. BMC Genomics. 2021;22(1):1–21.
4. Wang Y, Tang H, De Barry JD, Tan X, Li J, Wang X, Lee T-h, Jin H, Marler B, Guo H. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40(7):e49.
5. Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. Genome Biol. 2019;20(1):1–23.
6. Wang Y, Wang X, Tang H, Tan X, Ficklin SP, Feltus FA, Paterson AH. Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms. PLoS ONE. 2011;6(12):e28150.
7. Fang Y, Jiang J, Hou X, Guo J, Li X, Zhao D, Xie X. Plant protein-coding gene families: their origin and evolution. Front Plant Sci. 2022;13:995746.
8. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 2004;431(7008):569–73.
9. Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A. Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell. 2005;17(2):343–60.
10. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A. The Sorghum bicolor genome and the diversification of grasses. Nature. 2009;457(7229):551–6.
11. Wang Y, Li J, Paterson AH. MCScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans. Bioinformatics. 2013;29(11):1458–60.
12. Wang X, Shi X, Li Z, Zhu Q, Kong L, Tang W, Ge S, Luo J. Statistical inference of chromosomal homology based on gene colinearity and applications to Arabidopsis and rice. BMC Bioinformatics. 2006;7(1):1–13.
13. Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res. 2008;18(12):1944–54.
14. Kleyn PW, Fan W, Kovats SG, Lee JJ, Pulido JC, Wu Y, Berkemeier LR, Misumi DJ, Holmgren L, Charlat O. Identification and characterization of the mouse obesity gene tubby: a member of a novel gene family. Cell. 1996;85(2):281–90.
15. Carroll K, Gomez C, Shapiro L. Tubby proteins: the plot thickens. Nat Rev Mol Cell Biol. 2004;5(1):55–64.
16. Lai C-P, Lee C-L, Chen P-H, Wu S-H, Yang C-C, Shaw J-F. Molecular analyses of the Arabidopsis TUBBY-like protein gene family. Plant Physiol. 2004;134(4):1586–97.
17. Du F, Xu J-N, Zhan C-Y, Yu Z-B, Wang X-Y. An obesity-like gene MdTLP7 from apple (Malus× domestica) enhances abiotic stress tolerance. Biochem Biophys Res Commun. 2014;445(2):394–7.
18. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J. Pfam: the protein families database. Nucleic Acids Res. 2014;42(D1):D222–30.
19. Boggon TJ, Shan W-S, Santagata S, Myers SC, Shapiro L. Implication of tubby proteins as transcription factors by structure-based functional analysis. Science. 1999;286(5447):2119–25.
20. Bano N, Aalam S, Bag SK. Tubby-like proteins (TLPs) transcription factor in different regulatory mechanism in plants: a review. Plant Mol Biol. 2022;110(6):455–68.
21. Coleman D, Eicher E. Fat (fat) and tubby (tub): two autosomal recessive mutations causing obesity syndromes in the mouse. J Hered. 1990;81(6):424–7.
22. Ikeda S, He W, Ikeda A. Naggert JrK, North MA, Nishina PM: Cell-specific expression of tubby gene family members (tub, Tulp 1, 2, and 3) in the retina. Invest Ophthalmol Vis Sci. 1999;40(11):2706–12.
23. Wang Y, Seburn K, Bechtel L, Lee BY, Szatkiewicz JP, Nishina PM, Naggert JK. Defective carbohydrate metabolism in mice homozygous for the tubby mutation. Physiol Genomics. 2006;27(2):131–40.
24. Banerjee P, Kleyn PW, Knowles JA, Lewis CA, Ross BM, Parano E, Kovats SG, Lee JJ, Penchaszadeh GK, Ott J. TULP1 mutation in two extended Dominican kindreds with autosomal recessive retinitis pigmentosa. Nat Genet. 1998;18(2):177–9.
25. Gu S, Lennon A, Li Y, Lorenz B, Fossarello M, North M, Gal A, Wright A. Tubby-like protein-1 mutations in autosomal recessive retinitis pigmentosa. The lancet. 1998;351(9109):1103–4.
26. Hagstrom SA, North MA, Nishina PM, Berson EL, Dryja TP. Recessive mutations in the gene encoding the tubby-like protein TULP1 in patients with retinitis pigmentosa. Nat Genet. 1998;18(2):174–6.
27. Kou Y, Qiu D, Wang L, Li X, Wang S. Molecular analyses of the rice tubby-like protein gene family and their response to bacterial infection. Plant Cell Rep. 2009;28(1):113–21.
28. Wardhan V, Jahan K, Gupta S, Chennareddy S, Datta A, Chakraborty S, Chakraborty N. Overexpression of CaTLP1, a putative transcription factor in chickpea (Cicer arietinum L.), promotes stress tolerance. Plant Mol Biol. 2012;79(4):479–93.
29. Yulong C, Wei D, Baoming S, Yang Z, Qing M. Genome-wide identification and comparative analysis of the TUBBY-like protein gene family in maize. Genes & Genomics. 2016;38(1):25–36.
30. Mukhopadhyay S, Jackson PK. The tubby family proteins. Genome Biol. 2011;12(6):1–9.
31. Chia-Ping L, Jei-Fu S. Interaction analyses of Arabidopsis tubby-like proteins with ASK proteins. Botan Stud. 2012;53(4):447–58.
32. Lespinet O, Wolf YI, Koonin EV, Aravind L. The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res. 2002;12(7):1048–59.
33. Bano N, Fakhrah S, Mohanty CS, Bag SK. Genome-wide identification and evolutionary analysis of gossypium tubby-like protein (TLP) gene family and expression analyses during salt and drought stress. Front Plant Sci. 2021;12:667929.
34. Yang Z, Zhou Y, Wang X, Gu S, Yu J, Liang G, Yan C, Xu C. Genomewide comparative phylogenetic and molecular evolutionary analysis of tubby-like protein family in Arabidopsis, rice, and poplar. Genomics. 2008;92(4):246–53.
35. Wang T, Hu J, Ma X, Li C, Yang Q, Feng S, Li M, Li N, Song X. Identification, evolution and expression analyses of whole genome-wide TLP gene family in Brassica napus. BMC Genomics. 2020;21(1):1–14.

36. Chen H, Ge W. Identification, Molecular Characteristics, and Evolution of GRF Gene Family in Foxtail Millet (Setaria italica L.). Front Genet. 2021;12:727674–727674.
37. Ge W, Chen H, Zhang Y, Feng S, Wang S, Shang Q, Wu M, Li Z, Zhang L, Guo H, et al. Integrative genomics analysis of the ever-shrinking pectin methylesterase (PME) gene family in foxtail millet (Setaria italica). Funct Plant Biol. 2022;49:874–86.
38. Zhao T, Holmer R, de Bruijn S, Angenent GC, van den Burg HA, Schranz ME. Phylogenomic synteny network analysis of MADS-box transcription factor genes reveals lineage-specific transpositions, ancient tandem duplications, and deep positional conservation. Plant Cell. 2017;29(6):1278–92.
39. Zhang X, Li X, Zhao R, Zhou Y, Jiao Y. Evolutionary strategies drive a balance of the interacting gene products for the CBL and CIPK gene families. New Phytol. 2020;226(5):1506–16.
40. Wright F. The 'effective number of codons' used in a gene. Gene. 1990;87(1):23–9.
41. Chen H, Ji K, Li Y, Gao Y, Liu F, Cui Y, Liu Y, Ge W, Wang Z. Triplication is the main evolutionary driving force of NLP transcription factor family in Chinese cabbage and related species. Int J Biol Macromolecul. 2022;201:492–506.
42. Wu M, Nie F, Liu H, Zhang T, Li M, Song X, Chen W. The evolution of N6-methyladenosine regulators in plants. Methods. 2022;203:268–75.
43. Huang Z, Duan W, Song X, Tang J, Wu P, Zhang B, Hou X. Retention, molecular evolution, and expression divergence of the auxin/indole acetic acid and auxin response factor gene families in Brassica rapa shed light on their evolution patterns in plants. Genome Biol Evol. 2016;8(2):302–16.
44. Duan W, Ren J, Li Y, Liu T, Song X, Chen Z, Huang Z, Hou X, Li Y. Conservation and expression patterns divergence of ascorbic acid d-mannose/l-galactose pathway genes in Brassica rapa. Front Plant Sci. 2016;7:778.
45. Song X, Wang J, Ma X, Li Y, Lei T, Wang L, Ge W, Guo D, Wang Z, Li C. Origination, expansion, evolutionary trajectory, and expression bias of AP2/ERF superfamily in Brassica napus. Front Plant Sci. 2016;7:1186.
46. Song X, Ma X, Li C, Hu J, Yang Q, Wang T, Wang L, Wang J, Guo D, Ge W. Comprehensive analyses of the BES1 gene family in Brassica napus and examination of their evolutionary pattern in representative species. BMC Genomics. 2018;19(1):1–15.
47. Pei Q, Yu T, Wu T, Yang Q, Gong K, Zhou R, Cui C, Yu Y, Zhao W, Kang X. Comprehensive identification and analyses of the Hsf gene family in the whole-genome of three Apiaceae species. Horticult Plant J. 2021;7(5):457–68.
48. Pei Q, Li N, Yang Q, Wu T, Feng S, Feng X, Jing Z, Zhou R, Gong K, Yu T. Genome-wide identification and comparative analysis of ARF family genes in three Apiaceae species. Front Genet. 2020;11:590535.
49. Yu T, Bai Y, Liu Z, Wang Z, Yang Q, Wu T, Feng S, Zhang Y, Shen S, Li Q. Large-scale analyses of heat shock transcription factors and database construction based on whole-genome genes in horticultural and representative plants. Horticult Res. 2022;9:uhac035.
50. Li J, Qin M, Qiao X, Cheng Y, Li X, Zhang H, Wu J. A new insight into the evolution and functional divergence of SWEET transporters in Chinese white pear (Pyrus bretschneideri). Plant Cell Physiol. 2017;58(4):839–50.
51. Song X, Liu H, Shen S, Huang Z, Yu T, Liu Z, Yang Q, Wu T, Feng S, Zhang Y. Chromosome-level pepino genome provides insights into genome evolution and anthocyanin biosynthesis in Solanaceae. Plant J. 2022;110:1128–43.
52. Xiong Y, Fang J, Jiang X, Wang T, Liu K, Peng H, Zhang X, Zhang A. Genome-Wide Analysis of Multiple Organellar RNA Editing Factor (MORF) Family in Kiwifruit (Actinidia chinensis) Reveals Its Roles in Chloroplast RNA Editing and Pathogens Stress. Plants. 2022;11(2):146.
53. Zhang A, Xiong Y, Fang J, Jiang X, Wang T, Liu K, Peng H, Zhang X. Diversity and functional evolution of terpene synthases in Rosaceae. Plants. 2022;11(6):736.
54. Qiao X, Yin H, Li L, Wang R, Wu J, Wu J, Zhang S. Different modes of gene duplication show divergent evolutionary patterns and contribute differently to the expansion of gene families involved in important fruit traits in pear (Pyrus bretschneideri). Front Plant Sci. 2018;9:161.
55. Wang M, Yuan D, Gao W, Li Y, Tan J, Zhang X. A comparative genome analysis of PME and PMEI families reveals the evolution of pectin metabolism in plant cell walls. PLoS one. 2013;8(8):e72082.
56. Wang X, Gowik U, Tang H, Bowers JE, Westhoff P, Paterson AH. Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses. Genome Biol. 2009;10(6):1–18.
57. Fawcett JA, Maere S, Van De Peer Y. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Natl Acad Sci. 2009;106(14):5737–42.
58. Wang X, Wang J, Jin D, Guo H, Lee T-H, Liu T, Paterson AH. Genome alignment spanning major Poaceae lineages reveals heterogeneous evolutionary rates and alters inferred dates for key evolutionary events. Mol Plant. 2015;8(6):885–98.
59. Paterson A, Bowers J, Chapman B. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci. 2004;101(26):9903–8.
60. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. Synteny and collinearity in plant genomes. Science. 2008;320(5875):486–8.
61. Wang J, Sun P, Li Y, Liu Y, Yang N, Yu J, Ma X, Sun S, Xia R, Liu X. An overlooked paleotetraploidization in Cucurbitaceae. Mol Biol Evol. 2018;35(1):16–26.
62. Duan W, Huang Z, Song X, Liu T, Liu H, Hou X, Li Y. Comprehensive analysis of the polygalacturonase and pectin methylesterase genes in Brassica rapa shed light on their different evolutionary patterns. Sci Rep. 2016;6(1):1–14.
63. Shenton M, Fontaine V, Hartwell J, Marsh JT, Jenkins GI, Nimmo HG. Distinct patterns of control and expression amongst members of the PEP carboxylase kinase gene family in C4 plants. Plant J. 2006;48(1):45–53.
64. Hershberg R, Petrov DA. Selection on codon bias. Annu Rev Genet. 2008;42:287–99.
65. Carels N, Bernardi G. Two classes of genes in plants. Genetics. 2000;154(4):1819–25.
66. Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice. 2013;6(1):1–10.
67. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40(D1):D1178–86.
68 Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39(suppl_2):W29–37.
69. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 2010;39(suppl_1):D225–9.
70. Letunic I, Khedkar S, Bork P. SMART: recent updates, new developments and status in 2020. Nucleic Acids Res. 2021;49(D1):D458–60.
71. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10(1):1–9.
72. Wang W, Haberer G, Gundlach H, Gläßer C, Nussbaumer T, Luo M, Lomsadze A, Borodovsky M, Kerstetter R, Shanklin J. The Spirodela polyrhiza genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. Nat Commun. 2014;5(1):1–13.
73. Thompson JD, Gibson TJ, Higgins DG. Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinform. 2003;1:2.3. 1-2.3. 22.
74. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 1986;3(5):418–26.
75. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5(1):1–19.
76. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35(6):1547.
77. Chen H, Song X, Shang Q, Feng S, Ge W. CFVisual: an interactive desktop platform for drawing gene structure and protein architecture. BMC Bioinformatics. 2022;23(1):1–8.
78. Lee T-H, Tang H, Wang X, Paterson AH. PGDD: a database of gene and genome duplication in plants. Nucleic Acids Res. 2012;41(D1):D1152–8.
79. Chen K, Durand D, Farach-Colton M. NOTUNG: a program for dating gene duplications and optimizing gene family trees. J Comput Biol. 2000;7(3–4):429–47.
80. Yonekura-Sakakibara K, Hanada K. An evolutionary view of functional diversity in family 1 glycosyltransferases. Plant J. 2011;66(1):182–93.

81. Cao J, Li X, Lv Y, Ding L. Comparative analysis of the phytocyanin gene family in 10 plant species: a focus on Zea mays. Front Plant Sci. 2015;6:515.
82. Shao Z-Q, Xue J-Y, Wu P, Zhang Y-M, Wu Y, Hang Y-Y, Wang B, Chen J-Q. Large-scale analyses of angiosperm nucleotide-binding site-leucine-rich repeat genes reveal three anciently diverged classes with distinct evolutionary patterns. Plant Physiol. 2016;170(4):2095–109.
83. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Meth. 2021;18(4):366–8.
84. Zhao T, Schranz ME. Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. Proc Natl Acad Sci. 2019;116(6):2165–74.
85. Artur MAS, Zhao T, Ligterink W, Schranz E, Hilhorst HW. Dissecting the genomic diversification of late embryogenesis abundant (LEA) protein gene families in plants. Genome Biol Evol. 2019;11(2):459–71.
86. Derényi I, Palla G, Vicsek T. Clique percolation in random networks. Phys Rev Lett. 2005;94(16):160202.
87. Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature. 2005;435(7043):814–8.
88. Fortunato S. Community detection in graphs. Phys Rep. 2010;486(3–5):75–174.
89. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.
90. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021;49(W1):W293–6.
91. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–89.
92. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30(7):1575–84.
93. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS one. 2010;5(3):e9490.
94. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–91.
95. Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 1998;15(5):568–73.
96. Wickham H. ggplot2. Wiley Interdisciplin Rev. 2011;3(2):180–5.
97. Chen T, Zhang H, Liu Y, Liu Y-X, Huang L. EVenn: Easy to create repeatable and editable Venn diagrams and Venn networks online. J Genet Genomics. 2021;48(9):863–6.

Publisher's Note