Genome-wide identification and evolutionary analysis of the plant specific SBP-box transcription factor family

An-Yuan Guoa, Qi-Hui Zhua, Xiaocheng Gua, Song Geb, Ji Yangc,*, Jingchu Loua,⁎

a College of Life Sciences, Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, Peking University, Beijing, 100871, China
b State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, 100093, China
c Center for Evolutionary Biology, School of Life Sciences, Fudan University, Shanghai, 200433, China

A R T I C L E   I N F O

Article history:
Received 13 July 2007
Received in revised form 18 March 2008
Accepted 26 March 2008
Available online 9 April 2008

Keywords:
Phylogenetic analysis
miRNA
Conserved motif
Gene structure

A B S T R A C T

We made genome-wide analyses to explore the evolutionary process of the SBP-box gene family. We identified 120 SBP-box genes from nine species representing the main green plant lineages: green alga, moss, lycophyte, gnosisperm and angiosperm. A maximum-likelihood phylogenetic tree was constructed using the protein sequences of the DNA-binding domain of SBP-box genes (SBP-domain). Our results revealed that all SBP-box genes of green alga clustered into a single clade (CR group), while all genes from land-plants fell into two distinct groups. Group I had a single copy in each species except for poplar while group II had several members in each species and can be divided into several subgroups. The SBP-domain encoded by all SBP-box genes possesses two zinc fingers. The C-terminal zinc finger of both group I and group II had the same C2HC motif while their N-terminal zinc finger showed different signatures, C4 in group I and C3H in group II. The patterns of exon–intron structure in Arabidopsis and rice SBP-box genes were consistent with the phylogenetic results. A target site of microRNA miR156 was highly conserved among land-plant SBP-box genes. Our results suggested that the SBP-box gene family might have originated from a common ancestor of green plants, followed by duplication and divergence in each lineage including exon–intron loss processes.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Transcription factors (TFs) regulate and control gene expression in all living organisms. They are usually classified into different families and subfamilies based on the sequence of DNA-binding domains (Luscombe et al., 2000; Riechmann et al., 2005). SQUAMOSA is an Antirrhinum majus floral meristem identity MADS-box gene and the SQUA subfamily of MADS-box genes are critical in floral development as well as other physiological processes (Moreno et al., 2000; Riechmann et al., 2005). We made genome-wide analyses to explore the evolutionary process of the SBP-box gene family. We identified 120 SBP-box genes from nine species representing the main green plant lineages: green alga, moss, lycophyte, gnosisperm and angiosperm. A maximum-likelihood phylogenetic tree was constructed using the protein sequences of the DNA-binding domain of SBP-box genes (SBP-domain). Our results revealed that all SBP-box genes of green alga clustered into a single clade (CR group), while all genes from land-plants fell into two distinct groups. Group I had a single copy in each species except for poplar while group II had several members in each species and can be divided into several subgroups. The SBP-domain encoded by all SBP-box genes possesses two zinc fingers. The C-terminal zinc finger of both group I and group II had the same C2HC motif while their N-terminal zinc finger showed different signatures, C4 in group I and C3H in group II. The patterns of exon–intron structure in Arabidopsis and rice SBP-box genes were consistent with the phylogenetic results. A target site of microRNA miR156 was highly conserved among land-plant SBP-box genes. Our results suggested that the SBP-box gene family might have originated from a common ancestor of green plants, followed by duplication and divergence in each lineage including exon–intron loss processes.

© 2008 Elsevier B.V. All rights reserved.

Abbreviations: AA, Amino acid(s); CR, Chlamydomonas reinhardtii; EST, Expressed sequence tag; miRNA, microRNA; ML, Maximum likelihood; MP, Maximum-parsimony; NJ, Neighbor-joining; SBP, SQUAMOSA promoter binding protein; SPL, SQUAMOSA promoter binding protein like; TF, Transcription factor; UTR, Untranslated region.

⁎ Corresponding authors. J. Yang is to be contacted at tel.: +86 21 6564 3494. J. Luo, tel.: +86 10 62757281; fax: +86 10 62759001.
E-mail addresses: jyang@fudan.edu.cn (J. Yang), luojc@mail.cbi.pku.edu.cn (J. Luo).
site, has been identified as a target of miR156 (Arazi et al., 2005). Recently, Xie et al. identified 11 OsmiR156 targets from rice SBP-box genes and revealed tissue-specific interactions between OsmiR156 and OsSBP genes (Xie et al., 2006). However, it is unclear whether miRNA regulation is conserved in all land-plants and whether the SBP-box gene with a miRNA target site expand in each lineage.

Determining the phylogenetic relationships of the SBP-box gene family is an important step in elucidating the evolution and function divergence of this gene family. Phylogenetic analyses have been described for other plant TF families such as WRKY, MADS, GATA, AP2, DOF, etc. (Reyes et al., 2004; Wu et al., 2005; Zahn et al., 2005; Zhang and Wang, 2005; Moreno-Risueno et al., 2007; Shigyo et al., 2006; Shigyo et al., 2007). Cardon et al. (1999) identified 12 SBP-box genes from Arabidopsis and performed phylogenetic analysis for these Arabidopsis SBP-box genes together with other 12 SBP-box genes from A. majus, rice and maize. With more available plant genome sequences, comparison and phylogenetic analysis of SBP-box genes at genome scale are now possible. In this study, we made a genome-wide identification of SBP-box genes from 9 species representing the main plant lineages and performed phylogenetic analysis and classification to explore the evolution of SBP-box gene family. The feature of the exon–intron structure, the pattern of the conserved motifs, the role of the miRNA target, and the divergence of function are also discussed.

2. Materials and methods

2.1. Identification of SBP-box genes

We obtained the Arabidopsis SBP-box gene list from the DATF SBP-box gene family (Guo et al., 2005) which was built based on the Arabidopsis TAIR6 genome release (http://www.arabidopsis.org/). Oryza sativa ssp. japonica genome data were downloaded from the TIGR rice genome annotation database release 4 (Yuan et al., 2005) (http://rice.tigr.org/). We performed HMMER (http://hmmer.wustl.edu/) search using the Pfam profile PF03110 against japonica proteome sequences and refined the results manually to obtain the rice japonica SBP-box genes. We combined the BLAST search results generated from both TIGR maize gene index and TIGR maize database release 4.0 (http://maize.tigr.org/) with previously reported 8 members (Cardon et al., 1999) to identify maize SBP-box genes. We searched the draft genome sequences from PHYSCoBase (http://moss.nibb.ac.jp/cgi-bin/blast-assemble) for moss (Physcomitrella patens) and lycophyte (Selaginella moellendorffii) SBP-box genes. We obtained poplar (Populus trichocarpa) SBP-box genes from DPTF (http://dptf.cbi.pku.edu.cn), which was built based on the JGI poplar genome release 1.1 (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html). We downloaded green alga (Chlamydomonas reinhardtii) genome sequence from the US Joint Genome Institute (http://genome.jgi-psf.org/) and identified green alga SBP-box genes by HMMER search. Finally, we made BLAST search against the NCBI non-redundant and dbEST databases to obtain SBP-box genes in other species including pine and spruce. We used E value < 1e-5 as the cutoff in HMMER and BLAST searches since the SBP-box domain is very conserved and specific (http://www.sanger.ac.uk/Software/Pfam/data/jtml/seed/PF03110.shtml). Finally, we manually checked the search results to reduce hits with partial SBP-domain and other false positives.

2.2. Phylogenetic analysis

For the phylogenetic analysis, we considered only the amino acid sequence of the SBP-domain since no other regions could be aligned unambiguously for all the sequences available. We used ClustalW (v1.81) (Higgins et al., 1996) for the multiple sequence alignment of the SBP-domains with default settings and manually refined the alignment. We used PHYLIP (v3.6) (http://evolution.genetics.washing-
were different. (TAIR, http://www.Arabidopsis.org). We considered them as duplicates generated by recent duplication event. The locus At1g76580 which was named as AtSPL16 previously did not encode an SBP-domain in either cDNA clone (GenBank: NM_106308 and AY062670) or TAIR annotation. Interestingly, a complete SBP-domain was found by BLASTX at the upstream of the cDNA sequence. It was reported that AtSPL14 (At1g20980) and AtSPL16 (At1g76580) were duplicated gene pairs (Bowers et al., 2003; Blanc and Wolfe, 2004). We excluded At1g76580 since its SBP-domain was lost due to possible frame shift mutation. A complete SBP-domain was distributed in two neighboring japonica loci (LOC_Os11g30380 and LOC_Os11g30370) separated by a stop codon at the conserved intron position of the SBP-domain. We excluded these two genes in our analysis. We removed the extra fragment at the conserved intron position of the SBP-domain of OsSBP17 (LOC_Os09g31438) according to the available rice cDNA sequences (Kikuchi et al., 2003).

3.2. Phylogenetic relationships of SBP-box genes in all lineages

We constructed an unrooted maximum-likelihood (ML) phylogenetic tree for the 120 SBP-box genes from 9 species (Fig. 1) based on the amino acid sequences of their SBP-domains (see Supplementary material, Figure S1 and S2). In addition, using the neighbor-joining (NJ) and maximum-parsimony (MP) methods, we obtained trees with similar topology (data not shown). The tree topology and the corresponding phylogenetic relationships indicated that all proteins from green alga were grouped into the same clade (CR group), while those from land-plants were grouped into several other clades. Group I contained 6 land SBP-domains with a distinct feature that the zinc finger at the N-terminal consisted in four Cys residues while the N-terminal zinc finger of group II SBP-domains had a Cys3His motif (Fig. 2). In addition to the special zinc finger pattern, SBP-box genes in group I showed some features different from the other two groups.

<table>
<thead>
<tr>
<th>Lineage</th>
<th>Organism</th>
<th>Number</th>
<th>Nomenclature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alga</td>
<td>Chlamydomonas reinhardtii</td>
<td>7</td>
<td>CrSPL</td>
</tr>
<tr>
<td>Moss</td>
<td>Physcomitrella patens</td>
<td>14</td>
<td>PpSBP</td>
</tr>
<tr>
<td>Lycophyte</td>
<td>Selaginella moellendorfii</td>
<td>13</td>
<td>SmSBP</td>
</tr>
<tr>
<td>Gymnosperm</td>
<td>Pinus taiwana</td>
<td>3</td>
<td>PsSBP</td>
</tr>
<tr>
<td></td>
<td>Picea glauca</td>
<td>2</td>
<td>PgSBP</td>
</tr>
<tr>
<td>Dicots</td>
<td>Arabidopsis italiana</td>
<td>16</td>
<td>AtSPL</td>
</tr>
<tr>
<td></td>
<td>Populus trichocarpa</td>
<td>26</td>
<td>PtPSBP</td>
</tr>
<tr>
<td>Monocots</td>
<td>Oryza sativa</td>
<td>18</td>
<td>OsSBP</td>
</tr>
<tr>
<td></td>
<td>Zea mays</td>
<td>21</td>
<td>ZmSBP</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>120</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Number of SBP-box genes in nine representative plants

![Fig. 1. Unrooted maximum-likelihood tree of 120 complete SBP-domains. The tree was inferred by the maximum-likelihood (ML) method implemented in Molphy based on amino acid sequences of the SBP-domains. The dash line connecting the clades of group II indicates the uncertain relationship among different subgroups (bootstrap probability <70%). Scale bar corresponds to 0.1 amino acid substitution per residue. Colors denote different lineages, green: green alga, red: moss, blue: lycophyte, brown: gymnosperms, black: monocots, gray: dicots.](attachment:image.png)
Each species from *P. patens*, *S. moellendorfii*, spruce, to *Arabidopsis*, rice and maize had only one group I SBP-box gene except for poplar which had 2 members.

Group II contained the majority of the SBP-box genes from land-plants and it was further grouped into 7 subgroups (IIa–IIg) with high statistical supports (Fig. 1). Subgroup IIa and IIb contained genes from all land-plants. About half moss and lycophyte SBP-box genes fell into subgroup IIb which contained only one *Arabidopsis* and three rice genes. Subgroup IIc contained only vascular plants genes, and genes in subgroup IIa which contained only one *Arabidopsis* and three rice genes. Subgroup IIe contained only vascular plants genes, and genes in subgroup IIId–IIIf were all from seed plants. All three genes in subgroup IIId were from moss. The five gymnosperm SBP-box genes were classified into subgroup IIb and IIId. In most subgroups of group II, we could find two or more poplar SBP-box genes corresponding to one *Arabidopsis* gene, and more than one maize SBP-box genes grouped with one rice gene. SBP-box genes from the same lineage such as moss, lycophyte, gymnosperms and angiosperms tended to be clustered together.

### 3.3. Phylogenetic and gene structure analyses of Arabidopsis and rice SBP-box genes

We used the ML method to construct phylogenetic tree based on the SBP-domain amino acid sequences (Fig. 3a) of the 16 *Arabidopsis* and 18 rice SBP-box genes. The topology was similar to that constructed with 120 SBP-box genes from all lineages. The two genes of group I were grouped as one clade with high statistical support and the genes of group II were clustered into 6 subgroups (IIa–IIg). Furthermore, we made an analysis for the exon–intron structure of the *Arabidopsis* and rice SBP-box genes (Fig. 3b). Our results showed that genes in the same subgroup had similar exon–intron structure except for OsSBP1 in group IIc and OsSBP4 in subgroup IIe. Genes in group I and subgroup Ilb, except for OsSBP6, all had 10 exons and genes in other subgroups, except for OsSBP1, all contained less than 4 exons; while OsSBP1 and OsSBP6 contained 11 exons. The gene structure of OsSBP1 was similar to that of subgroup IIa rather than subgroup IIc (Fig. 3b). Proteins of subgroup IIa and OsSBP1 all had a long C-terminus with conserved sequences of more than 600 amino acid residues (Fig. 3c). Sequence comparison showed that amino acids encoded by the third and fourth exon of OsSBP1 and OsSBP6 were similar to the sequence encoded by the third exon of other genes in subgroup IIa. Genes in subgroup IIId all had 4 exons except for OsSBP4, which missed one exon and one intron.

Our analyses demonstrated that the DNA-binding domain of all land-plants was encoded by two exons except for moss *PsSBP2*. The intron position was highly conserved with the splicing site before the dipeptide Phe-His of the conserved CQQC[S/G][R/K]FH octapeptide. The intron phases of the two exons encoding the SBP-domain were also conserved. We found that most of the rice SBP-box genes had long introns. The average intron length of *Arabidopsis* SBP-box genes was 124bp, shorter than that of the whole *Arabidopsis* genome (168 bp) calculated from *Arabidopsis* genome TAIR6 release, while the average intron length of rice SBP-box genes was 520 bp, longer than that of the whole rice genome (393 bp) calculated from TIGR rice genome release 4.0. Most of them are putative miniature inverted-repeat transposable elements (MITEs) or retrotransposons which can be found in the TIGR Oryza Repeat Database (http://rice.tigr.org/tdb/e2k1/osa1/blastsearch.shtml).

### 3.4. MicroRNA target site and conserved motifs in SBP-box genes

Rhoades et al. (2002) reported that a complementary site of miRNA *miR156/157* in some SBP-box genes and predicted 10 *Arabidopsis* SBP-box genes as potential *miR156* targets. We also found this conserved functional site either in the last coding exon or the 3’-UTR of subgroups IIc–IIf *Arabidopsis* and rice SBP-box genes (Table 2). The miRNA target site was almost completely conserved except for 1 mismatch in OsSBP3. The miRNA target sites in exons of different genes all encoded a conserved peptide ALSLLS. BLAST search indicated that this conserved hexapeptide also existed in SBP-box genes of other species, such as maize and *Medicago truncatula*. *AtSPL3/4/5* and *OsSPL13* in subgroup IIId had only two exons and this miRNA target...
site located in their 3′-UTRs. OsSBP4 and OsSBP11 was a duplicated gene pair and the miRNA target site located in the 3′-UTR of OsSBP4 and the last exon of OsSBP11.

We made MEME search for conserved protein motifs flanking the SBP-domains and the results revealed some conserved motifs among several subgroups (Fig. 3c). Proteins in the same subgroup shared similar number and pattern of conserved motifs. Fig. 4 shows the sequence logo of the top four conserved motifs. Table 3 lists the occurrence of these 4 motifs in Arabidopsis, rice, moss and lycophyte. Motif 1 was found in most subgroups except for subgroup IIa of Arabidopsis and rice. Motif 2 encoded by the miRNA complementary sequence also existed in moss subgroup IIg and lycophyte subgroup IIc. Motif 3 was only found in subgroup IIa of Arabidopsis and rice. Motif 4 was predicted in subgroup IIg of moss and subgroup IIIf of Arabidopsis and rice. Physiological function of these motifs remains to be investigated though the potential miRNA target site of motif 2 was reported.

Proteins in group I, subgroup IIa and OsSBP1 contained a long C-terminus. A database search revealed that the protein sequences of the long terminus were SBP specific. All these SBP-box genes with long C-terminus including moss and lycophyte SBP-box genes in group I, together with CrSPL1 shared a highly conserved 59 AA motif about 120

![Fig. 3. Clustering of 16 Arabidopsis and 18 rice SBPs by three different approaches (a) Maximum-likelihood phylogenetic analysis reconstructed by Molphy. Scale bar corresponds to 0.1 amino acids substitutions per residue. (b) Exon and intron structure. Filled boxes: SBP-domains; white boxes: other exon regions; lines: introns. Numbers 0, 1, and 2: intron phases. The length of the boxes and lines are scaled based on the length of genes except for OsSBP6 and OsSBP9 with long introns denoted by “//”. (c) MEME motif search results aligned based on the DNA-binding domain represented as white boxes with gene names. Conserved motifs are indicated in numbered color boxes. Boxes marked “Long” indicate the long C-terminus.](https://example.com/fig3.png)
AA downstream the SBP-domain (see Supplementary material, Figure S3). This conserved motif had an intron at a conserved position. We observed an Ankyrin (IPR002110) motif about 500 AA downstream the SBP-domain (see Supplementary material, Figure S3). This conserved motif had an intron at a conserved position. We observed an Ankyrin (IPR002110) motif about 500 AA downstream to the SBP-domain in subgroup IIa and rice SBP-box genes of group IIa to 2–3 in moss group II and 0–2 in angiosperms. Genes of angiosperm subgroup IIc and moss PpSBP3 had the same exon–intron structure and conserved motifs. Genes with three exons in subgroup IIc might have evolved by intron loss, while genes with two exons in subgroup IIc might have been formed by degenerating the last exon to the 3′-UTR which retained the miRNA target site. Based on the above evidence, we propose that the diversity of SBP-box gene structure was mainly caused by gene duplication followed by intron and exon loss and it is still an undergoing process in angiosperm SBP-box genes.

4. Discussion

4.1. Evolution of gene structure

The exon–intron structure of three moss SBP-box genes (PpSBP1: AJ968320; PpSBP3: AJ968318; PpSBP4: AJ968319) gave an evidence of exon–intron loss in the evolution of the SBP-box gene structure. PpSBP1 contained two exons at the 5′-end flanking the SBP-box and PpSBP4 had a short exon in the same region (Fig. 5c). On the other hand, PpSBP3 of subgroup IIc did not have the corresponding exon at the 5′-terminus. Furthermore, PpSBP3 had only two exons at the 3′-end flanking the SBP-box while both PpSBP1 and PpSBP4 had three exons in the same region. The gene structure of these three moss SBP-box genes may provide some hints for the gene structure of group Ila and other subgroups in group II in angiosperms. The ancestor SBP-box gene in land-plants might have some exons at both 5′- and 3′-termini flanking the SBP-box. One or two exons upstream of the SBP-box coding region had remained in some moss genes, but they were all lost in angiosperms. The downstream exons of the SBP-box coding region after the SBP-box might also have suffered from exon loss events and the number of exons was reduced from 8 (group I and IIa) to 2–3 in moss group II and 0–2 in angiosperms. Genes of angiosperm subgroup IIc and moss PpSBP3 had the same exon–intron structure and conserved motifs. Genes with three exons in subgroup IIc might have evolved by intron loss, while genes with two exons in subgroup IIc might have been formed by degenerating the last exon to the 3′-UTR which retained the miRNA target site. Based on the above evidence, we propose that the diversity of SBP-box gene structure was mainly caused by gene duplication followed by intron and exon loss and it is still an undergoing process in angiosperm SBP-box genes.

4.2. Origin and evolution of SBP-box genes

It has been proposed that SBP-box genes are plant specific (Cardon et al., 1999). Sequence similarity search against available EST databases and genome sequences including brown algae (taxid:2870), red algae (taxid:2763), blue-green algae (taxid:1117), and golden alga (taxid:2825) suggested that SBP-box genes existed only in green plants. The earliest SBP-box genes we identified in this study were from the genome sequence of C. reinhardtii, a model organism representing the Chlorophyta (green algae). Our results indicated that SBP-box genes were plant specific and might originate predating the divergence of the green algae and the ancestor of land-plants. Based on the results obtained from phylogenetic analysis, gene structure comparison and motif search, we propose a model to account for the evolution of the SBP-box gene family (Fig. 5a).

Our results showed that all SBP-box genes from land-plants were clustered into group I and II while all seven SBP-like genes identified from green alga fell into a separate clade (Fig. 1). In SBP-box genes of the land-plants, the intron position in the SBP-domain as well as the intron phases of the two exons encoding the SBP-domain was conserved. This conserved intron position of SBP-domain indicated that all land-plant SBP-box genes might have originated from a common ancestor. Phylogenetic analysis suggested that SBP-box genes diversified into group I and II before the land-plants started to diverge but after the divergence of green algae from the last common ancestor of land-plants. Interestingly, it has also been reported that the plant specific DOF TF family and the AP2 TF subfamily have the similar pattern of origination and evolution (Moreno-Risueno et al., 2007; Shigyo et al., 2006).

CisPL1 and proteins of group I and subgroup Ila shared the similar pattern of a long C-terminus and a highly conserved motif (see Section 3.4 and Supplemental material, Figure S3). Genes of group I and Ila had many more exons than that of other subgroups with several continuous zero phase introns (Fig. 5b), which might be lost more easily (Roy and Gilbert, 2005). Based on the above evidence, we
assume that the SBP ancestor gene of land-plants might have a complex gene structure with many exons and two zinc fingers C3H and C2HC. It then duplicated and diverged into two ancestor genes of group I and II. The ancestor genes of group I have evolved to the group I genes by altering the first zinc finger from C3H to C4. The group II ancestor gene might have duplicated into three copies. The first copy might evolve to subgroup IIb by losing exons and introns. The second copy might evolve to subgroup Ila and the third copy might be the ancestor of subgroups IIc–Ilg. As group I and subgroup Ila and Ilb contain genes from all land-plants (Fig. 1), we infer that these duplication events might predate the divergence of moss and vascular plant lineage.

Genes of subgroups Ilc–Ilg all had a conserved miRNA target site in their exon or 3′-UTR. We suggest that the miRNA target site might exist in the ancestor of these subgroups. The ancestor gene of subgroup Ilc–Ilg might suffer intron and exon loss and obtain a miRNA target site, then evolve into three clades: Ilc, Ilg and Ild–Ilf (Fig. 1). Subgroup Ilc contained genes of both lycophyte and angiosperms, while subgroups Ild–Ilf contain genes from seed plants only and subgroup Ilg was a mosaic of genes from all land-plants (Fig. 1). We deduce that these duplication events might predate the divergence of moss and vascular plant lineage.

4.3. Duplication of SBP-box genes

Duplication at both gene and genome levels has been and continues to be a pervasive process and contributes to the origin of biological novelty in evolution (Adams and Wendel, 2005). Gene duplication in angiosperm has been reported in many TF families, such as AP2, MADS, DOF, etc. (Zahn et al., 2005; Moreno-Risueno et al., 2007; Shigyo et al., 2006). Some duplicated SBP-box gene pairs (AtSPL10 and AtSPL11, AtSPL4 and AtSPL5, AtSPL1 and AtSPL2, OsSBP10 and OsSBP5, OsSBP11 and OsSBP4, OsSBP12 and OsSBP3) in Arabidopsis and rice have been found in genome analyses (Bowers et al., 2003; Blanc and Wolfe, 2004; Paterson et al., 2004; Wang et al., 2005b). Our analyses demonstrated that SBP-box genes duplicated and diversified in all species during their evolution. SBP-box genes from the same lineage tended to be clustered together in the phylogenetic tree, suggesting that they duplicated after the divergence of the lineages such as moss, lycophyte, gymnosperms and angiosperms (Fig. 1). In most subgroups of group II, two or more poplar SBP-box genes were found along with one Arabidopsis gene indicating that SBP-box genes in poplar experienced duplications after the divergence of poplar and Arabidopsis.

4.4. Function divergence of SBP-box genes

The difference of exon–intron structure and the divergence of amino acid sequence among different subgroups provide us with some hints that SBP transcription factors may have a variety of physiological functions. To date, several important and divergent biological processes regulated by SBP-box genes have been reported, such as flower and fruit development (Klein et al., 1996; Cardon et al., 1997; Wang et al., 2005a; Manning et al., 2006), architecture formation (Becraft et al., 1990; Unte et al., 2003; Stone et al., 2005), sporogenesis (Unte et al., 2003), response to copper and fungal toxin (Eriksson et al., 2004; Stone et al., 2005), as well as control of GA level (Zhang et al., 2006).

Each species in land-plants had only one SBP-box gene in group I except for poplar in which two members were predicted. All group I genes were very conserved and clustered into a separate clade with high statistical support on the phylogenetic tree. The Arabidopsis member AtSPL7 in group I was found to have the highest expression intensity in xylem obtained by Gene Atlas on Genevestigator (https://www.genevestigator.ethz.ch/at/). It would be interesting to explore the exact role of group I SBP-box genes in all green plants by functional characterization.

The variety of subgroups within group II reflected a big spectrum of structural and functional diversity of this group. We found 6 moss and 6 lycophyte SBP-box genes but only one Arabidopsis, 2 poplar and 3 rice genes in subgroup I. Arabidopsis genes in group subgroup I and was reported to involve in sporogenesis (Unte et al., 2003). AtSPL3, a member of subgroup Ild is a putative regulator of the MADS-box TF genes and constitutive expression of AtSPL3 results in early flowering (Cardon et al., 1997). Three AtSPL3 homologs in A. majus (AmSBP1 and AmSBP2) and silver birch (Betula pendula) bind to MADS-box gene regulating flower development (Huisjer et al., 1992; Klein et al., 1996; Lannenpaa et al., 2004). A tomato AtSPL3 homolog (LeSPL-CNFR) is critical for normal fruit development and ripening (Manning et al., 2006). These evidences suggest that SBP-box genes in this subgroup of seed plants play a critical role in flower and fruit development through regulating MADS-box genes.
4.5. Conservation of miRNA target site in SBP-box genes

MicroRNAs play important roles in gene expression regulation and miRNA targets have been found in many TF families, including SBP, MYB, NAC, ARF, CCAAT, GRAS, and AP2 (Rhoades et al., 2002). For example, in the SBP-box gene family, tissue-specific interactions between OsmiR156 and OsSBP target genes were found in rice (Xie et al., 2006), and moss PsSBP3 has also been reported to contain a miR156 target site (Arazi et al., 2005). MI156 over expression causes a moderate delay in flowering and a severe decrease of apical dominance through regulating SBP-box genes (Schwab et al., 2005). In the AP2 TF family, the miR172 target site was conserved in gymnosperm and angiosperm of AP2 homologs (Shigyo et al., 2006). In our case, a miR156 target site was found in many SBP-box genes of moss, lycophyte, and angiosperms (Table 3), suggesting that the regulatory interaction between miR156 and SBP-box genes exists before the divergence of moss from the vascular plants.

Our analysis showed that the SBP-box genes with the miRNA target site existed across many subgroups (Ic–III) in angiosperms, suggestive of the conservation of the miRNA target site because of its functional importance. More importantly, this miRNA target site would move to 3′-UTRs of genes when exons with this site degenerated. Interestingly, we found only one or few genes in moss and lycophyte but many in angiosperms with this target site, indicating that miRNA regulation is more prevalent in angiosperms than other lineages. All these suggest that the regulation of miRNAs on TFs is an ancient and important regulatory mechanism.

Acknowledgements

Our deep appreciation goes to Graham Seymour for providing us with the sequence of LeSPL-CNR and to the TAIR curators for checking AtSPL13 and AtSPL17. We thank Ge Gao, Kun He and Xiaoli Shi, Zhenyin Xie, He Zhang for helpful discussions. This study was supported by grants (NSFC: 30370092, 30121003 and 90408015, 973: 2003CB715900, 863: 2006AA02Z334, MOE: 20060390012).

Appendix A. Supplementary data


References