植 物 分 类 与 资 源 学 报 2012，34（1）：33～46
Plant Diversity and Resources DOI：10.3724/SP.J.1143.2012.11117

miR156 靶基因 *SPL* 在植物中的系统分布和进化模式分析*

凌立贞[1,2]，张书东[1**]
（1 中国科学院昆明植物研究所中国西南野生生物种质资源库，云南 昆明 650201；
2 中国科学院研究生院，北京 100049）

摘要： *Squamosa promoter-binding protein-like genes*（*SPLs*）在植物发育过程中具有重要作用。很多 *SPLs* 被 miR156 调节，然而，对于它们在植物中的系统分布和进化模式还知之甚少。本文对 9 个测序物种（藻类，苔藓，石松，单子叶和双子叶植物）的 183 个 *SPLs* 进行了生物信息学分析。结果表明 miR156 应答元件（MREs）仅在陆生植物 *SPLs* 中发现，藻类中不存在。系统进化分析显示陆生植物 *SPLs* 分为两大分支：group I 和 group II。MiR156 靶基因仅分布于 group II，表明它们有着共同的祖先。Group II 进一步分为 7 个亚支（IIa-IIg），miR156 靶基因分布在除 IId 外的其余 6 个亚支的特定 *SPLs*。系统分类与基因结构的相关性反映了 *SPL* 靶基因结构上的变化。在进化过程中，它们可能发生外显子的丢失且伴随 MRE 的丢失。另外，基因重复对 *SPL* 靶基因的丰度变化影响很大，尤其是被子植物与低等植物分歧后它们数量明显增加。以拟南芥为模式植物分析发现串联重复和片段重复是 *SPL* 靶基因扩张的主要机制。

关键词： 系统分析；基因重复；基因结构；microRNA；转录因子
中图分类号： Q 78 **文献标识码：** A **文章编号：** 2095-0845(2012)01-033-14

Unraveling the Distribution and Evolution of miR156-targeted *SPLs* in Plants by Phylogenetic Analysis

LING Li-Zhen[1,2], ZHANG Shu-Dong[1**]
(1 *Germplasm Bank of Wild Species*, *Kunming Institute of Botany*, *Chinese Academy of Sciences*, Kunming 650201, China;
2 *Graduate University of Chinese Academy of Sciences*, Beijing 100049, China)

Abstract: *Squamosa promoter-binding protein-like* genes (*SPLs*) are critical during plant development and mostly regulated by miR156. However, little is known about phylogenetic distribution and evolutionary patterns of miR156-targeted *SPLs*. In this study, 183 *SPLs* from nine genome-sequenced species representing algae, bryophytes, lycophyte, monocots, and eudicots were computationally analyzed. Our results showed that miR156 responsive elements (MREs) on *SPLs* were present in land plants but absent from unicellular green algae. Phylogenetic analysis revealed that miR156-targeted *SPLs* only distributed in group II not group I of land plants, suggesting they originated from a common ancestor. In addition, group II were further divided into seven subgroups (IIa-IIg) and miR156-targeted *SPLs* distributed in some specific members of *SPLs* from six subgroups except subgroup IId. Such distribution pattern was well elucidated by gene structure evolution of miR156-targeted *SPLs* based on the correlation of phylogenetic classification and gene structure. They could suffer from the exon loss events combined with MREs loss during evolution. Moreover, gene duplication contributed to the abundance of miR156-targeted *SPLs*, which had significantly increased after angiosperms and lower plants split. With *Arabidopsis* as the model species, we found segmental and tandem gene duplications predominated during miR156-targeted *SPLs* expansion. Taken together, these results pro-

* Fundation items: The Chinese Academy of Sciences Large-Scale Scientific Facility (2009-LSF-GBOWS-01)
** Author for correspondence; E-mail: sdchang@mail.kib.ac.cn; Tel +86-871-5223137
Received date: 2011-08-16, Accepted date: 2011-11-07
作者简介：凌立贞（1979-）女，博士，主要从事植物进化研究。E-mail: linglizhen@mail.kib.ac.cn

vide better insights in understanding the function diversity and evolution of miR156-targeted *SPLs* in plants.

Key words: Phylogenetic analysis; Gene duplication; Gene structure; MicroRNA; Transcription factor

Abbreviations: CDS, coding sequence; CR, *Chlamydomonas reinhardtii*; DBD, DNA-binding domain; miRNAs, microRNAs; MITEs, miniature inverted repeat transposable elements; ML, maximum-likelihood; MREs, miR156 responsive elements; NJ, Neighbor-Joining; NLS, nuclear localization signal; SBP, *Squamosa* promoter-Binding Protein; *SPL*, *Squamosa promoter-binding protein-like* gene

Squamosa promoter-binding protein-like genes (*SPLs*) encode plant-specific transcription factors (TFs) that share a highly conserved *Squamosa* promoter Binding Protein (SBP) domain and recognize similar target DNA sequences. This SBP-domain spans 79 amino acids residues and features a sequence-specific DNA-binding domain (DBD). The DBD contains two zinc-binding sites assembled as Cys-Cys-His-Cys (Cys2HisCys) and Cys-Cys-Cys-His (Cys_3His) (Yamasaki *et al.*, 2004) and a highly conserved bipartite nuclear localization signal (NLS) in C-terminal (Birkenbihl *et al.*, 2005). It has been proved that the SBP-domain specifically binds to sequences containing a palindromic GTAC core motif (Birkenbihl *et al.*, 2005; Cardon *et al.*, 1997).

Elevated studies have described the functions of members of SBP-box genes in different plant organisms through analysis of either their loss-of-function or gain-of-function mutants. It is known that *SPLs* are critical in diverse biological processes, including seed germination and seedling development (Martin *et al.*, 2010), leaf development (Moreno *et al.*, 1997), phase transition (Gandikota *et al.*, 2007; Wang *et al.*, 2009; Wu and Poethig, 2006), fruit ripening (Manning *et al.*, 2006), copper homeostasis (Kropat *et al.*, 2005; Yamasaki *et al.*, 2009) and grain yield (Jiao *et al.*, 2010; Miura *et al.*, 2010). In fact, it is difficult to point out the exact functions of *SPL* transcription factors in development because of their extreme genetic redundancy and the regulatory complexity. Recently, the variety of elegant approaches elucidated the regulatory mode of *SPLs* and miR156 at different stages of plant development. Their interplay provides the paradigms for how these *SPLs* exert their functions in development. For example, the low-level expression of *SPLs* in miR156-overexpress mutant prolonged the juvenile phase in both maize (Chuck *et al.*, 2007) and *Arabidopsis* (Wu and Poethig, 2006). Another case is the validation of miR156-miR172 gene regulation cascades regulated by *SPL9* from *Arabidopsis* juvenile to adult phase transition. In this case, evidence has been obtained for the direct regulation of miR172b by *SPL9*, a miR156 target (Wu *et al.*, 2009). Over the past few years, many researchers have been working to reveal the functions of the miR156-regulated developmental programs through analyzing the spatio-temporal expression patterns of miR156 and its targets, as well as characterizing the mutations in *Arabidopsis* and maize. The regulatory functions of *SPL* transcription factors in relation to miR156 were documented in several reviews (Chen *et al.*, 2010; Fornara and Coupland, 2009; Nonogaki, 2010; Poethig, 2010).

A large number of work has shown that miR156 families and their targeted *SPLs* are conserved throughout land plants. With respect to miR156 families, the evolutionary study has benefited from large-scale small RNA sequencing/cloning project from many species, in particular the most ancient land plants (e. g. moss). The comparision of the mature miR156 sequences showed that miR156 family was conserved between core eudicots and mosses, elaborating their presence in the earliest common ancestor of land plants. In addition, the targeted *SPLs* for miR156 families are also conserved in plants. For example, a conserved SBP protein target of miR156 has been cloned from the moss *Physcomitrella patens* and shown to be cleaved within the predicted target site (Arazi *et al.*, 2005). More recently, Guo and his colleagues have demonstrated that there is nearly a perfect conservation of the

miR156 target site in *SPLs* for all land plants analyzed but not conserved in the unicellular green alga *Chlamydomonas reinhardtii* (Guo *et al.*, 2008).

Current evidences indicate that considerable divergence of the functions of *SPLs* (including miR156-targeted *SPLs*) exists in plants. For example, in *Arabidopsis SPL3*, *SPL4* and *SPL5* appear to function mostly in the control of flowering time and phase change (Fornara and Coupland, 2009; Wu and Poethig, 2006), whereas *SPL9* and *SPL15* have strong effects on leaf initiation (Schwarz *et al.*, 2008). In addition, studies displayed SBP-box genes were diversified during evolution by analyzing gene structures, phylogeny, and motif elements (Guo *et al.*, 2008; Riese *et al.*, 2007; Yang *et al.*, 2008). Take motif elements for example, some of them was conserved between moss and seed plants (Guo *et al.*, 2008; Riese *et al.*, 2007), whereas others are species-specific after the split of monocotyledon and dicotyledon (Yang *et al.*, 2008). Although the previous studies illustrated the diversity of SBP-box gene family in plants during evolution, they did not detail the evolutionary pathway of miR156-targeted *SPLs*. For example, when such a large set of important MREs has been established in the *SPLs*? Why some of SBP-box genes are targeted by miR156, while others were not. What are the differences between miR156-targeted *SPLs* and non-targeted *SPLs* during evolution? These aforementioned questions intrigue us to glean about the evolutionary information of targeted *SPLs* over long evolutionary timescales. The survey of miR156-targeted *SPLs* occurrence in plants and mapping this information onto the comprehensive plant phylogeny will update our knowledge on their origin and facilitate interpretation of evolutionary pathway and function divergence among distantly related plant species. In addition, the complete sequencing of numerous plant species genome (in particular, the green algae and moss) promotes the comprehensive collection of information on the SBP-box genes. Currently, two integrative transcription factor libraries exploited are available online, which documented SBP-box TF family and other TF families in lower and higher plants (He *et al.*, 2010; Perez-Rodriguez *et al.*, 2009). These resources allow us to perform extensive phylogenetic analyses for miR156-targeted *SPLs* and explore evolutionary history based on their phylogenetic distribution.

Materials and methods

SPL sequences collection

The protein, domain and mRNA sequences of SBP-box genes were downloaded from PlnTFDB v3.0 (Riano-Pachon *et al.*, 2007). The collected sequences included nine genome-sequenced species: one alga (*Chlamydomonas reinhardtii*), one moss (*Physcomitrella patens*), one lycophyte (*Selaginella moellendorffii*), three eudicots (*Arabidopsis thaliana*, *Populus trichocarpa* and *Vitis vinifera*) and three monocots (*Oryza sativa* subsp. *japonica*, *Zea mays* and *Sorghum bicolor*) (Table 1 and Supplementary Table 1). Sequence data for gene and CDS (Coding Sequence) were downloaded from DOE Joint Genome Institute (JGI) (http://www.jgi.doe.gov/) and several species genome annotation databases: The *Arabidopsis* Information Resource (TAIR) 10 genome release (http://www.arabidopsis.org/), TIGR rice genome annotation database release 6.1 (http://rice.plantbiology.msu.edu/), and maize sequence genome database release 5b.60 (http://www.maizesequence.org/index.html). The transcript sequences of grape SBP-box genes were conducted blast analysis to obtain their corresponding gene sequences and CDS in Phytogome v6.0 (http://www.phytozome.net). Some obsolete locus identifiers and new added *SPLs* uniformly adopted the locus identifiers from JGI or species genome annotation database. A total of 183 SBP-box genes were obtained and the complete catalog of them is available in Supplementary Table 1, including the obsolete or new added genes.

Prediction of miR156 responsive elements within *SPL* genes

Mature miR156 sequences of eight species (not

found in green algae) were downloaded from the miRBase database (Release 17.0) (Kozomara and Griffiths-Jones, 2011). They were used to predict *SPL* targets by using miRU with default settings (Zhang, 2005). To further increase the stringency of predicted miR156 targets, we used empirical parameters as a second filter (Schwab *et al.*, 2005). These algorithms were designed to reflect molecular target recognition mechanisms that are assumed to apply to miRNA target recognition. The empirical parameters used in this study were as follows: no mismatch at positions 10 and 11; no more than one mismatch at positions 2-12; no more than two consecutive mismatches downstream of position 13; the total number of mismatch no more than 3 with minor modification. By applying the above rules, our analysis led to the prediction of 61 *SPL* genes as the putative targets for miR156 family (Table 1 and Appendix 1).

Sequence alignment and phylogenetic analysis

Protein and domain sequences from the above nine species were initially aligned using CLUSTALX (Thompson *et al.*, 1997) and manually adjusted in Se-Al software v2.0 (http://evolve.zoo.ac.uk) whenever necessary. Only the SBP-box domains were used for the phylogenetic analysis, because the protein sequences showed no consensus sequences when SBP-box domains were masked. We used PHYLIP (v3.6) (http://www.bioinformatics.uthscsa.edu/www/phylip/) to construct the neighbor-joining (NJ) and maximum-likelihood (ML) tree following Guo's method (Guo *et al.*, 2008). Support values were assessed using 1000 replicate bootstrap tests, only the clades with bootstrap value higher than 50 were shown.

Intron/exon structure and sequence logo analysis

The CDS and genomic sequences of *SPL* genes were used to derive intron/exon structure with Gene Structure Display Server (GSDS, http://gsds.cbi.pku.edu.cn/). The sequence logos were performed using the WebLogo at the URL: http://weblogo.berkeley.edu/logo.cgi.

Chromosomal distribution and duplication analysis

The location of SBP-box genes on chromosomes in *Arabidopsis* was mapped by the Chromosome Map Tool at TAIR (http://arabidopsis.org/jsp/ChromosomeMap/tool.jsp webcite). Gene duplications and their presence on duplicated chromosomal segments were investigated using "Paralogous in *Arabidopsis thaliana*" with the default parameters set (Blanc *et al.*, 2003; Vision *et al.*, 2000; Wang *et al.*, 2008). Only the blocks containing SBP-box genes were retained, and then genes detected were mapped on the chromosomes and linked to each other by lines manually.

Results and discussion

MREs are specific to *SPLs* of land plant lineages

A total of 183 *SPL* genes were obtained from nine species, which represented the main lineages of the green plants: green alga (*C. reinhardtii*), moss (*P. patens*), lycophyte (*S. moellendorffii*), monocots (rice, sorghum and maize) and eudicots (*Arabidopsis*, grape and poplar) (Table 1 and Appendix 1). These genomes have been fully sequenced and all the putative members of the SBP-box gene family have been identified according to their domain structure (Perez-Rodriguez *et al.*, 2009). For example, the *SPLs* from green algae had reached to 23, which was 3 times more than those of the previous report (Guo *et al.*, 2008) when the genome sequence has not been released (Table 1). Therefore, these species open the possibility for a comprehensive analysis of MREs within *SPL* genes. High-confidence prediction

Table 1 The number of SBP-box genes in nine representative plants

Lineage	Organism	*SPLs*	Targeted *SPLs*
Alga	*Chlamydomonas reinhardtii*	23	0
Moss	*Physcomitrella patens*	14	5
Lycophyte	*Selaginella moellendorffii*	11	0
Monocots	*Oryza sativa*	19	12
	Sorghum bicolor	19	9
	Zea mays	33	13
Eudicots	*Arabidopsis thaliana*	17	11
	Populus trichocarpa	29	6
	Vitis vinifera	18	5
Total		183	61

of miR156 targets were performed by miRU based on sequence complementarity and evolution conservation (Zhang, 2005). To further increase the stringency of predicted miR156 targets, we used empirical parameters as a second filter (Schwab *et al.*, 2005). These parameters considered more algorithm features instead only sequence complementarity and conservation (see Materials and Methods in detail). Finally, 61 out of 183 *SPL* genes were the putative target genes for miR156 family with high probability (Table 1 and Appendix 1). We roughly estimated the accuracy of putative targets by seeking confirmations for experimental data. For example, all the putative targets in *Arabidopsis* and rice have been experimentally validated by several independent laboratories (Li *et al.*, 2010; Xie *et al.*, 2006; Xing *et al.*, 2010), indicating that our prediction of conserved miR156 targets was highly accurate. The predicted results showed that MREs were not found in green algae and *Selaginella moellendorffii*, while they were observed in other seven land plants. With respect to green algae, we noticed that no miR156 homologous have been identified after publishing its genome sequence (Worden *et al.*, 2009). To further affirm our prediction, we used the members of miR156 from all other land plants to predict MREs within *SPLs* of green algae. The result indicated that there were still no MREs found in *SPLs* of green algae. Meanwhile, previous studies indicated that no universal miRNA regulatory pathways (including miR156-regulatory pathway) were existed among land plants and green algae (Guo *et al.*, 2008; Molnar *et al.*, 2007). Therefore, we can conclude that the miR156 targets were indeed not appeared in unicellular green algae. On the contrary, the miR156 homologous and *SPL* genes were indentified in *Selaginella moellendorffii*, but the MREs were not predicted in our analysis. One possible explanation is the interactive sites miR156 and *SPLs* had more than four mismatches and did not serve as MREs by using prediction criteria in this study (data not shown). All together, these above analyses concluded that miR156-regulatory pathway had arisen after the divergence multicellular land plants and unicellular green algae.

Phylogenetic distribution of miR156 targeted-*SPLs* in land plants

To understand the evolution history of these targeted *SPLs*, we constructed an unrooted neighbor-joining (NJ) tree for all the *SPLs* of land plants (Fig. 1). In addition, we obtained another tree with similar topology using maximum-likelihood (ML) method (data not shown). As shown in Fig. 1, all SBP-box gene sequences of land plants were resolved into two major clades (group I and group II). Ten *SPL* genes with the SBP-domain of four Cys residues from moss, lycophyte and several flowering plants formed group I (Fig. 1 and Fig. 2: A). A large number of *SPLs* from each land plant lineage were clustered into group II, where they were further divided into seven subgroups (IIa-IIg). The group II had the SBP-domain with a Cys_3 His motif, which was different from group I but same to CR group (Fig. 2: B, C). Based on the phylogenetic data, we speculated that the last common ancestor of land plants had at least two classes of SBP-box genes. Inspection of miR156 targets displayed distribution in group II but not in group I (Fig. 1), indicating that they originated from a common ancestor and had arisen after the divergence of group I and group II.

Similarly, an uneven distribution of the miR156 targets on different subgroups of group II was also apparent (Fig. 1 and Table 2). At first, with the exception of subgroup IId, targeted *SPLs* were widely existed in remaining six subgroups. It was obvious that these targeted *SPLs* restrictedly distributed in some members of six branches (Fig. 1). More strikingly, we found a lineage-specific distribution pattern of targeted *SPLs*. For example, two targeted *SPLs* in subgroup IIe were both from moss, whereas subgroup IIa, IIb, IIf only contained angiosperm targeted *SPLs* (Fig. 1 and Table 2). A similar distribution pattern of miR172 binding sites, a downstream regulatory factor of miR156 targets, that were

only restricted to some member sequences of the *eu-AP2* group of the *AP2*-like family was also reported (Kim *et al.*, 2006). The restricted distribution of miR156 targets suggested their involvement in targeting of selected SBP-box genes in distinct lineages and therefore in the regulation of particular functions of lineage-specific characters. Secondly, the abundance of miR156-targeted *SPLs* in angiosperm lineages was largely different on evolutionary timescale. If angiosperm *SPLs* clustered with lower plant *SPLs* (e.g. moss *SPLs*), suggesting that these angiosperm *SPLs* were early evolved and vice verse. Among seven subgroups, angiosperm *SPLs* and lower plants *SPLs* shared four subgroups (IIc, IId, IIe and IIf), where 20 out of 70 *SPLs* were targeted by miR156. However, the remaining three subgroups (IIa, IIb and IIg) only possessed angiosperm *SPL* genes, where 35 targeted *SPLs* were detected among 55 *SPLs* (Table 2). By comparing the abundance of targeted *SPLs* in angiosperms across subgroups over different evolutionary timescales, we concluded that targeted *SPLs* of angiosperms mainly increased after angiosperms and lower plants split. More importantly, we found a majority of targeted *SPLs* enriched in gene pairs among different angiosperm lineages of group II. Therefore, it implied that gene duplication lead to the increasing of miR156 targets in angiosperms after the divergence of angiosperms from lower plants.

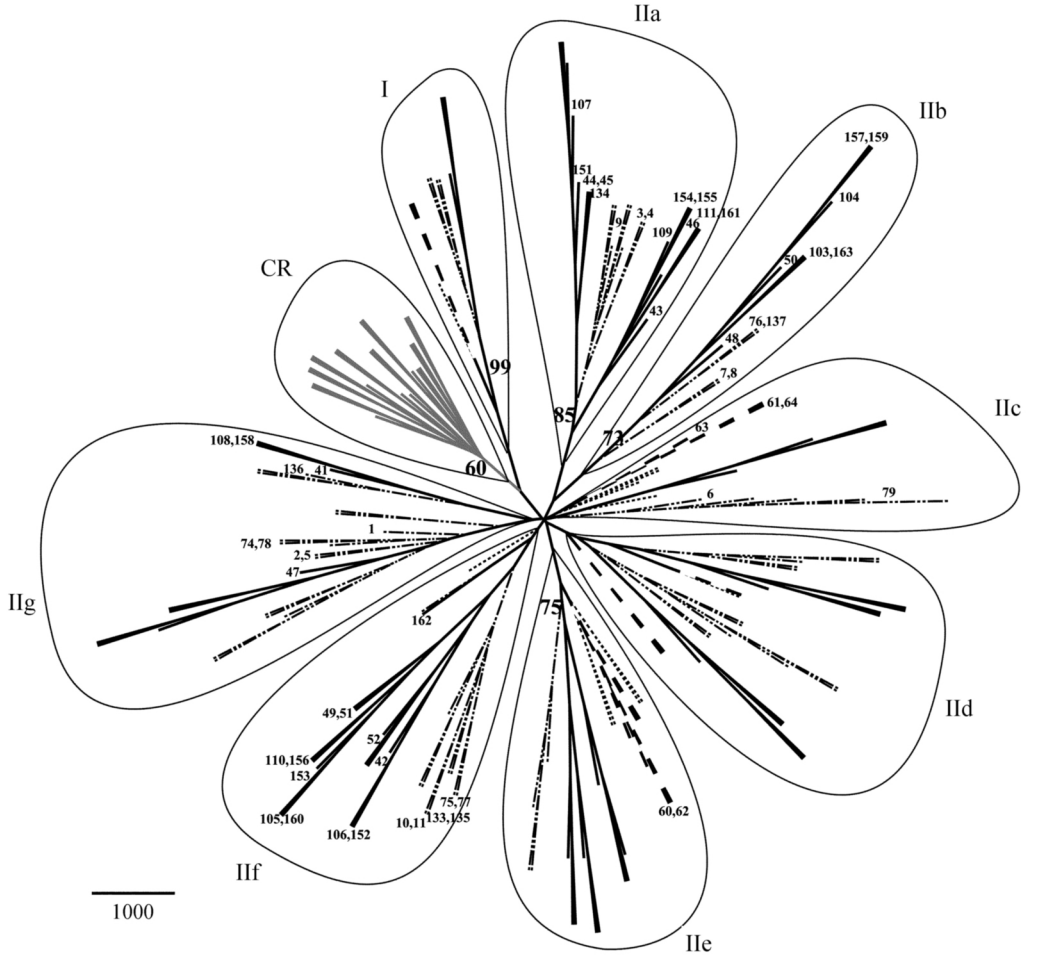

Fig. 1 Phylogenetic tree of SBP-box genes across different species. SBP-box domain sequences of nine plant species were analyzed; an unrooted tree was constructed using Neighbour-Joining (NJ) method, bootstrap 1000 replicates. ——: green algae, - - - -: moss, ·····: lycophyte, ——: monocots, —·—·—: eudicots. Note: The digits inside the branches indicate the support values and those outside the branches indicate the number of miR156-targeted *SPLs* (see Appendix 1 in detail)

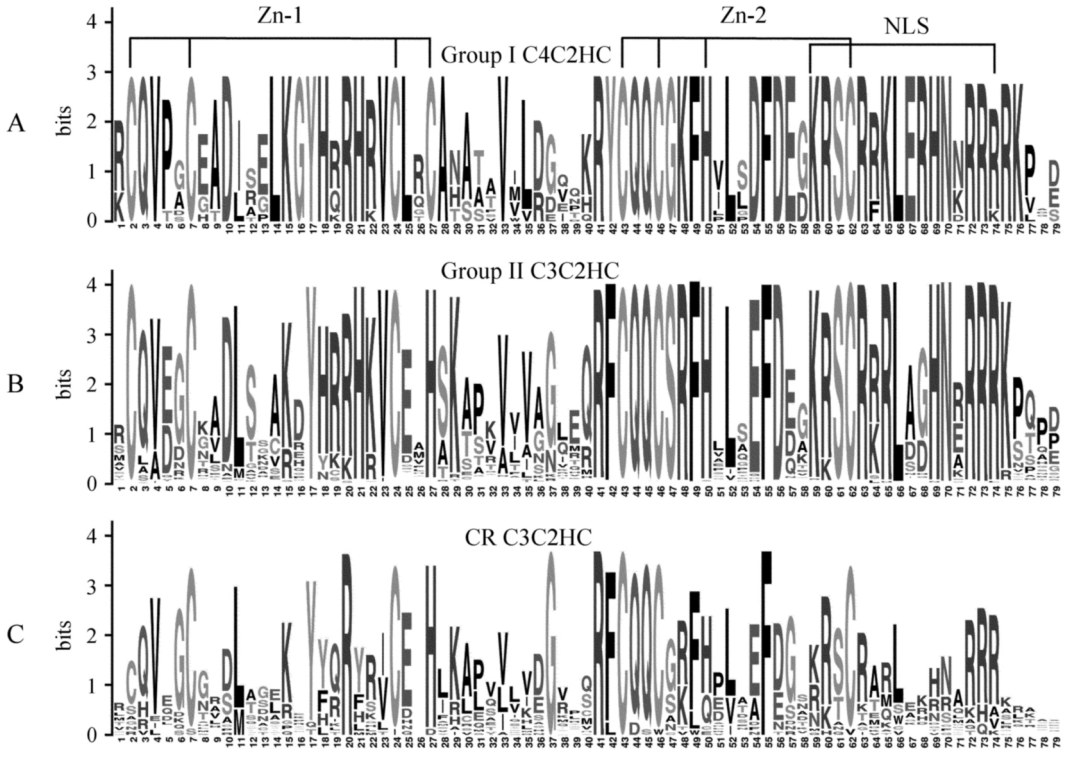

Fig. 2　Sequence logos of the SBP-box domain. The sequence logos of SBP-box domain of group I (A), group II (B), CR group (C). The overall height of letter at each position represents the degree conservation. The two conserved zinc finger structures and NLS are indicated in three different groups

Table 2　The distribution of miR156-targeted *SPLs* on each subgroup of group II

Subgroup	No. of *SPLs*	Lower plants[a]	Higher plants[b]
Subgroup IIa	22	0 (0)	22 (15)
Subgroup IIb	11	0 (0)	11 (11)
Subgroup IIc	16	7 (3)	9 (2)
Subgroup IId	27	4 (0)	23 (0)
Subgroup IIe	24	10 (2)	14 (0)
Subgroup IIf	25	1 (0)	24 (18)
Subgroup IIg	25	0 (0)	25 (10)
Total	150	23 (5)	128 (56)

[a] and [b] The digit in bracket indicates the number of miR156-targeted *SPLs* in lower plants and higher plants, respectively

Gene structure analysis of miR156-targeted *SPLs*

The phylogenetic distribution patterns of miR156 targets could shed light on the evolutionary pathway that shaped their history. To investigate this possibility, we analyzed and compared the gene structure between the targeted *SPLs* and non-targeted *SPLs* because gene structure is an important indicator to classify the different genes. As such, we reconstructed an NJ tree based on 36 SBP-box proteins from eudicots (*Arabidopsis*) and monocots (*Oryza sativa*) and carried out intron/exon structure analysis (Fig. 3: A). It is notable that the intron/exon structure correlated with the classification of *SPL* genes based on the phylogenetic analysis. For example, *SPL* genes in subgroup IId and group I had 11 and 10 exons respectively, while all *SPL* genes of subgroup IIa and IIf had four and three exons respectively (Fig. 3: B). The apparent correlation between intron/exon structures and the classes of *SPL* genes was probably due to the expansion of *SPLs* in each clade by ancient and recent duplication events. The alternative possibility was that the *SPL* genes intron/exon structures could have certain level of stability at the late stages of evolution of angiosperms. Therefore, this good correlation between phylogenetic relationship and gene structure was contributed to understanding the evolution of gene structure of targeted *SPLs* and interpreting their distribution patterns.

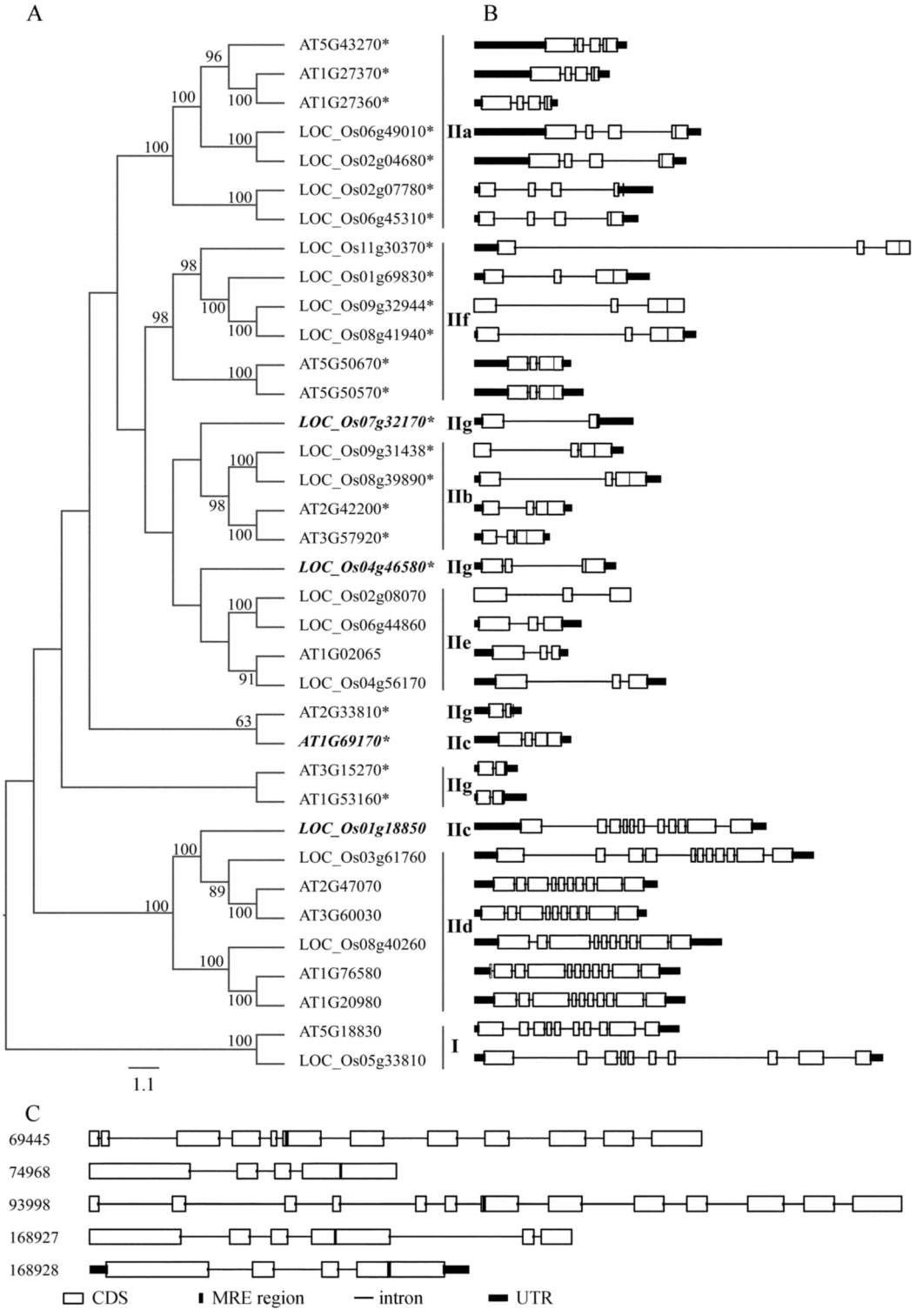

Fig. 3 Intron/exon structure in conjunction with phylogenetic tree for the *Arabidopsis* and rice SBP-box proteins and the structure of targeted *SPLs* in moss. (A) Schematic diagram of phylogenetic tree reconstructed from a complete alignment of 17 *Arabidopsis* and 19 rice SBP-box proteins. (B) Intron/exon structures of SBP-box genes of *Arabidopsis* and rice. (C) Intron/exon structures of targeted *SPLs* in moss. The genes marked the star in the phylogenetic tree were regulated by miR156. As shown in the legend, blank boxes stand for CDS, horizontal lines stand for the introns, black boxes are UTR regions, and vertical bars indicate the position of MREs. The phylogenetic relationships of groups and subgroups were presented in Figure 1. Four SBP-box proteins not assigned the clades corresponding to the domain tree were bold and italic

As shown in Fig. 3B, the coding sequences of SBP-box genes were interrupted by a variable number of exons ranging from two to eleven. The targeted *SPL* genes from different subgroups had similar structure features with two to four exons. In addition, all the targets belonged to a monophyletic clade. By contrast, non-targeted *SPLs* were the most divergent in gene structures and could be classified into two classes according to the phylogenetic position. One class of non-targeted *SPLs* lied outside of the monophyletic clade that the targeted *SPLs* belonged to, such as the genes from subgroup IId. This class of non-targeted *SPLs* contained more exons (at least ten exons) than the targeted *SPLs* but was similar to *SPL* genes of group I. Therefore, we speculate that the targeted *SPLs* might suffer from exon loss events during evolution. Furthermore, moss, an early-branching species of land plants, could provide a window into the early evolution of targeted *SPLs* in land plants. Fig. 3C shows that a portion of targeted *SPLs* in moss possessed the ancient gene structure with exons ranging from 6 to 13, such as targeted *SPLs* 69445, 93998 and 168927. These exons of miR156-targeted *SPLs* might be lost at different dimensions. At first, previous studies suggested that SBP-domains lied in the first two exons and possessed the conserved intron position (Guo *et al.*, 2008; Xie *et al.*, 2006). The authors found that part of moss SBP-box genes had some exons at the upstream of SBP-domain, providing the evidence of exons loss from 5'-end flanking of SBP-domain. In our study, we found the MREs within the above three targeted *SPLs* in moss lied in exon regions excluding the last ones. By contrast, most MREs were located in the last exon and some of them began to reside in 3' UTR regions. These results indicated the exon of targeted *SPLs* might also be lost from the 3'-end regions. A mechanistic explanation for these scenarios suggested that the exons might be lost from the 3'-portion of *SPLs* because of homologous recombination of their cDNAs (Derr, 1998; Mourier and Jeffares, 2003).

The second class of non-targeted *SPLs* (e. g. the genes from subgroup IIe) had the similar gene structure to targeted *SPLs* and also possessed no more than four exons (Fig. 3: A, B). However, they embedded within the same monophyletic clade as targeted *SPLs*. One impossible explanation was that these non-targeted *SPLs* might be originally targeted by miR156 followed by the loss of miR156 binding sites. To test this hypothesis, we further analyzed the phylogenetic relationship across targeted *SPLs* because the paraphyly of miRNA targets on the phylogenetic tree may account for MRE loss. Indeed, five targeted *SPL* genes (e. g. LOC_Os08g39890 and LOC_Os09g31438 from subgroup IIb) and one target gene (LOC_Os04g46580) from subgroup IIg formed paraphyletic branch (Fig. 3: A). However, LOC_Os04g46580 and non-targeted *SPL* genes (e. g. LOC_Os02g08070 and LOC_Os04g56170) in subgroup IIe clustered each other in a branch. Such distribution pattern of MREs suggested a loss of miR156 targeting or alternatively a gain of miR156 targeting in closely related genes. This could be evidence of loss of a MRE after duplication event, because the latter scenario was less likely unless recombinational events or gene conversion events were involved. Overall, these analyses revealed targeted *SPLs* mainly experienced the exon loss events following by some MREs loss during evolution.

Gene duplication of miR156-targeted *SPLs* in *Arabidopsis*

Apart from the relatedness of gene structure, gene duplication was also an important factor to influence the distribution pattern of targeted *SPLs*. As shown in Fig. 3A, more than an half of SBP-box genes constituted gene pairs, such as 12 paralogous gene pairs and 2 orthologous gene pairs indentified based on protein analysis. We observed the paralogous gene pairs in each lineage were mainly regulated by miR156. For example, 5 out of 7 paralogous gene pairs were miR156 targets in *Arabidopsis*. This result suggested that the duplication events in respective lineage were the main resource of targeted

SPLs and influenced the abundance of them on phylogenetic tree. Therefore, it is important to study the duplication mechanisms to interpret the distribution pattern. In our study, we focused on *Arabidopsis*. This species genome has undergone at least two large-scale segmental duplication events, which had great impact on amplification of members of a gene family (including targeted *SPLs*) in the genome. One was the recent polyploidy duplication, which occurred before *Arabidopsis* and *Brassica rapa* split about 24-40 Mya. The other was an older duplication between chromosomal blocks after the divergence of monocot-eudicot around 120 Mya (Blanc *et al.*, 2003; Bowers *et al.*, 2003; Vision *et al.*, 2000). Considering these factors, we investigated SBP-box family gene duplication and distribution on all five *Arabidopsis* chromosomes. The recent segmental polyploidy duplicated blocks were explored by the "Paralogons in *Arabidopsis thaliana*" search engine (Wang *et al.*, 2008).

Fig. 4 showed that there were three pairs of recent duplicated blocks containing SBP-box genes. Both regions on chromosome 1 containing AT1G20980 and AT1G76580 were duplicated segmental block pairs. The region containing AT1G53160 on chromosome 1 and the region containing AT3G15270 on chromosome 3 were duplicated segmental block pairs. The regions on chromosome 2 and on chromosome 3 comprised two duplicated segmental block pairs, such as AT2G42200 and AT3G57920, AT2G47070 and AT3G60030. Among four segmental pairs, there were two duplicated gene pairs targeted by miR156. All of these segmentally duplicated genes were also found to be paralogous in the phylogenetic analysis as shown in Fig. 1. The results indicated that segmental duplication was a major way for SBP-box gene birth (in particular the targeted *SPLs*) for *Arabidopsis*.

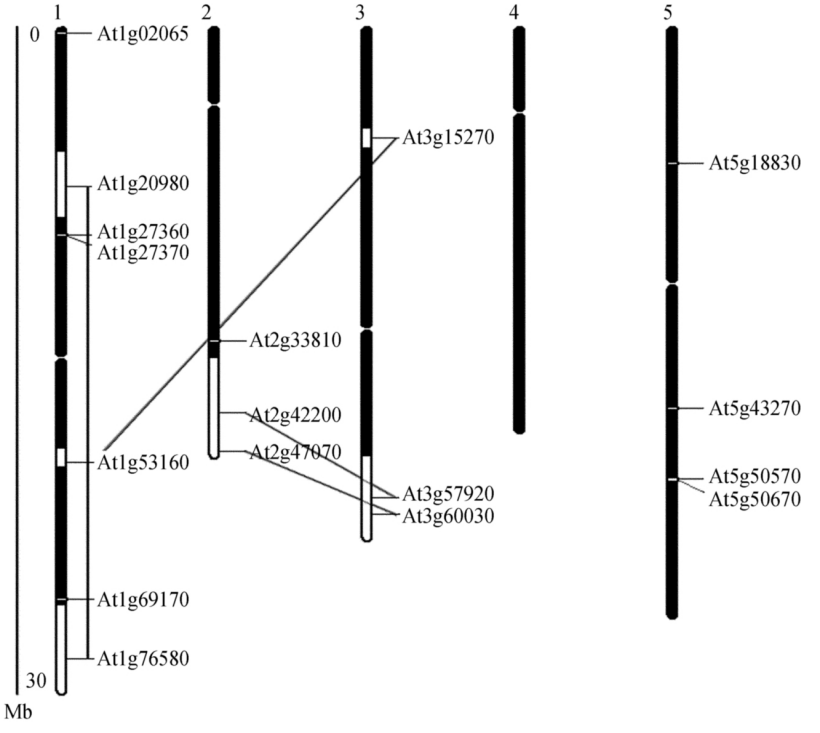

Fig. 4 Chromosomal distribution and duplication events for *Arabidopsis* SBP-box genes. Diagram of five chromosomes of *Arabidopsis* was depicted, 17 SBP-box family genes were distributed on these chromosomes. Only the duplicated regions containing SBP-box genes are shown. Black lines connect corresponding sister gene pairs in duplicated blocks (Blank boxes). AT1G27360 and AT1G27370, AT5G50570 and AT1G50670 are clustered as tandem repeats

Besides, two tandem duplication events were also found on chromosome 1 and 5, respectively. *AT1G27360* and *AT1G27370* were two genes with high similarity of DNA sequence and only 1 kb distance on the chromosome 1. The other two genes, *AT5G50570* and *AT1G50670* had almost consensus similarity, although they depart from about 31 kb distance. The two gene pairs were targeted by miR156. All together, large-scale segmental duplication and tandem duplication events in *Arabidopsis* increased the abundance of targeted *SPLs* and appeared to have exclusively contributed to the current complexes of the targeted *SPLs* and their gene family.

Acknowledgements: The authors kindly thank Dr. Yuxiao Zhang (Kunming Institute of Botany, Chinese Academy of Sciences) for revising the manuscript. The authors also thank Professor Aizhong Liu (Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences) for his constructive advices.

References:

Arazi T, Talmor-Neiman M, Stav R *et al.*, 2005. Cloning and characterization of micro-RNAs from moss [J]. *The Plant Journal*, **43:** 837—848

Birkenbihl RP, Jach G, Saedler H *et al.*, 2005. Functional dissection of the plant-specific SBP-domain: overlap of the DNA-binding and nuclear localization domains [J]. *Journal of Molecular Biology*, **352:** 585—596

Blanc G, Hokamp K, Wolfe KH, 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome [J]. *Genome Research*, **13:** 137—144

Bowers JE, Chapman BA, Rong J *et al.*, 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events [J]. *Nature*, **422:** 433—438

Cardon GH, Hohmann S, Nettesheim K *et al.*, 1997. Functional analysis of the *Arabidopsis thaliana* SBP-box gene SPL3: a novel gene involved in the floral transition [J]. *The Plant Journal*, **12:** 367—377

Chen X, Zhang Z, Liu D *et al.*, 2010. SQUAMOSA promoter-binding protein-like transcription factors: star players for plant growth and development [J]. *Journal of Integrative Plant Biology*, **52:** 946—951

Chuck G, Cigan AM, Saeteurn K *et al.*, 2007. The heterochronic maize mutant Corngrass1 results from overexpression of a tandem microRNA [J]. *Nature Genetics*, **39:** 544—549

Derr LK, 1998. The involvement of cellular recombination and repair genes in RNA-mediated recombination in *Saccharomyces cerevisiae* [J]. *Genetics*, **148:** 937—945

Fornara F, Coupland G, 2009. Plant phase transitions make a SPLash [J]. *Cell*, **138:** 625—627

Gandikota M, Birkenbihl RP, Hohmann S *et al.*, 2007. The miRNA156/157 recognition element in the 3' UTR of the *Arabidopsis* SBP box gene SPL3 prevents early flowering by translational inhibition in seedlings [J]. *The Plant Journal*, **49:** 683—693

Guo AY, Zhu QH, Gu X *et al.*, 2008. Genome-wide identification and evolutionary analysis of the plant specific SBP-box transcription factor family [J]. *Gene*, **418:** 1—8

He K, Guo AY, Gao G *et al.*, 2010. Computational identification of plant transcription factors and the construction of the PlantTFDB database [J]. *Methods in Molecular Biology*, **674:** 351—368

Jiao Y, Wang Y, Xue D *et al.*, 2010. Regulation of OsSPL14 by OsmiR156 defines ideal plant architecture in rice [J]. *Nature Genetics*, **42:** 541—544

Kim S, Soltis PS, Wall K *et al.*, 2006. Phylogeny and domain evolution in the APETALA2-like gene family [J]. *Molecular Biology and Evolution*, **23:** 107—120

Kozomara A, Griffiths-Jones S, 2011. miRBase: integrating microRNA annotation and deep-sequencing data [J]. *Nucleic Acids Research*, **39:** D157

Kropat J, Tottey S, Birkenbihl RP *et al.*, 2005. A regulator of nutritional copper signaling in Chlamydomonas is an SBP domain protein that recognizes the GTAC core of copper response element [J]. *Proceedings of the National Academy of Sciences of the United States of America*, **102:** 18730—18735

Li YF, Zheng Y, Addo-Quaye C *et al.*, 2010. Transcriptome-wide identification of microRNA targets in rice [J]. *The Plant Journal*, **62:** 742—759

Manning K, Tor M, Poole M *et al.*, 2006. A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening [J]. *Nature Genetics*, **38:** 948—952

Martin RC, Liu PP, Goloviznina NA *et al.*, 2010. microRNA, seeds, and Darwin?: diverse function of miRNA in seed biology and plant responses to stress [J]. *Journal of Experimental Botany*, **61:** 2229—2234

Miura K, Ikeda M, Matsubara A *et al.*, 2010. OsSPL14 promotes panicle branching and higher grain productivity in rice [J]. *Nature Genetics*, **42:** 545—549

Molnar A, Schwach F, Studholme DJ *et al.*, 2007. miRNAs control gene expression in the single-cell alga *Chlamydomonas reinhardtii* [J]. *Nature*, **447:** 1129

Moreno MA, Harper LC, Krueger RW *et al.*, 1997. Liguleless1 encodes a nuclear-localized protein required for induction of ligules and auricles during maize leaf organogenesis [J]. *Genes & Development*, **11:** 616—628

Mourier T, Jeffares DC, 2003. Eukaryotic intron loss [J]. *Science*, **300**: 1393

Nonogaki H, 2010. microRNA gene regulation cascades during early stages of plant development [J]. *Plant and Cell Physiology*, **51**: 1840—1846

Perez-Rodriguez P, Riano-Pachon DM, Correa LG *et al.*, 2009. PlnTFDB: updated content and new features of the plant transcription factor database [J]. *Nucleic Acids Research*, **38**: D822—D827

Poethig RS, 2010. The past, present, and future of vegetative phase change [J]. *Plant Physiol*, **154**: 541—544

Riano-Pachon DM, Ruzicic S, Dreyer I *et al.*, 2007. PlnTFDB: an integrative plant transcription factor database [J]. *BMC Bioinformatics*, **8**: 42

Riese M, Hohmann S, Saedler H *et al.*, 2007. Comparative analysis of the SBP-box gene families in *P. patens* and seed plants [J]. *Gene*, **401**: 28—37

Schwab R, Palatnik JF, Riester M *et al.*, 2005. Specific effects of microRNAs on the plant transcriptome [J]. *Developmental Cell*, **8**: 517—527

Schwarz S, Grande AV, Bujdoso N *et al.*, 2008. The microRNA regulated SBP-box genes SPL9 and SPL15 control shoot maturation in *Arabidopsis* [J]. *Plant Molecular Biology*, **67**: 183—195

Thompson JD, Gibson TJ, Plewniak F *et al.*, 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools [J]. *Nucleic Acids Research*, **25**: 4876—4882

Vision TJ, Brown DG, Tanksley SD, 2000. The origins of genomic duplications in *Arabidopsis* [J]. *Science*, **290**: 2114—2117

Wang D, Guo Y, Wu C *et al.*, 2008. Genome-wide analysis of CCCH zinc finger family in *Arabidopsis* and rice [J]. *BMC Genomics*, **9**: 44

Wang JW, Czech B, Weigel D, 2009. miR156-regulated *SPL* transcription factors define an endogenous flowering pathway in *Arabidopsis thaliana* [J]. *Cell*, **138**: 738—749

Worden AZ, Lee JH, Mock T *et al.*, 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas [J]. *Science*, **324**: 268—272

Wu G, Park MY, Conway SR *et al.*, 2009. The sequential action of miR156 and miR172 regulates developmental timing in *Arabidopsis* [J]. *Cell*, **138**: 750—759

Wu G, Poethig RS, 2006. Temporal regulation of shoot development in *Arabidopsis thaliana* by miR156 and its target SPL3 [J]. *Development*, **133**: 3539—3547

Xie K, Wu C, Xiong L, 2006. Genomic organization, differential expression, and interaction of SQUAMOSA promoter-binding-like transcription factors and microRNA156 in rice [J]. *Plant Physiology*, **142**: 280—293

Xing S, Salinas M, Hohmann S *et al.*, 2010. miR156-Targeted and nontargeted SBP-box transcription factors act in concert to secure male fertility in *Arabidopsis* [J]. *The Plant Cell*, **22**: 3935—3950

Yamasaki H, Hayashi M, Fukazawa M *et al.*, 2009. SQUAMOSA promoter Binding Protein-Like7 Is a central regulator for copper *Homeostasis* in *Arabidopsis* [J]. *The Plant Cell*, **21**: 347—361

Yamasaki K, Kigawa T, Inoue M *et al.*, 2004. A novel zinc-binding motif revealed by solution structures of DNA-binding domains of *Arabidopsis* SBP-family transcription factors [J]. *Journal of Molecular Biology*, **337**: 49—63

Yang Z, Wang X, Gu S *et al.*, 2008. Comparative study of SBP-box gene family in *Arabidopsis* and rice [J]. *Gene*, **407**: 1—11

Zhang Y, 2005. miRU: an automated plant miRNA target prediction server [J]. *Nucleic Acids Research*, **33**: W701—W704

Appendix 1 The catalog of 183 SBP-box genes in nine species

No.	Species	Locus ID	Target site position	No.	Species	Locus ID	Target site position
1	*Arabidopsis thaliana*	AT2G33810	3' UTR	14		AT2G47070	
2	(Arabidopsis)	AT3G15270	3' UTR	15		AT3G60030	
3		AT1G27360	CDS	16		AT5G18830	
4		AT1G27370	CDS	17		AT1G76580	
5		AT1G53160	CDS	18	*Chlaymydomonas reinhardtii*	93505	
6		AT1G69170	CDS	19	(green algae)	96716	
7		AT2G42200	CDS	20		101247	
8		AT3G57920	CDS	21		101657	
9		AT5G43270	CDS	22		105679	
10		AT5G50570	CDS	23		106739	
11		AT5G50670	CDS	24		108149	
12		AT1G02065		25		108444	
13		AT1G20980		26		115124	

Continued

No.	Species	Locus ID	Target site position	No.	Species	Locus ID	Target site position
27		115254		72		90199	
28		118761		73		97909	
29		120852		74	*Populus trichocarpa*	743829	3' UTR
30		121606		75	(Poplar)	570289	CDS
31		121939		76		576281	CDS
32		170753		77		733659	CDS
33		171833		78		755123	CDS
34		186869		79		769914	CDS
35		195928		80		179090	
36		288620		81		179183	
37		290479		82		197948	
38		291579		83		216243	
39		405089		84		226094	
40		414856		85		235814	
41	*Oryza sativa* subsp. *japonica*	LOC_Os04g46580	3'UTR	86		245406	
42	(rice)	LOC_Os01g69830	CDS	87		263406	
43		LOC_Os02g04680	CDS	88		267542	
44		LOC_Os02g07780	CDS	89		274234	
45		LOC_Os06g45310	CDS	90		286316	
46		LOC_Os06g49010	CDS	91		286321	
47		LOC_Os07g32170	CDS	92		298307	
48		LOC_Os08g39890	CDS	93		409154	
49		LOC_Os08g41940	CDS	94		412443	
50		LOC_Os09g31438	CDS	95		415293	
51		LOC_Os09g32944	CDS	96		560022	
52		LOC_Os11g30370	CDS	97		647067	
53		LOC_Os01g18850		98		656549	
54		LOC_Os02g08070		99		656553	
55		LOC_Os03g61760		100		798319	
56		LOC_Os04g56170		101		832886	
57		LOC_Os05g33810		102		833398	
58		LOC_Os06g44860		103	*Sorghum bicolor*	4160487	CDS
59		LOC_Os08g40260		104	(sorghum)	4160700	CDS
60	*Physcomitrella patens*	69445	CDS	105		4748489	CDS
61	(moss)	74968	CDS	106		5003160	CDS
62		93998	CDS	107		5003651	CDS
63		168927	CDS	108		5042307	CDS
64		168928	CDS	109		5054656	CDS
65		8925		110		5059026	CDS
66		19787		111		5062217	CDS
67		19788		112		4112095	
68		29422		113		4163165	
69		29851		114		4785561	
70		74970		115		4814910	
71		83876		116		4862557	

Continued

No.	Species	Locus ID	Target site position	No.	Species	Locus ID	Target site position
117		4974192		151	*Zea mays*	GRMZM2G163813	3'UTR
118		4985600		152	(Maize)	GRMZM2G040785	CDS
119		5047795		153		GRMZM2G061734	CDS
120		5059084		154		GRMZM2G065451	CDS
121		5060486		155		GRMZM2G097275	CDS
122	*Selaginella moellendorffii*	17777		156		GRMZM2G101511	CDS
123	(Lycophyte)	28598		157		GRMZM2G126018	CDS
124		28626		158		GRMZM2G148467	CDS
125		28629		159		GRMZM2G307588	CDS
126		28630		160		GRMZM2G390470	CDS
127		28635		161		GRMZM2G414805	CDS
128		49859		162		GRMZM2G450128	CDS
129		59543		163		GRMZM2G460544	CDS
130		59991		164		GRMZM2G024760	
131		79699		165		GRMZM2G036297	
132		437670		166		GRMZM2G058588	
133	*Vitis vinifera*	GSVIVT00002776001	CDS	167		GRMZM2G067624	
134	(grape)	GSVIVT00017032001	CDS	168		GRMZM2G080065	
135		GSVIVT00017953001	CDS	169		GRMZM2G081127	
136		GSVIVT00019157001	CDS	170		GRMZM2G098557	
137		GSVIVT00025360001	CDS	171		GRMZM2G101499	
138		GSVIVT00002800001		172		GRMZM2G102758	
139		GSVIVT00002959001		173		GRMZM2G106798	
140		GSVIVT00003071001		174		GRMZM2G109354	
141		GSVIVT00004625001		175		GRMZM2G113779	
142		GSVIVT00008511001		176		GRMZM2G126827	
143		GSVIVT00018616001		177		GRMZM2G133279	
144		GSVIVT00019158001		178		GRMZM2G133646	
145		GSVIVT00019711001		179		GRMZM2G138421	
146		GSVIVT00019851001		180		GRMZM2G156621	
147		GSVIVT00027720001		181		GRMZM2G156756	
148		GSVIVT00028195001		182		GRMZM2G168229	
149		GSVIVT00030009001		183		GRMZM2G169270	
150		GSVIVT00037879001					

Data resource: Green alge, Moss, Lycophyte, Poplar and Sorghum (http://www.jgi.doe.gov/genome-projects/, the release version is 4.0, 1.1, 1.0 and 1.0 for the first four species, respectively); Grape (http://www.phytozome.net, v6.0); *Arabidopsis* (http://www.arabidopsis.org/, release 10); Rice (http://rice.plantbiology.msu.edu/, v6.1); Maize (http://www.maizesequence.org/index.html, release 5b.60). The gene sequences and CDS of SBP-box genes of each species were downloaded from the above databases. All the transcripts, protein and domain sequences were downloaded from PlnTFDB (v3.0) (http://plntfdb.bio.uni-potsdam.de/v3.0/).

Noting: the 10 obsolete locus identifiers of *SPL* genes: GRMZM2G006850, GRMZM2G015007, GRMZM2G020881, GRMZM2G075639, GRMZM2G090058, GRMZM2G108162, GRMZM2G114243, GRMZM2G145615, GRMZM2G154844 and GRMZM2G160932. Another 6 new added SPL genes: GRMZM2G307588, GRMZM2G390470, GRMZM2G414805, GRMZM2G450128, GRMZM2G460544 and AT1G76580. All the altered genes were from maize except for AT1G76580 (from *Arabidopsis*).