Systematic characterization of branch point binding protein, splicing factor 1, gene family in plant development and stress responses

Kai-Lu Zhang
Nanjing Forestry University
Zhen Feng
Nanjing Forestry University
Jing-Fang Yang
Central China Normal University
Tian Yuan
Shandong Agricultural University
Di Zhang
The Chinese University of Hong Kong
Ge-Fei Hao
Central China Normal University
Guang-Fu Yang
Central China Normal University
Yan-Ming Fang
Nanjing Forestry University
Jianhua Zhang
The Chinese University of Hong Kong
Caie Wu
Nanjing Forestry University
Mo-Xian Chen
Nanjing Forestry University
Fu-Yuan Zhu (fyzhu@njfu.edu.cn)
Nanjing forestry university https://orcid.org/0000-0002-8276-0550

Research article

Keywords: Alternative splicing, gene expression, phylogenetics, plants, promoter, splicing factor.

Posted Date: May 28th, 2019

DOI: https://doi.org/10.21203/rs.2.9838/v1

Version of Record: A version of this preprint was published at BMC Plant Biology on August 18th, 2020.
See the published version at https://doi.org/10.1186/s12870-020-02570-6.

Abstract

Among eukaryotic organisms, the splicing of nuclear precursor messenger RNA (pre-mRNA) is a process of introns excision and sequentially joining of exons, leading multi-exonic genes to generate multiple splicing isoforms at transcription level. This process is carried out by a super-protein complex defined as spliceosome. Specifically, splicing factor 1/branchpoint binding protein (SF1/BBP) is a single protein that can bind to the intronic branchpoint sequence (BPS), connecting 5' and 3' splice site binding complexes during early spliceosome assembly. The molecular function of this protein has been extensively investigated in yeast, metazoan and mammals. However, their counterparts in plants are seldomly reported. To this end, we conducted a systematic characterization of SF1 gene family across plant lineage. In this work, a total of 92 sequences from 59 plant species were identified. Phylogenetic relationships of these sequences were constructed and subsequent bioinformatic analysis suggested that this family is likely originated from an ancient gene transposition duplication event. Most plant species were shown to maintain a single copy of this gene. Furthermore, an additional RNA binding motif (RRM) existed in most members of this gene family in comparison to their animal and yeast counterparts, indicating their potential role conserved in plant lineage. Our analysis presents general feature of gene and protein structure of this splicing factor family and will provide fundamental information for further functional studies in plants.

Background

In eukaryotes, alternative splicing (AS) removes noncoding intronic sequences of precursor messenger RNA (pre-mRNA) and sequentially joins neighboring exons to generate messenger RNAs (Reed, 2000;Tom and Bosiljka, 2002). The resulting products of AS greatly contribute to post-transcriptional regulation, biological complexity and proteome diversity in eukaryotes (Graveley, 2001;Mo and Manley, 2009;Wahl et al., 2009). Given that on average, approximately 8 exons of each transcript in human transcriptome and the degenerative nature of corresponding splice sites (Graveley, 2001), pre-mRNA splicing is sophistically catalyzed by a multi-megadalton protein complex, spliceosome, consisting of five (U1, U2, U4, U5 and U6) small nuclear ribonucleoprotein particles (snRNPs) and over 100 spliceosomal proteins (Wahl et al., 2009). Furthermore, the early assembly of spliceosome complex E or the commitment complex is an ATP-independent process and contains U1 snRNPs, SF1 and U2 snRNP auxiliary factors (U2AF large and U2AF small subunits) (Mount et al., 1983;Michaud and Reed, 1991). Subsequently, the pre-spliceosome complex A is formed by replacing SF1 with SF3b155/SAP155 of U2 snRNPs (Gozani et al., 1998;Staley and Guthrie, 1998;Will et al., 2014). Stepwise assembly of the following spliceosome during the splicing reaction has been reported as well (Shih-Peng et al., 2003;Makarova et al., 2014), however, splice site recognition is the critical step during early assembly of spliceosome. The current model describes the binding of U1 snRNP and U1 snRNA of a short stretch of 6 nucleotides at 5' splice site, splicing factor 1 (SF1)/mammalian branch point binding protein (mBBP) at branch point and U2 snRNP auxiliary factors at 3' splice site (Margherita et al., 2011). These three *cis*-elements are necessary but usually insufficient to define a specific exon–intron boundary. Thus, additional splicing enhancers or silencers located at exons

and introns may allow the recognition of genuine splice sites during early spliceosome assembly (Jana et al., 2004).

Importantly, SF1 preferentially binds to the intron branch point sequence (BPS) which is adjacent to the binding site (polypyrimidine tract, Py) of U2AF large subunits (mammal U2AF65 and fission yeast U2AF59), bridging U1 and U2AF to form an intermediate lariat structure (Zamore et al., 1992;Reed, 2000). In particular, the SF1 is characterized by the presence of two types of RNA binding motifs at N-terminus, a K homology/Quaking 2 (KH/QUA2) domain originated from the human heterogeneous ribonucleoprotein (hnRNP) K protein (Gibson et al., 1993;Siomi et al., 1993) and one or two zinc knuckle motif(s) ($CX_2CX_4HX_4C$, X represents any amino acid), and a proline-rich region at C-terminus (Arning et al., 1996;Abovich and Rosbash, 1997). Intriguingly, the yeast KH domain specifically binds to the BPS of pre-mRNAs with a Gly-Pro-Arg-Gly motif and the variable loop of the KH domain (Liu et al., 2001) and is necessary for spliceosome assembly (Rain et al., 1998), while the first but not the second zinc knuckle domain in yeast has been demonstrated to bind RNA with high affinity (Garrey et al., 2006). Moreover, the stability of SF1–U2AF65–RNA complex could be further affected by the phosphorylation status of several SF1 serine residues (Ser20, Ser80 and Ser82) *in vitro* (Manceau et al., 2010). The proline-rich region of SF1 is able to interact with U1 snRNP Prp40/FBP11 in yeast and human, respectively (Abovich and Rosbash, 1997;Lin and Lu, 2004). With the aspect of its interaction partner, U2AF large subunit, N-terminal of SF1 interacts with its non-canonical RNA recognition motifs (RRM) or U2AF homology motif (UHM) (Rain et al., 1998;Selenko et al., 2003), whereas the other two RRMs of U2AF large subunit binds to the Py region (Sickmier et al., 2006).

Previous study in fission yeast (*Schizosaccharomyces pombe*) suggests that the initial co-recognition of the branch site and 3’ splice site is pivotal for correct splicing of target pre-mRNAs (Sasaki-Haraguchi et al., 2015). Because the importance of splice site recognition for gene expression and protein diversity, SF1 has been demonstrated to play essential role among a number of eukaryotic species including human (*Homo sapiens*), mice (*Mus musculus*), budding yeast (*Saccharomyces cerevisiae*), common fruit fly (*Drosophila melanogaster*) and roundworm (*Caenorhabditis elegans*) (Abovich and Rosbash, 1997;Mazroui et al., 1999;Tanackovic and Kramer, 2005;Shitashige et al., 2010). For example, in human, missense mutation of splicing factors which are responsible for splice site recognition, such as SF1, has been linked to tumorigenesis (Kenichi et al., 2011). Similarly, heterozygous SF1 (+/-) knockdown mice is susceptible to colon tumorigenesis treated by an organotrophic carcinogen, azoxymethane (Shitashige et al., 2010) and SF1 has been found to associate with beta-catenin/TCF4 complex, suggesting its role in carcinogenesis (Miki et al., 2007). In contrast, knockdown of SF1 is able to suppress the development of germ cell tumor in mice (Zhu et al., 2010), indicating its tissue dependency in cancer research. Furthermore, the molecular function of SF1 has been extensively studied in yeast. For instance, a sf1 mutant strain causes frequent exon skipping in fission yeast (Noriko et al., 2007). And SF1 has been proposed to recognize sub-optimal sequences in specific introns and to accumulate pre-mRNA with aberrant splicing in nucleus (Vincent et al., 2004). However, increasing evidences indicate that this protein is a regulator on splice site recognition specifically during alternative splicing by targeting a subset of

genes (Tanackovic and Kramer, 2005;Noriko et al., 2007;Margherita et al., 2011). This hypothesis is supported by the facts that knockdown of SF1 in both yeast and human extracts only slightly affect the splicing outcome (Guth and Valcárcel, 2000). And RNAi approach on this gene has been demonstrated to not affect the splicing pattern of several splicing marker genes (Tanackovic and Kramer, 2005).

In comparison to studies in human and yeast, few reports have been published related to plant *SF1* genes. Similar function of *Arabidopsis SF1* gene has been proposed in an early study at 2014 (Jang et al., 2014). This plant SF1 homologue has been reported to be responsible for splicing of a group of transcripts. And the loss-of-function mutant (*atsf1-2*) of this gene leads to abnormal development (early flowering and dwarfism) and ABA or heat stress sensitivity in *Arabidopsis* (Jang et al., 2014;Lee et al., 2017). Subsequently, the domain structure and its functional relationship has been substantially investigated (Lee et al., 2017), suggesting that the RRM domain is crucial to maintain its function in plants. However, no related investigations have been performed on the phylogenetic analysis of plant *SF1* genes and their regulations. To this end, we systematically identified 92 *SF1* sequences from 59 plant species, ranging from algae to higher plants. Meanwhile, the gene and protein structure, potential regulation at promoter region and the expression pattern of these genes have been further investigated. In this study, we hypothesize that SF1 is structurally different from their counterparts of animals and yeast, but is conserved among lower and higher plants, indicating its specific role in alternative splicing at branch point recognition.

Methods

Sequence acquisition and identification of plant *SF1* genes

The *Arabidopsis thaliana* SF1 protein sequence (AT5G51300) was used to search similar sequences in all the available plant species from Phytozome v12.1 database (https://phytozome.jgi.doe.gov/pz/portal.html) (Goodstein et al., 2012) by performing BLASTp program with an e-value cutoff = $1e^{-10}$. Then the retrieved protein sequences were examined and filtered using the HMMER score (Johnson et al., 2010), which contained PF16275 (Splicing factor 1 helix-hairpin domain, SF1-HH), PF00013 (K Homology domain, KH_1) and PF00076 (RNA recognition motif, RRM_1). Finally, 92 putative *SF1* sequences from 59 plant species were identified. Detailed information including groups, plant species, common names and number of SF1 homologs reported for each plant species for subsequent analysis were listed in Table S1. Subcellular locations prediction of identified SF1 proteins was carried out using WoLF PSORT (https://wolfpsort.hgc.jp/).

Construction of molecular phylogenetic tree of plant *SF1* genes

Protein sequences of aforesaid plant *SF1* genes were extracted from Phytozome v12.1 database for phylogenetic relationship analysis. The one with the longest coding sequences was chosen for genes with multiple different splicing isoforms. Then multiple SF1 protein sequences were aligned by

performing Muscle v3.8 software with default settings (Edgar, 2004). The molecular phylogenetic tree of plant *SF1* genes was then conducted using the maximum likelihood method (ML, JTT+G+I model) via PhyML v3.0 program (Guindon et al., 2010). FigTree v1.4.3 was used to visualize and edit the phylogenetic tree.

Gene structure, protein domain and multiple Em for motif elicitation (MEME) analysis

Required genomic, cDNA, peptide sequences and all *SF1* gene structures were downloaded from Phytozome v12.1 database. Corresponding intron phases were conducted using the online program Gene Structure Display Server 2.0 (GSDS2.0) (http://gsds.cbi.pku.edu.cn) (Hu et al., 2014). Correlation analysis of SF1 exons were performed by using piece2 webserver (http://www.bioinfogenome.net/piece/search.php? tdsourcetag=s_pctim_aiomsg) (Wang et al., 2016). SF1 protein sequences were used to search for matching Pfam families using the HMMER website (https://www.ebi.ac.uk/Tools/hmmer/) (Finn et al., 2015). Then protein domain patterns were drawn by using TBtools software (Chen et al., 2018) according to the full Pfam resultant table. Conserved motifs of plant *SF1* cDNA sequences and protein sequences were analyzed on MEME online program (http://meme-suite.org/tools/meme) (Bailey et al., 2009), considering 10 maximum most conserved motifs predicted for each sequence and leaving other settings on the default parameters.

Motif prediction in promoter regions of plant *SF1* genes

The 1.5-kb 5'-flanking sequences of plant *SF1* genes were extracted from genomic data available in Phytozome database. Prediction of plant putative cis-elements was performed by using online server PlantCARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/) (Lescot et al., 2002a). Motifs related to tissue-specific expression, internal hormones and external environmental stresses responses, were selected for further analysis and discussion.

Expression analysis base on microarray datasets

Expression data of fractional *SF1* gene family members, including tissue-specificity and stress responses, were extracted from eFP browser series of the Bio-analytic Resource for plant biology (http://bar.utoronto.ca/) (Kiana et al., 2005). Expression values of selected plant *SF1* genes were logarithmic transformed (log_2) to generate visualize expression differences heatmaps by using the BAR HeatMapper Plus tool program (http://bar.utoronto.ca/ntools/cgi-bin/ntools_heatmapper_ plus.cgi).

Analysis of protein-protein interaction network and structural conservation

Proteion-protein interaction network was generated by STRING website (https://string-db.org) (Damian et al., 2015) using representative protein sequence of *Arabidopsis*. The basic settings were employed as the following: meaning of network edges, evidence (line color indicates the type of interaction evidence); and active interaction sources, experiments.

There are three domains in the *Arabidopsis* SF1 protein. The phosphorylation and U2AF65 binding of the n-terminal domain of splicing factor 1 during 3 splice site recognition of *Homo sapiens* (PDBID: 2M0G, identity: 36%, E-value: 7E-17) was similar to K Homology domain. The structure for recognition of the intron branch site RNA by splicing factor 1 of *Homo sapiens* (PDBID: 1K1G, identity: 47%, E-value: 9E-27) can be used as the template of splicing factor 1 helix-hairpin domain. Therefore, the homely modeling was performed by modeller (Madhusudhan et al., 2014) based on two crystal structures. The amino acid conservation scores were calculated using ConSurf Web server based on ML method (Ashkenazy et al., 2016). The 3D model and multiple sequence alignment (Figure S4) were provided as input attributes. Figures were drawn based on Pymol (Yuan et al., 2017).

Analysis gene structure evolution with ortholog group of *SF1* genes

The reconstruction of the evolutionary history of the structure of the plant *SF1* family of orthologous genes was carried out by searching AT5G51300.1 in PIECE 2 sever (http://www.bioinfogenome.net/piece/index.php), which provided exon-intron display for orthologous genes from gene structure data sets linked to the phylogenetic tree.

Results

Sequence identification and phylogenetic analysis of the plant *SF1* gene family

To identify *SF1* gene family in plants, we carried out a BLAST search by using *Arabidopsis* (AT5G51300) amino acid sequence against Phytozome database (v12.1). After filtering sequence without *SF1* signature or truncated sequences, a total of 92 sequences from 59 plant species were retrieved, roughly classifying as 7 algae, 5 bryophyta, 1 basic angiosperm, 21 monocots, and 58 eudicots (Table S1). Specifically, the only copies were the four in one species of SF1 plants, which is *Eutrema salsugineum* (salt cress) (Table S1). In particular, three copies of *SF1* genes were observed among five species, including *Panicum virgatum* (Switchgrass), *Triticum aestivum* (common wheat), *Daucus carota* (Carrot), *Kalanchoe laxiflora* (milky widow's thrill) and *Salix purpurea* (purple osier willow). In addition, 20 plant species contained two copies and 33 species, including model plants *Arabidopsis*, possessed only one copy of plant *SF1s*, respectively. The relatively larger number of SF1 genes and higher number of plant species in this work demonstrated the universality and complexity of *SF1* gene family. And retrieved sequences of 59 plant species provided us a more complete information to analyze the phylogenetic

relationship of SF1 gene family. Subsequently, a rooted phylogenetic tree was constructed based on the above-mentioned 92 protein sequences by using maximum likelihood method, and the tree's bootstrap was represented by a color gradient (Figure 1). In general, all SF1 protein sequences were clustered into four major clades including alga (in yellow), other land plants (in green), monocots (in pink) and eudicots (in blue), and there is one species (*Amborella trichopoda*) belonged to basic angiosperm (shown in colourless). The clear topology and overall high bootstrap values supporting each clade indicated the validity of this phylogenetic reconstruction of plant *SF1* family (Figures 1 and 2, left panel), which also mirrored the similar evolutionary trend from lower plants to higher plants by other studies. For example, the genes of algae in yellow branch were representative member of lineage that diverged before the evolution of land plants, which formed the basal part of this phylogenetic tree. In blue branch, as model eudicot (*Arabidopsis*), two sequences from *Arabidopsis lyrate*, one sequence from *Arabidopsis halleri* and one sequence from *A. thaliana* formed a small clade, consistent with their closer phylogenetic relationships while the BS value of the *A. lyrate* (AL2G22290.t1 and AL8G25570.t1) sequences were not too high. In addition, Cagra.3782s0026.1.p from *Capsella grandiflora* and Carubv10025900m from *C. rubella* with the *Arabidopsis* sequences formed a subclade, which all from Brassicaceae (Figure 1 and Table S1). Usually, some homologous SF1 sequences from the same species were clustered in a same small branch and next to each other, including cashew, soybean, apple, woodland strawberry, quinoa, carrot, colorado blue columbine, maize, common wheat, cereal grass, moss and bog moss (Figure 1 and Table S1). In contrast, some homologous SF1 members from the same species were clustered into the different subclades, such as purple osier willow, poplar, eastern cottonwood, salt cress, potato diploid kalanchoe, milky widow's thrill, hall's panicgrass, switchgrass, green algae and volvox (Figure 1 and Table S1).

Gene structure and conserved motif analysis

It is necessary to compare the exon-intron organization and conserved motifs of plant *SF1* gene family to clarify their evolutionary process and potential function. The gene structure models of *SF1* genes were attached to the phylogenetic tree (Figure 2), and corresponding intron phase of each was also displayed (Figure 2, Table S2). Figure 2 (middle, panel) shows that the gene length and structure of each member of *SF1* family has exited significant difference. For example, the gene structure of 23 members of 92 *SF1* family genes did not contain intron sequences, accounting for 15.7% of the total number of members, and 48 sequences of *SF1* genes had 2 exon-1 intron organization, accounting for 52.2% of all genes. While some genes from algae had multiple exons, one of them containing the most exons (19 exons). Furthermore, sequences located in the same subclade may have different gene structures, for example, two sequences from *Zea mays* (maize) (Zm00008a037777_P01, 3 exons and Zm00008a007621_P01, 4 exons) were observed to have distinctive gene structures. The dissimilation of gene structure of each member of *SF1*s was serious, which may influence the differentiation of their gene function.

Further investigation on conserved motifs by using Multiple Em for Motif Elicitation (MEME) search tool demonstrated that most of *SF1* genes (79 sequences) exhibited similar sequence signatures and same orders, which all contained the 10 analyzed motifs, except one sequence of *Micromonas pusilla* (50949) had different position (Figure 2, right panel). Despite no obvious differences of identified conserved motifs were found among basal angiosperm, monocots and eudicots, sequences from same species may have different motifs (Figure 2). For example, Aqcoe5G406900.1.p and Aqcoe7G039300.1.p from eudicots *Aquilegia coerulea* had 10 motifs and 9 motifs, respectively. The same situation was found in *D. carota*, DCAR_006843, DCAR_008506 and DCAR_004968 had 10 motifs 9 motifs and 10 motifs, respectively. Intriguingly, the CDS length of DCAR_008506 was longest. Notably, some sequences from algae and moss had fewer conserved motifs. For example, in bryophyta, the sequence of *Physcomitrella patens* (Pp3c7_10890V3.1.p and Pp3c11_24710V3.1.p), *Sphagnum fallax* (Sphfalx0015s0077.1.p and Sphfalx0010s0197.1.p) and *Marchantia polymorpha* (Mapoly0009s0189.1.p) had nine motifs. In algal plants, the sequence of *Volvox carteri* (Vocar.0007s0345.1.p and Vocar.0008s0294.1.p) and *Chlamydomonas reinhardtii* (Cre12.g553750.t1.1 and Cre09.g386731.t1.1) had also 9 motifs, while sequence of 145219 and 62857 from *Micromonas* had 7 motifs and 6 motifs, respectively, suggesting algal sequence diversity. Further correlation analysis of *SF1* exon regions were carried out to elucidate the gain/loss of introns. Correlations between two transcripts were shown in Figure 3, providing additional information for phylogenetic analysis.

Analysis of protein domain and conserved motifs in peptides

The protein domains were analyzed by using above selected peptide sequences, whose annotations were splicing factor-related and conserved protein motifs were predicted according to the retrieved peptide sequences by MEME analysis (Figure 4). Consequently, all of the SF1 peptides were predicted to have one N-terminal domain annotated as SF1_HH N-terminal followed by a KH domain and a C-terminal domain namely as RNA recognition motif (RRM) (Figure 4, middle panel). Interestingly, in algae, 3 peptides from *M. pusilla* (145219), *V. carteri* (Vocar.0008s0294.1.p) and *C. reinhardtii* (Cre09.g386731.t1.1) had two RRM domains. SF1 proteins from all species range from 499 to 1583 amino acids and average of all them are approximately around 710 amino acids (Table S3). Consistently, most of them are about 700 to 800. Subcellular locations prediction showed that the subcellular location of majority of SF1 proteins were in nuclear (86, 93.4%) (Table S3). Moreover, proteins of 30147.m014250 (*Ricinus communis*) and Migut.F01191.1.p (*Mimulus guttatus*) were located in vacuolar; proteins of Traes_2DL_6F03F05FA.4 (*T. aestivum*) and 145219 (*M. pusilla*) were predicted in cytoplasmic; proteins of GSMUA_Achr5P25100_001 (*Musa acuminata*) and Cre09.g386731.t1.1 (*C. reinhardtii*) were located chloroplast and endoplasmic reticulum, respectively.

MEME analysis for SF1 peptide sequences to predict a total of ten conserved motifs, presented as colored boxes, which are able to cover most area of the protein (Figure 4, right panel). Further analysis showed that 77 peptides have all ten motifs, accounting for about 83.7% of all SF1 protein sequences

analyzed in the study. Interestingly, all sequences from moss have ten conserved motifs in the analysis, suggesting the conservation of SF1 proteins in bryophyta. Furthermore, almost eudicots had ten conserved motifs except *Anacardium occidentale* (Anaoc.0018s0425.1.p) and *C. grandiflora* (Cagra.3782s0026.1.p) without motif 2 and *Malus domestica* (MDP0000558834), *Fragaria vesca* (mrna21192.1-v1.0-hybrid) and *Brassica rapa* (Brara.C01481.1.p) without motif 10, while most monocots had eight conserved motifs. In contrast, algal plants showed the least degree of conservation in this work, implying the divergence of plant SF1 proteins in algae. Motifs that all algae shared were motif 3, motif 4, motif 5 and motif 9.

Analysis of promoter and tissue-specific expression of *SF1* genes

To further analyze the regulation of plant *SF1* genes at transcription level, the 1.5-kb upstream sequences of plant *SF1* genes were obtained by Phytozome database, then the *cis*-elements of each promoter were identified via using PlantCARE program (Table S4) (Lescot et al., 2002b). Consequently, a total of 108 motifs were predicted. Generally, eight *cis*-elements related to tissue-specific expression among them were selected (Figure 5 and Table S4), including, HD-Zip1 for differentiation of the palisade mesophyll cells, RY-element regulating seed-specific expression, AACA_motif and GCN4_motif involved in endosperm expression, CAT-box, CCGTCC-box, dOCT, and OCT for meristem expression. Further analysis showed that there were only 50 promoters of *SF1* genes which had tissue-specific regulatory *cis*-elements. Particularly, the CAT-box and CCGTCC-box turned up at the highest frequency and with largest numbers in the promoters of *SF1* genes. Both of them regulated meristem-specific expression and played the key roles during development and growth of plants. Consistently, purple false brome (*Brachypodium distachyon*) of monocots not only had CAT-box and CCGTCC-box, but also was highly expressed in young leaves, internode, adventitious roots and roots (Figure 5 and Figure S2). However, no motifs were found to link the high expression of two *SF1*s of *Glycine max* (soybean) in SAM and root-tip (Figure S1). In addition, AACA_motif was only detected in *Solanum tuberosum* (PGSC0003DMP400032853) of potato, suggesting its specific role in regulating endosperm-specific negative expression. While HD-Zip 1 was present in Podel.03G113200.1.p of *Populus deltoides* (eastern cottonwood) and Spipo17G0046100 of *Spirodela polyrhiza* (greater duckweed). RY-element was detected in promoter of the dicot model plant *Arabidopsis*, and low expression was also reported in dry seed in *Arabidopsis* (Figure 6), suggesting that RY-element involved in seed-specific negative expression of *Arabidopsis*. Moreover, it was found that the expression levels in the same tissue type showed significant differences during different growth stages, such as expression level in stamen of flower stage 15 of *Arabidopsis* was obviously higher than the other flower development stages. However, it was observed that the expression levels of different growth stage of *Solanum lycopersicum* were not only similar but lower, and no motifs were found in the promoter in tomato (Figures 5 and S1). Furthermore, the different expression patterns were detected in several *SF1* genes with multiple copies (Figures 6 and S1). For instance, the similar tissue expression profile was detected in two *SF1* homologues from dicot *Populus trichocarpa* (Potri.001G126400.1 and

Potri.003G107200.1) (Figure S1). In contrary, two *SF1* genes of *S. tuberosum* and *G. max* showed differential expression pattern (Figures 6 and S1).

Analysis of promoter and internal and external hormones expression of *SF1* genes

In the long-term evolution and development, plants have gradually formed mechanisms of adaptation and resistance to adversity to maintain their life and sustain growth. In order to understand the regulatory mechanisms of internal and external stimuli on plant *SF1*s, *cis*-acting elements involved in hormone and stress from PlantCARE database were studied (Figure 7, Table S4). Finally, 19 hormone and stress-related motifs were selected from predicted 108 motifs. There are hormone-related motifs including abscisic acid (ABRE), auxin (AuxRE, AuxRE-core, TGA-box, TGA-element), ethylene (ERE), gibberellin (GARE-motif, P-box, TATC-box), MeJA (CGTCA-motif, TGACG-motif), salicylic acid (TCA-element) and stress-related motifs such as low-temperature (LTR), drought (MBS), wound (WUN-motif) and anoxic (ARE, GC-motif). Almost each *SF1* sequence had a great diversity of *cis*-elements in their promoter regions except some sequences such as Araha.13031s0002.1. and Traes_2AL_3D6729692.1 had no one motif due to the sequences contain 'N' or no promoter, suggesting that multiple hormones-mediated signaling pathways were closely related to *SF1* plants resistance. Statistics showed that more than half of *SF1* promoters contained ABRE, CGTCA-motif, TGACG-motif and ARE, respectively. Moreover, external hormone signals could also affect the abundance of *SF1* transcripts (Figure S3). For example, in *Arabidopsis* (AT5G51300.1), MJ (methyl jasmonate) inhibited its expression (Figure 7) and treatments with other hormones like ACC (a precursor of ethylene), auxin (IAA), ABA and GA (gibberellin) would regulate the expression of AT5G51300.1.

Analysis of protein-protein interaction network and structural conservation

Protein-protein interaction (PPI) network analysis can study the working principle of protein in biological system, the molecular mechanism of biological signal and energy metabolism, and the functional relationship between proteins systematically. In this study, we generated protein-protein interaction networks of SF1 protein according to representative protein sequence of *Arabidopsis* (AT5G51300) by STRING database based on experiment (Figure 8A). Finally, the ten predicted functional partners of SF1 protein were obtained, including CDC5 (AT1G09770.1), AT1G10580 (AT1G10580.1), ATU2AF65A (AT4G36690.1), AT2G33440 (AT2G33440.1), AT2G33435 (AT2G33435.1), AT1G60900 (AT1G60900.1), AT1G60830 (AT1G60830.1), MAC3B (AT2G33340.1), MAC3A (AT1G04510.1), AT1G31870 (AT1G31870.1) (Figure 8A). CDC5, MAC3A and MAC3B all are the component of the MAC complex, both of them may be involved in pre-mRNA splicing and DNA repair and probably regulate defense responses through transcriptional control, which is essential for plant innate immunity. AT1G10580 is pre-mRNA-processing factor 17 and AT1G31870 is splicing factor CWC26, both of them participated in RNA splicing and pre-mRNA processing. AT2G33440, AT2G33435 and AT1G60830 are RNA recognition motif-containing protein, whose main molecular function are involved in pre-mRNA splice site binding.

ATU2AF65A and AT1G60900 are splicing factor U2af large subunit A and B, respectively, which are necessary for the splicing of pre-mRNA. And AT5G51300 (splicing factor-like protein 1) was already demonstrated to be necessary for the splicing of pre-mRNA and required during development and for abscisic acid (ABA) responses. In general, SF1 protein and its functional partners are involved in RNA splicing and pre-mRNA processing more or less, and some of them also possess the function that defense response to bacterium (Figure 8A).

The *A. thaliana* SF1 protein includes three domains, splicing factor 1 helix-hairpin domain (residue: 126-237), KH domain (residue: 244-330) and RNA recognition motif (residue: 482-552). The multiple-sequence alignment revealed that the conservations of these domains are relatively high (Figure S4), suggesting the similar function of these genes. Furthermore, a 3D model of splicing factor 1 helix-hairpin domain and KH domain were reconstructed according two crystal structures by using homology modeling approach (Figure 8B). The first domain forms a secondary, hydrophobic interface with U2AF65(UHM) (Yun et al., 2013). The second one is present in a wide variety of nucleic acid-binding proteins (García-Mayoral et al., 2007). Therefore, superimposed the crystal structure of U2AF65 (2M0G) and RNA (1K1G) on the structure from homely modeling to observe the interaction. The residues with higher ConSurf Grade are more conservative. The ConSurf Grade of 198 (74.4%) residues over 7, and the ConSurf Grade of 111 (41.7%) residues over 9. More importantly, the binding domain of RNA is highly conservative (Figure 8B). All the import residues with ConSurf Grade higher than 7, except for Val288. And the residues at position 288 with similar physiochemical properties, including Val and Ile. For another domain, it is not conservative as the splicing factor 1 helix-hairpin domain with a loop interacting with U2AF65. However, the important residues are in relatively higher ConSurf Grade, only two residues (Lys146 and Asp147) are in lower ConSurf Grade than 7. In the lower plants, these two residues are replaced by Ile, Gly, Tyr, Thr, Ala and Gly, Ser, His, separately. At the same time, they are lost in many species. Therefore, the functions of these domains are conservative. And RNA binding domain is much more conserved than U2AF65 binding domain, especially in lower plant.

Discussion

It is well acknowledged that mature mRNA is formed by sequentially ligating adjacent exons to maintain a particular reading frame for protein translation (Sasaki-Haraguchi et al., 2015). In human, nearly all the annotated protein-coding genes undergo alternative splicing (Qun et al., 2008;Wang et al., 2008). And in plants, over 80% of intron-containing genes exhibit splicing isoforms (Zhu et al., 2017;Chen et al., 2019). Furthermore, the process of splicing is tightly regulated by initial recognition of splice site during early spliceosome assembly. Therefore, proteins which are responsible for this recognition are important to be studied and provide valuable targets for genetic control of splicing in eukaryotes (Kotake et al., 2012;Uehara et al., 2017). To this end, the branch point binding protein SF1, which connects both 5’ and 3’ splice site determination complex, emerges as crucial component on splice site choices.

Comparison of structural and functional conservation among plant *SF1* genes

In this study, we systematically characterized 92 plant *SF1* genes from 59 different species. Although over 50% (34/59) of these gene family maintained one copy of *SF1* gene, 26 plant species contained multiple *SF1* members (Table S1), suggesting their functional redundancy. Intriguingly, most of *SF1* genes had one single exon encoding target protein product except for several algal sequences (Figure 2), indicating that an ancient gene transposition duplication event may influence the evolution of this gene family across plant lineage (Hofberger et al., 2015). At the molecular aspect, SF1 is an important component to mediate early spliceosome assembly and splice site recognition. Therefore, substantial investigations have been carried out to elucidate its molecular function in both animals and plants. For example, the primary amino acid sequences and domain architecture of SF1 proteins have been reported to be conserved among eukaryotic organisms such as yeast, human, metazoans and plants (Abovich and Rosbash, 1997;Berglund et al., 1997;Mazroui et al., 1999;Jang et al., 2014). SF1 proteins normally are characterized by three domains named as KH/QUA2, zinc finger and RRM (Lee et al., 2017). However, plant SF1 proteins have been documented to contain an additional RRM domain but lacks UHM-specific features (Lee et al., 2017). Previous study has demonstrated that truncated plant SF1 protein without RRM domain still has sufficient activity for pre-mRNA splicing in response to ABA treatment (Lee et al., 2017). Thus, the potential function of this additional domain in planta needs to be further investigated. Furthermore, post-translational modification such as serine phosphorylation by KIS kinase has been reported to enhance the assembly of SF1–U2AF65–RNA tri-complex (Manceau et al., 2010;Yun et al., 2013) or to recruit other splicing factors during splice site recognition (Abovich and Rosbash, 1997;Ingham et al., 2005).

Functional diversification of plant *SF1* genes revealed by their expression patterns

SF1 is considered to be a pivotal component connecting 5’ and 3’ splice sites definition complex. Furthermore, substantial evidence has proved that SF1 plays crucial roles during splice site recognition among a variety of eukaryotic organisms (Tanackovic and Kramer, 2005;Noriko et al., 2007;Margherita et al., 2011). However, their roles on cell viability remains argumentative. Accumulating evidence suggest that SF1 may not be essential for viability and only controls subsets of genes in plants and animals (Guth and Valcárcel, 2000;Zhu et al., 2010), indicating an alternative mechanism may exist in addition to SF1-mediated splice site recognition (Valcárcel et al., 1996;Haihong and Green, 2006;Margherita et al., 2011). Furthermore, the function of SF1 could be further affected by cell/tissue/organ specificity. For example, mouse *SF1* transcripts were detected in brain and heart, implying their tissue-specific regulation at transcription level (Zhu et al., 2010). And SF1 was highly expressed in differentiated villous cells, but was not observed in adenoma or undifferentiated intestinal crypt cells of the intestinal epithelium (Miki et al., 2007). In plants, interestingly, SF1 has been found to be involved in a number of plant developmental processes and stress responses (Jang et al., 2014;Lee et al., 2017). In particular, SF1 has been observed to influence flowering time and leaf size in *Arabidopsis*, coinciding its relative high expression in flower

parts and leaves (Figure 6, left panel). Furthermore, transcripts of SF1 were unevenly distributed in several monocots and eudicots (Figures 6, S1 and S2), suggesting their potential role during plant development in these species.

In comparison to tissue specificity, more *cis*-elements involved in hormone and stress responses were observed within promoter regions of plant *SF1* genes (Figure 7 and Table S5), indicating their putative role in response to internal and external stimuli. The *Arabidopsis* SF1 has been demonstrated to participate in ABA signaling (Jang et al., 2014;Lee et al., 2017), coinciding the presence of an ABRE motif at its own 5'-flanking region (Figure 7). Furthermore, *Arabidopsis SF1* was induced by IAA at 1 hour after the treatment and repressed by MeJA (MJ). The AuxRR-core and CGTCA-motifs observed at its promoter region may be responsible for this regulation (Figure 7). However, experimental data is required to further strengthen this hypothesis in future functional investigations.

Composition of splice site determination complex reveals diverged mechanism to define exon-intron boundary among eukaryotes

In general, eukaryotic SF1s have similar molecular function to mediate early splice site recognition. Specifically, *Arabidopsis* SF1 has been proposed to have similar function in comparison to its yeast or metazoan counterparts (Jang et al., 2014;Lee et al., 2017). However, different eukaryotic organism may evolve their own recognition mechanism during early spliceosome assembly through SF1. First of all, the target BPS of SF1 is distinct in yeast in comparison to those sequences in animals and plants. In particular, yeast intronic BPS is a conserved seven-nucleotide sequence (UACUAAC), whereas the mammalian SF1 has been reported to bind more degenerate sequence (YNCURAY; N, any nucleotide; R, A or G; Y, C or U) (Keller and Noon, 1984). No conserved BPS has been observed in nematodes and plants at this stage (Lorkovi? et al., 2000;Long and H Robert, 2009). It poses the question of how SF1 recognize the BPS in these organisms and whether the additional RRM in plants contributes to this recognition (Jang et al., 2014). Second, the different coordinative mechanisms are present in a variety of organisms. For example, as the interaction partner of SF1 to coordinate 3' splice site recognition, mammalian U2AF65 interacts with U2AF small subunit (U2AF35). Similar interaction complex has been found in fission yeast, *S. pombe*, except for the small U2AF subunit named as U2AF23 (Tao et al., 2014). In contrast, the budding yeast lacks of a U2AF35-like small U2AF factor, the other two proteins (BBP/SF1 and Mud2p/U2AF65) are proposed to form a stable complex during splicing (Qiang et al., 2008). Furthermore, splicing reaction in animals requires the binding of U2AF65 to Py sequence downstream of BPS, while either of these two components are not necessary for yeast splicing (Abovich et al., 1994;Rutz and Seraphin, 1999). Intriguingly, plant showed a distinct splicing pattern in comparison to animals. For example, a high proportion of intron-retention events has been observed in plants, whereas exon skipping is the dominant AS type in animals (Qiang et al., 2008). And SF1 has been proposed to enhance splicing efficiency of introns containing weakly conserved 3' splice sites in *C. elegans* (Long et al., 2011). Therefore, it is

tempting to speculate that this difference may result from different SF1-centered splice site recognition between animals and plants.

Conclusions

In this work, we comprehensively identified 92 *SF1* sequences from 59 plant species, ranging from algae to eudicots. Subsequent phylogenetic and expression analysis have been carried out to elucidate the conservation and functional regulation of this gene family. By considering the connecting role of SF1 during splice site recognition, we hypothesize that plant SF1s may have overlap but also distinct function of their animal counterparts. Understanding the molecular mechanism of this protein family in plants provides intriguing possibility to manipulate crop traits through genetic control of plant splicing.

Declarations

FUNDING

This work was supported by the National Natural Science Foundation of China (NSFC31701341), the NJFU project funding (GXL2018005), the Natural Science Foundation of Guangdong Province (2018A030313030), the Natural Science Foundation of Hunan Province, Shenzhen Virtual University Park Support Scheme to CUHK Shenzhen Research Institute and Hong Kong Research Grant Council (AoE/M-05/12, AoE/M-403/16, GRF14160516, 14177617, 12100318).

AVAILABILITY OF DATA AND MATERIALS

The data sets are included within the article and its Additional files.

AUTHOR CONTRIBUTIONS

F.Y.Z., M.X.C., C.W. designed experiments. K.L.Z., Z.F., J.F.Y. performed experiments. K.L.Z., J.F.Y., Y.T., M.X.C. analysed data. K.L.Z., M.X.C. wrote the manuscript. G.F.H., G.F.Y., Y.M.F., J.H.Z. critically commented and revised it.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

COMPETING INTERESTS

The authors have no conflicts of interest to declare

References

Abovich, N., ., Liao, X.C., and Rosbash, M., . (1994). The yeast MUD2 protein: an interaction with PRP11 defines a bridge between commitment complexes and U2 snRNP addition. *Genes Dev* 8**,** 843-854.

Abovich, N., ., and Rosbash, M., . (1997). Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. *Cell* 89**,** 403-412.

Arning, S., ., Grüter, P., ., Bilbe, G., ., and Kr?Mer, A., . (1996). Mammalian splicing factor SF1 is encoded by variant cDNAs and binds to RNA. *Rna-a Publication of the Rna Society* 2**,** 794-810.

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., and Bental, N. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research* 44**,** W344-W350.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* 37**,** W202-W208.

Berglund, J.A., Chua, K., Abovich, N., Reed, R., and Rosbash, M. (1997). The Splicing Factor BBP Interacts Specifically with the Pre-mRNA Branchpoint Sequence UACUAAC. *Cell* 89**,** 781.

Chen, C., Rui, X., Hao, C., and He, Y. (2018). TBtools, a Toolkit for Biologists integrating various HTS-data handling tools with a user-friendly interface.

Chen, M.X., Zhu, F.Y., Wang, F.Z., Ye, N.H., Gao, B., Chen, X., Zhao, S.S., Fan, T., Cao, Y.Y., Liu, T.Y., Su, Z.Z., Xie, L.J., Hu, Q.J., Wu, H.J., Xiao, S., Zhang, J., and Liu, Y.G. (2019). Alternative splicing and translation play important roles in hypoxic germination in rice. *J Exp Bot* 70**,** 817-833.

Damian, S., Andrea, F., Stefan, W., Kristoffer, F., Davide, H., Jaime, H.C., Milan, S., Alexander, R., Alberto, S., and Tsafou, K.P. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research* 43**,** D447.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32**,** 1792-1797.

Finn, R.D., Jody, C., William, A., Miller, B.L., Wheeler, T.J., Fabian, S., Alex, B., and Eddy, S.R. (2015). HMMER web server: 2015 update. *Nucleic Acids Research* 43**,** 30-38.

García-Mayoral, M.F., Hollingworth, D., Masino, L., Díaz-Moreno, I., Kelly, G., Gherzi, R., Chou, C.F., Chen, C.Y., and Ramos, A. (2007). The Structure of the C-Terminal KH Domains of KSRP Reveals a Noncanonical Motif Important for mRNA Degradation. *Structure* 15**,** 485-498.

Garrey, S.M., Rodger, V., and J Andrew, B. (2006). An extended RNA binding site for the yeast branch point-binding protein and the role of its zinc knuckle domains in RNA binding. *Journal of Biological Chemistry* 281**,** 27443-27453.

Gibson, T.J., Thompson, J.D., and Heringa, J., . (1993). The KH domain occurs in a diverse set of RNA-binding proteins that include the antiterminator NusA and is probably involved in binding to nucleic acid. *Febs Letters* 324**,** 361-366.

Goodstein, D.M., Shengqiang, S., Russell, H., Rochak, N., Hayes, R.D., Joni, F., Therese, M., William, D., Uffe, H., and Nicholas, P. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* 40**,** D1178-D1186.

Gozani, O., Potashkin, J., and Reed, R. (1998). A potential role for U2AF-SAP 155 interactions in recruiting U2 snRNP to the branch site. *Mol Cell Biol* 18**,** 4752-4760.

Graveley, B.R. (2001). Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics* 17**,** 100-107.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*.

Guth, S., ., and Valcárcel, J., . (2000). Kinetic role for mammalian SF1/BBP in spliceosome assembly and function after polypyrimidine tract recognition by U2AF. *Journal of Biological Chemistry* 275**,** 38059-38066.

Haihong, S., and Green, M.R. (2006). RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans. *Genes & Development* 20**,** 1755-1765.

Hofberger, J.A., Nsibo, D.L., Govers, F., Bouwmeester, K., and Schranz, M.E. (2015). A complex interplay of tandem- and whole-genome duplication drives expansion of the L-type lectin receptor kinase gene family in the brassicaceae. *Genome Biol Evol* 7**,** 720-734.

Hu, B., Jin, J., Guo, A.Y., Zhang, H., Luo, J., and Gao, G. (2014). GSDS 2.0: an upgraded gene feature visualization server. *Bioinformatics* 31**,** 1296.

Ingham, R.J., Karen, C., Caley, H., Sabine, D., Lim, C.S.H., Joanna, Y., Kadija, H., Judith, R., Gerald, G., and Geraldine, M. (2005). WW domains provide a platform for the assembly of multiprotein networks. *Molecular & Cellular Biology* 25**,** 7092-7106.

Jana, K., Sophie, H.M., Angela, K.M., and Igor, V. (2004). Branch site haplotypes that control alternative splicing. *Human Molecular Genetics* 13**,** 3189-3202.

Jang, Y.H., Park, H.-Y., Lee, K.C., Thu, M.P., Kim, S.-K., Suh, M.C., Kang, H., and Kim, J.-K. (2014). A homolog of splicing factor SF1 is essential for development and is involved in the alternative splicing of pre-mRNA in Arabidopsis thaliana. *Plant Journal* 78**,** 591-603.

Johnson, L.S., Eddy, S.R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *Bmc Bioinformatics* 11**,** 431.

Keller, E.B., and Noon, W.A. (1984). Intron splicing: a conserved internal signal in introns of animal pre-mRNAs. *Proceedings of the National Academy of Sciences of the United States of America* 81**,** 7417-7420.

Kenichi, Y., Masashi, S., Yuichi, S., Daniel, N., Yasunobu, N., Ryo, Y., Yusuke, S., Aiko, S.O., Ayana, K., and Masao, N. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* 478**,** 64.

Kiana, T., Brady, S.M., Ryan, A., Eugene, L., and Provart, N.J. (2005). The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. *Plant Journal* 43**,** 153-163.

Kotake, Y., Sagane, K., Owa, T., Mimorikiyosue, Y., Shimizu, H., Uesugi, M., Ishihama, Y., Iwata, M., and Mizui, Y. (2012). Splicing factor SF3b as a target of the antitumor natural product pladienolide. *Nature Chemical Biology* 3**,** 570.

Lee, K.C., Yun, H.J., Kim, S.K., Park, H.Y., Thu, M.P., Lee, J.H., and Kim, J.K. (2017). RRM domain of Arabidopsis splicing factor SF1 is important for pre-mRNA splicing of a specific set of genes. *Plant Cell Reports* 36**,** 1-13.

Lescot, M., Dehais, P., Thijs, G., Marchal, K., Moreau, Y., Van De Peer, Y., Rouze, P., and Rombauts, S. (2002a). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res* 30**,** 325-327.

Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van De Peer, Y., Rouzé, P., and Rombauts, S. (2002b). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic acids research* 30**,** 325-327.

Lin, K., and Lu, R., Wy (2004). The WW domain-containing proteins interact with the early spliceosome and participate in pre-mRNA splicing in vivo. *Molecular & Cellular Biology* 24**,** 9176-9185.

Liu, Z., ., Luyten, I., ., Bottomley, M.J., Messias, A.C., Houngninou-Molango, S., ., Sprangers, R., ., Zanier, K., ., Kr?Mer, A., ., and Sattler, M., . (2001). Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science* 294**,** 1098-1102.

Long, M., and H Robert, H. (2009). Mutations in the Caenorhabditis elegans U2AF large subunit UAF-1 alter the choice of a 3' splice site in vivo. *Plos Genetics* 5**,** e1000708.

Long, M., Zhiping, T., Yanling, T., Sebastian, H., and H Robert, H. (2011). In vivo effects on intron retention and exon skipping by the U2AF large subunit and SF1/BBP in the nematode Caenorhabditis elegans. *Rna-a Publication of the Rna Society* 17**,** 2201-2211.

Lorkovi?, Z.J., Kirk, D.A., Wieczorek, Lambermon, M.H., and Filipowicz, W., . (2000). Pre-mRNA splicing in higher plants. *Trends in Plant Science* 5**,** 160-167.

Madhusudhan, M.S., Marti-Renom, M.A., Eswar, N., John, B., Pieper, U., Karchin, R., Shen, M.Y., and Sali, A. (2014). Comparative Protein Structure Modeling. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* 47**,** 5.6.1.

Makarova, O.V., Makarov, E.M., and Lührmann, R., . (2014). The 65 and 110 kDa SR-related proteins of the U4/U6.U5 tri-snRNP are essential for the assembly of mature spliceosomes. *Embo Journal* 20**,** 2553-2563.

Manceau, V., Swenson, M., Caer, J.L., Sobel, A., and Kielkopf, C., A (2010). Major phosphorylation of SF1 on adjacent Ser-Pro motifs enhances interaction with U2AF(65). *Febs Journal* 273**,** 577-587.

Margherita, C., Nicolas, A., Goranka, T., Mihaela, Z., and Angela, K.M. (2011). Analysis of in situ pre-mRNA targets of human splicing factor SF1 reveals a function in alternative splicing. *Nucleic Acids Research* 39**,** 1868.

Mazroui, R., ., Puoti, A., ., and Kr?Mer, A., . (1999). Splicing factor SF1 from Drosophila and Caenorhabditis: presence of an N-terminal RS domain and requirement for viability. *Rna-a Publication of the Rna Society* 5**,** 1615-1631.

Michaud, S., ., and Reed, R., . (1991). An ATP-independent complex commits pre-mRNA to the mammalian spliceosome assembly pathway. *Genes & Development* 5**,** 2534.

Miki, S., Yasuyoshi, N., Masashi, I., Kazufumi, H., Masaya, O., Setsuo, H., and Tesshi, Y. (2007). Involvement of splicing factor-1 in beta-catenin/T-cell factor-4-mediated gene transactivation and pre-mRNA splicing. *Gastroenterology* 132**,** 1039-1054.

Mo, C., and Manley, J.L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* 10**,** 741-754.

Mount, S.M., Pettersson, I., Hinterberger, M., Karmas, A., and Steitz, J.A. (1983). The U1 small nuclear RNA-protein complex selectively binds a 5' splice site in vitro. *Cell* 33**,** 509-518.

Noriko, H., Tomoko, A., David, F., and Tokio, T. (2007). Mutations in the SF1-U2AF59-U2AF23 complex cause exon skipping in Schizosaccharomyces pombe. *Journal of Biological Chemistry* 282**,** 2221-2228.

Qiang, W., Li, Z., Bert, L., and Rymond, B.C. (2008). A BBP-Mud2p heterodimer mediates branchpoint recognition and influences splicing substrate abundance in budding yeast. *Nucleic Acids Research* 36**,** 2787-2798.

Qun, P., Ofer, S., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* 40**,** 1413-1415.

Rain, J.C., Rafi, Z., ., Rhani, Z., ., Legrain, P., ., and Kr?Mer, A., . (1998). Conservation of functional domains involved in RNA binding and protein-protein interactions in human and Saccharomyces cerevisiae pre-mRNA splicing factor SF1. *Rna-a Publication of the Rna Society* 4**,** 551-565.

Reed, R. (2000). Mechanisms of fidelity in pre-mRNA splicing. *Curr Opin Cell Biol* 12**,** 340-345.

Rutz, B., and Seraphin, B. (1999). Transient interaction of BBP/ScSF1 and Mud2 with the splicing machinery affects the kinetics of spliceosome assembly. *RNA* 5**,** 819-831.

Sasaki-Haraguchi, N., Ikuyama, T., Yoshii, S., Takeuchi-Andoh, T., Frendewey, D., and Tani, T. (2015). Cwf16p Associating with the Nineteen Complex Ensures Ordered Exon Joining in Constitutive Pre-mRNA Splicing in Fission Yeast. *Plos One* 10**,** e0136336.

Selenko, P., Gregorovic, G., Sprangers, R., Stier, G., Rhani, Z., Krämer, A., and Sattler, M. (2003). Structural basis for the molecular recognition between human splicing factors U2AF65 and SF1/mBBP. *Molecular Cell* 11**,** 965-976.

Shih-Peng, C., Der-I, K., Wei-Yü, T., and Soo-Chen, C. (2003). The Prp19p-associated complex in spliceosome activation. *Science* 302**,** 282-281.

Shitashige, M., Satow, R., Honda, K., Ono, M., Hirohashi, S., and Yamada, T. (2010). Increased susceptibility of Sf1(+/-) mice to azoxymethane-induced colon tumorigenesis. *Cancer Science* 99**,** 1862-1867.

Sickmier, E.A., Frato, K.E., Shen, H., Paranawithana, S.R., Green, M.R., and Kielkopf, C.L. (2006). Structural Basis for Polypyrimidine Tract Recognition by the Essential Pre-mRNA Splicing Factor U2AF65. *Molecular Cell* 23**,** 49-59.

Siomi, H., ., Matunis, M.J., Michael, W.M., and Dreyfuss, G., . (1993). The pre-mRNA binding K protein contains a novel evolutionarily conserved motif. *Nucleic Acids Research* 21**,** 1193.

Staley, J.P., and Guthrie, C., . (1998). Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* 92**,** 315-326.

Tanackovic, G., and Kramer, A. (2005). Human splicing factor SF3a, but not SF1, is essential for pre-mRNA splicing in vivo. *Molecular Biology of the Cell* 16**,** 1366-1377.

Tao, H., Josep, V., and Query, C.C. (2014). Pre-spliceosome formation in S.pombe requires a stable complex of SF1-U2AF(59)-U2AF(23). *Embo Journal* 21**,** 5516-5526.

Tom, M., and Bosiljka, T. (2002). Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* 418**,** 236-243.

Uehara, T., Minoshima, Y., Sagane, K., Sugi, N.H., Mitsuhashi, K.O., Yamamoto, N., Kamiyama, H., Takahashi, K., Kotake, Y., and Uesugi, M. (2017). Selective degradation of splicing factor CAPERα by anticancer sulfonamides. *Nature Chemical Biology* 13**,** 675.

Valcárcel, J., ., Gaur, R.K., Singh, R., ., and Green, M.R. (1996). Interaction of U2AF65 RS region with pre-mRNA branch point and promotion of base pairing with U2 snRNA [corrected]. *Science* 273**,** 1706-1709.

Vincent, G., Olivier, G., Micheline, F.R., Alper, R., Alain, J., and Ulf, N. (2004). Nuclear retention of unspliced mRNAs in yeast is mediated by perinuclear Mlp1. *Cell* 116**,** 63-73.

Wahl, M.C., Will, C.L., and Reinhard, L. (2009). The spliceosome: design principles of a dynamic RNP machine. *Cell* 136**,** 701-718.

Wang, E.T., Rickard, S., Shujun, L., Irina, K., Lu, Z., Christine, M., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456**,** 470-476.

Wang, Y., Xu, L., Thilmony, R., You, F.M., Gu, Y.Q., and Coleman-Derr, D. (2016). PIECE 2.0: an update for the plant gene structure comparison and evolution database. *Nucleic Acids Research* 45**,** 1015.

Will, C.L., Schneider, C., ., Macmillan, A.M., Katopodis, N.F., Neubauer, G., ., Wilm, M., ., Lührmann, R., ., and Query, C.C. (2014). A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site. *Embo Journal* 20**,** 4536-4546.

Yuan, S., Chan, H.C.S., and Hu, Z. (2017). Using PyMOL as a platform for computational drug design. *Wiley Interdisciplinary Reviews Computational Molecular Science* 7**,** e1298.

Yun, Z., Tobias, M., Ivona, B., Thomas, K., Hyun-Seo, K., Peijian, Z., Nina, M.U., Sieber, S.A., Angela, K.M., and Michael, S. (2013). Structure, phosphorylation and U2AF65 binding of the N-terminal domain of splicing factor 1 during 3'-splice site recognition. *Nucleic Acids Research* 41**,** 1343-1354.

Zamore, P.D., Patton, J.G., and Green, M.R. (1992). Cloning and domain structure of the mammalian splicing factor U2AF. *Nature* 355**,** 609-614.

Zhu, F.Y., Chen, M.X., Ye, N.H., Shi, L., Ma, K.L., Yang, J.F., Cao, Y.Y., Zhang, Y., Yoshida, T., and Fernie, A.R. (2017). Proteogenomic analysis reveals alternative splicing and translation as part of the abscisic acid response in Arabidopsis seedlings. *Plant Journal* 91**,** 518-533.

Zhu, R., Heaney, J., Nadeau, J.H., Ali, S., and Matin, A. (2010). Deficiency of splicing factor 1 suppresses the occurrence of testicular germ cell tumors. *Cancer Research* 70**,** 7264-7272.

Figures

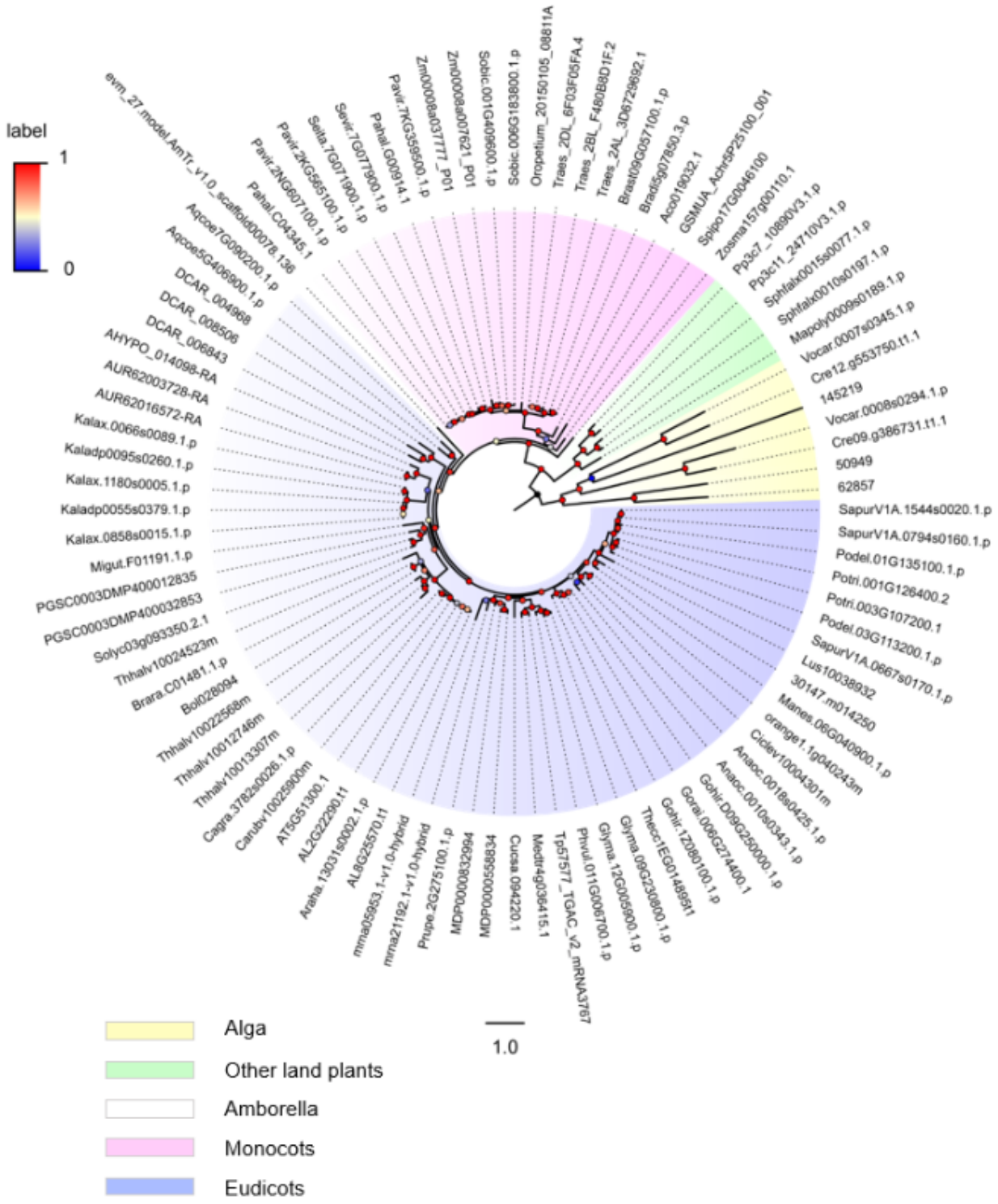


Figure 1

Circle phylogenetic tree representation of the available plant SF1 gene family. Phylogenetic analysis of plant SF1 gene family was carried out by using software PhyML v3.037 (maximum-likelihood, JTT+G model). Totally, 92 protein sequences from 59 plant species were chosen to calculate the phylogenetic relationship for tree construction. Bootstrap values are labeled at each major branch. The corresponding

information of each transcript such as species name, common name, number of identified transcripts and their transcript ID (nomenclature) are shown in Table S1 (taxonomies based on APG-Ⅳ system).

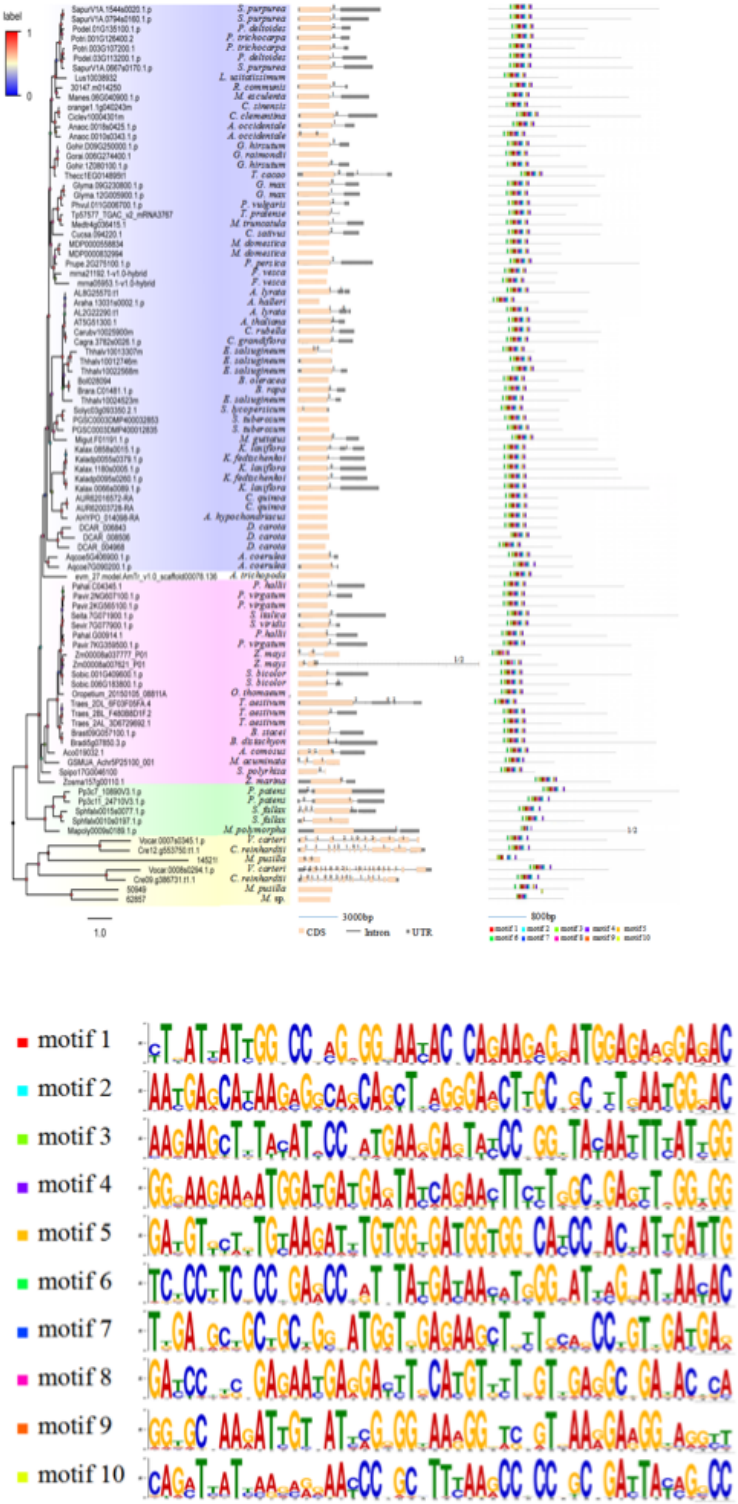


Figure 2

Gene structure comparisons and conserved motif identification among plant SF1 genes. Genomic organization (middle panel) and identified cDNA conserved motifs (right panel) by multiple Em for motif elicitation (MEME) analysis are shown against the vertical phylogenetic tree (left panel). Intron phase 0, 1

and 2 were showed on gene structure. The conserved sequence of ten identified motifs represented by different coloured boxes are listed below. The black vertical lines represent break at that particular branch. Some long genes were reduced to one half of the original to fit this picture.

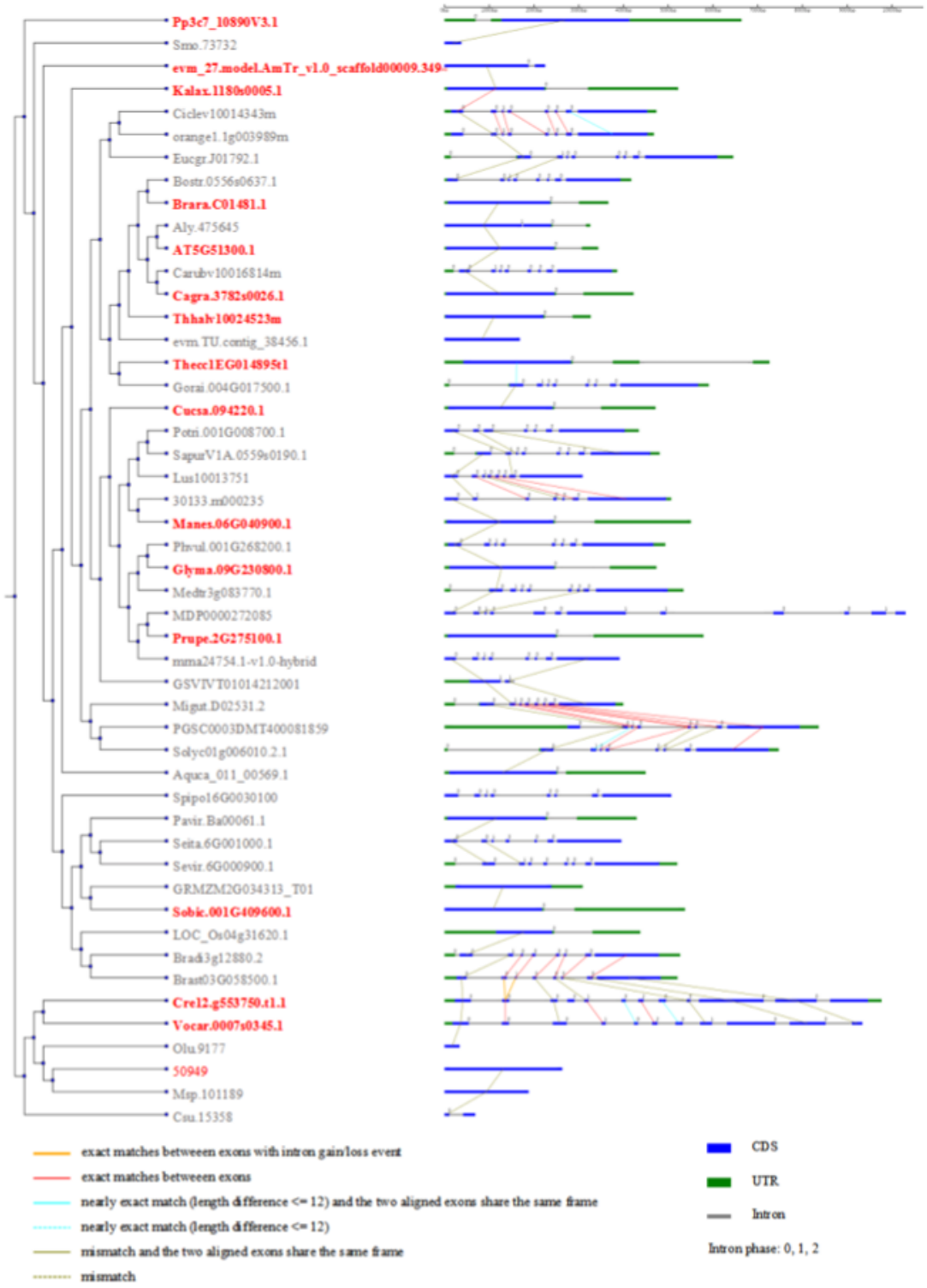


Figure 3

Analysis gene structure evolution with ortholog group of SF1 genes. Exon-intron structure and intron phase (right panel) are linked to the plant species tree (left panel). Genes with red color represent the

members of the plant SF1 genes. Different colored line means different exon comparison result between species.

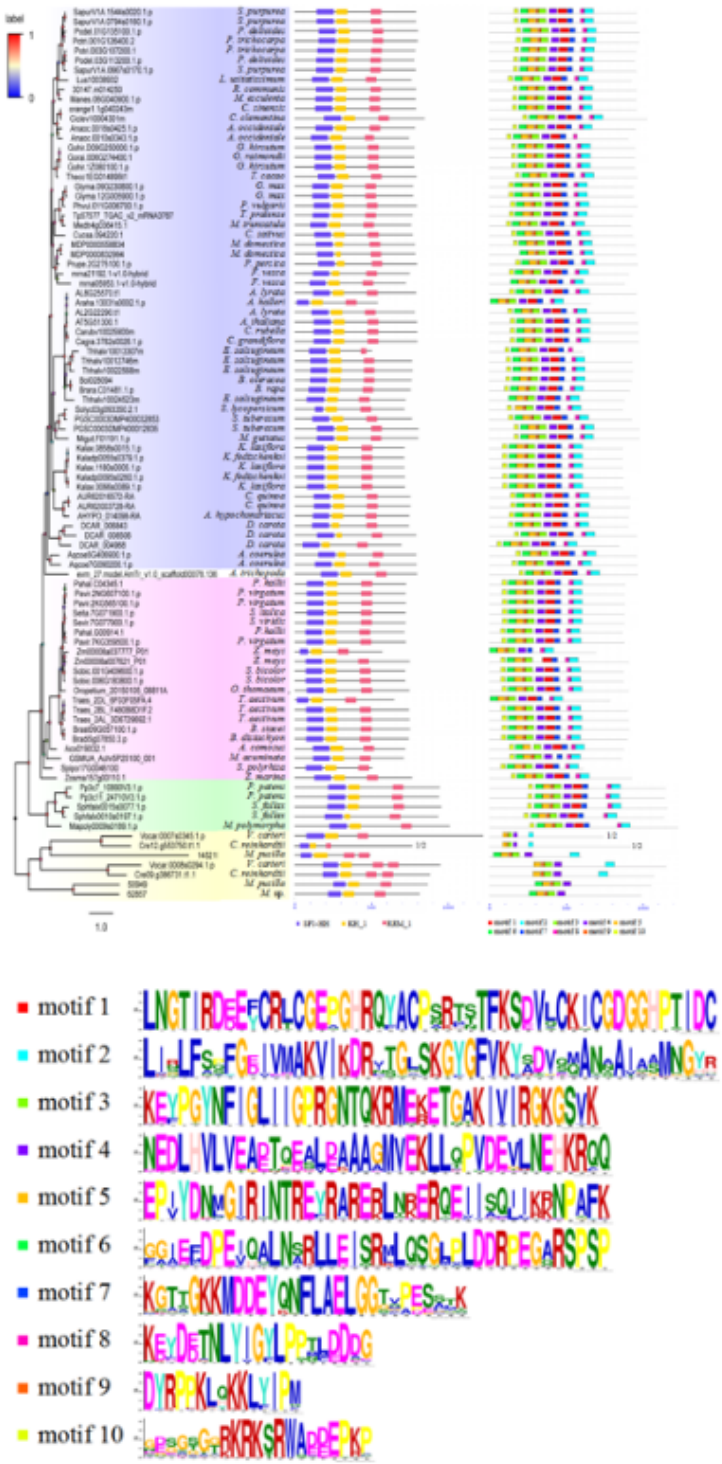


Figure 4

Comparisons of protein domain and conserved motif identification among plant SF1 genes. Protein domain (middle panel) and identified protein conserved motifs (right panel) by MEME analysis are shown against the vertical phylogenetic tree (left panel). The conserved sequence of ten identified motifs

represented by different coloured boxes are listed below. The black vertical lines represent break at that particular branch.

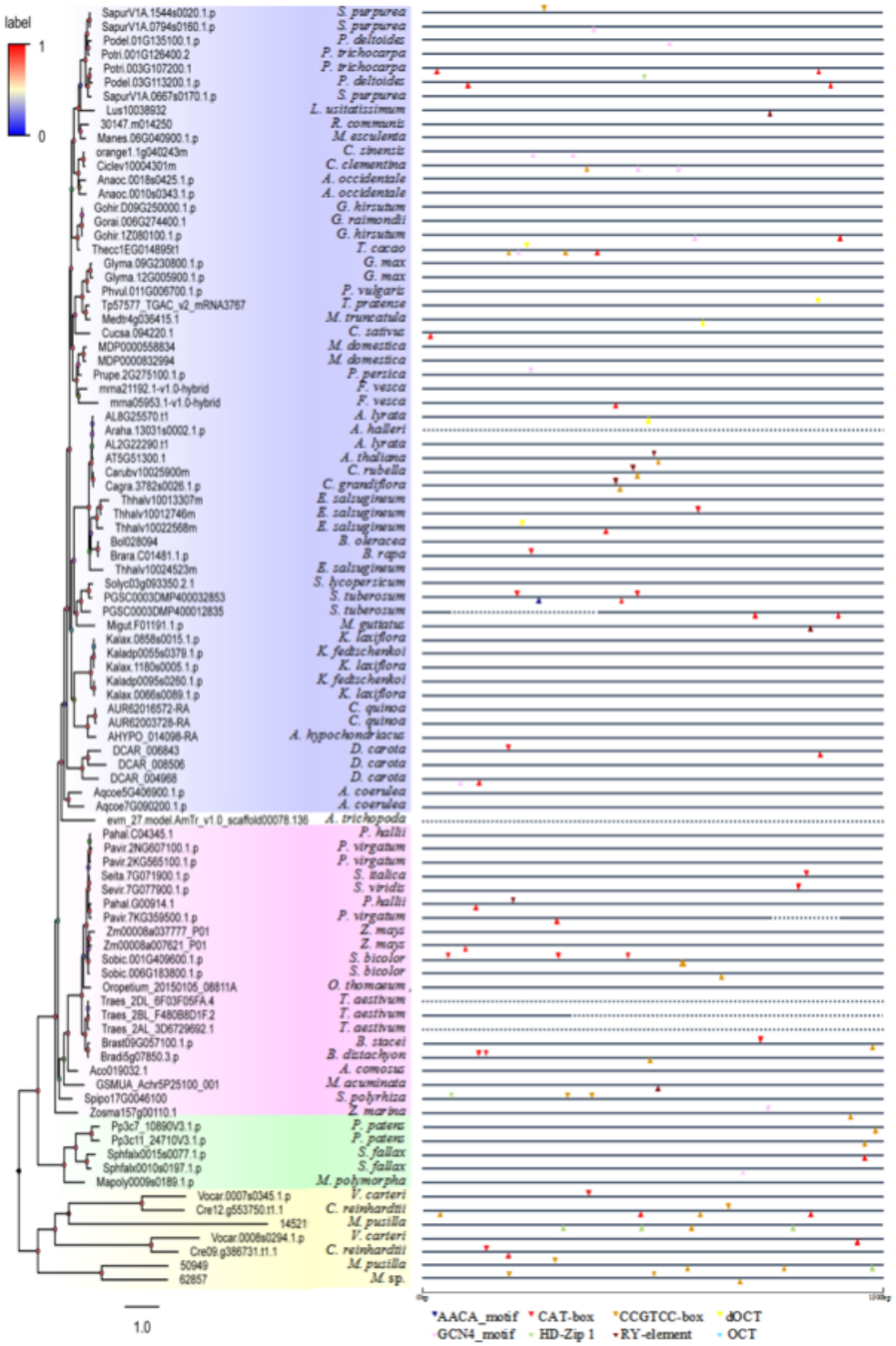


Figure 5

Analysis of motifs related tissue-specificity in the plant SF1 promoter regions. Eight identified motifs are represented by squares with various colors. These cis-acting motifs are labeled along the 1.5-kb promoter region (straight line) isolated from each plant SF1 gene according to their relative nucleotide positions to

transcript start sites. The full line represented the regions with explicit basic pairs. Gray dotted line represented the region of annexed base N or no sequences. Motifs at the positive strand are labeled above the line, whereas motifs at the negative strand are marked below the line. AACA-motif: involved in endosperm-specific negative expression; CAT-box: cis-acting regulatory element related to meristem expression; CCGTCC-box: cis-acting regulatory element related to meristem specific activation; dOCT: cis-acting regulatory element related to meristem specific activation; OCT: cis-acting regulatory element related to meristem specific activation; GCN4_motif: cis-regulatory element involved in endosperm expression; HD-Zip1: element involved in differentiation of the palisade mesophyll cells; RY-element: cis-acting regulatory element involved in seed-specific regulation. The black vertical lines represent break at that particular branch.

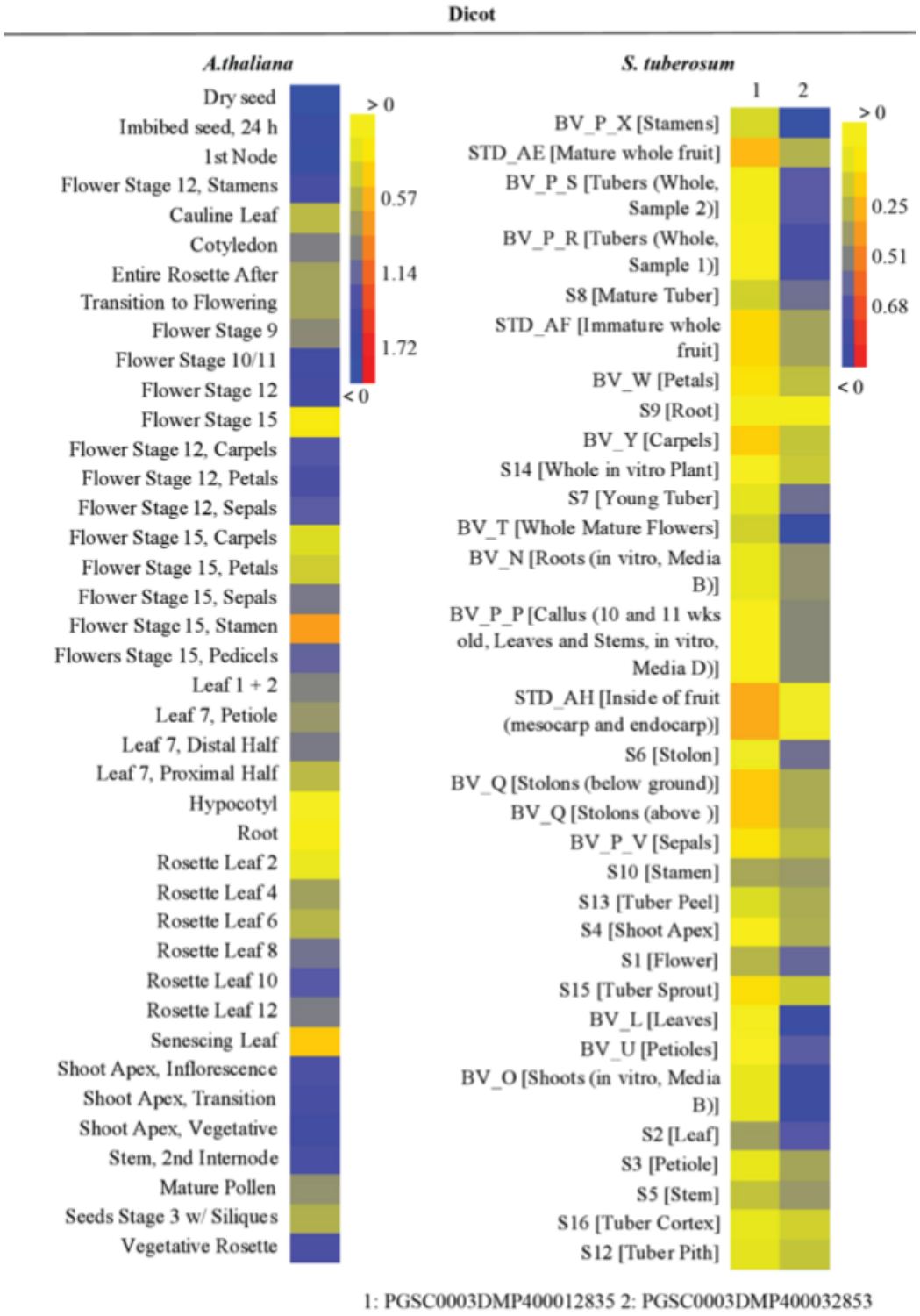


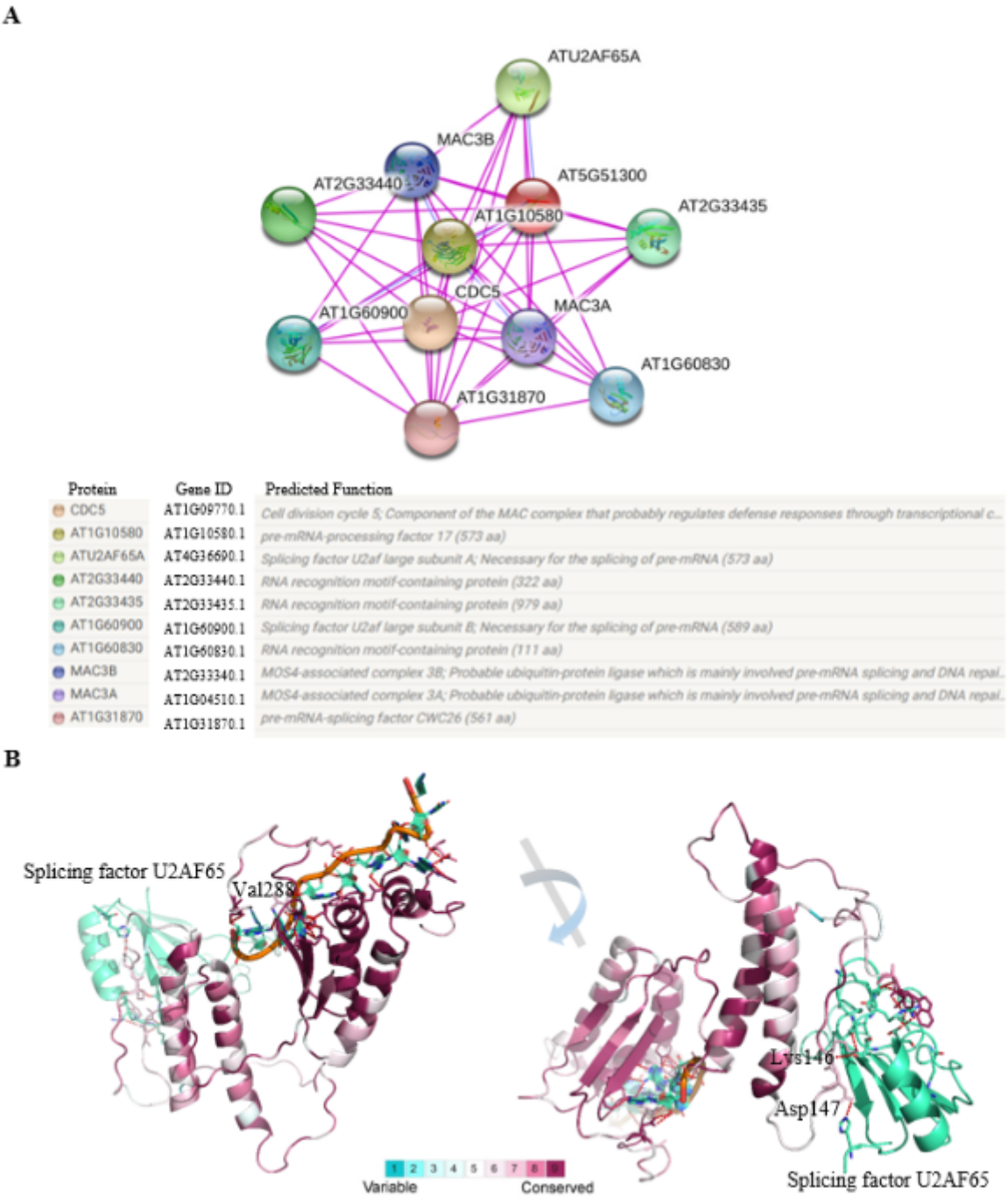
Figure 6

Expression patterns of Arabidopsis and Solanum tuberosum (potato) SF1 genes. Expression data were obtained from plant eFP browser microarray datasets and transformed by Log2 conversion and presented as heatmap. Red colour presents high levels of transcript abundance and blue presents low transcript abundance.

Figure 7

Analysis of motifs-related hormone and stresses in the plant SF1 promoter regions. Nineteen identified motifs are represented by symbols with different colors. These cis-acting motifs are labeled along the 1.5-kb promoter region (straight line) isolated from each plant SF1 gene according to their relative nucleotide positions to transcript start sites. The full line represented the regions with explicit basic pairs. Gray dotted line represented the region of annexed base N or no sequences. Motifs at the positive strand are

labeled above the line, whereas motifs at the negative strand are marked below the line. ABRE: cis-acting element involved in the abscisic acid responsiveness; ARE: cis-acting regulatory element essential for the anaerobic induction; AT-rich sequence: element for maximal elicitor-mediated activation (2copies); AuxRE: part of an auxin-responsive element; AuxRR-core: cis-acting regulatory element involved in auxin responsiveness; CGTCA-motif: cis-acting regulatory element involved in the MeJA-responsiveness; ERE: ethylene-responsive element; GARE-motif: gibberellin-responsive element; GC-motif: enhancer-like element involved in anoxic specific inducibility; LTR: cis-acting element involved in low-temperature responsiveness; MBS: MYB binding site involved in drought-inducibility; P-box: gibberellin-responsive element; TATC-box: cis-acting element involved in gibberellin-responsiveness; TCA-element: cis-acting element involved in salicylic acid responsiveness; TC-rich repeats: cis-acting element involved in defense and stress responsiveness; TGA-box: part of an auxin-responsive element; TGACG-motif: cis-acting regulatory element involved in the MeJA-responsiveness; TGA-element: auxin-responsive element; WUN-motif: wound-responsive element. The black vertical lines represent break at that particular branch.



Protein	Gene ID	Predicted Function
CDC5	AT1G09770.1	Cell division cycle 5; Component of the MAC complex that probably regulates defense responses through transcriptional c...
AT1G10580	AT1G10580.1	pre-mRNA-processing factor 17 (573 aa)
ATU2AF65A	AT4G36690.1	Splicing factor U2af large subunit A; Necessary for the splicing of pre-mRNA (573 aa)
AT2G33440	AT2G33440.1	RNA recognition motif-containing protein (322 aa)
AT2G33435	AT2G33435.1	RNA recognition motif-containing protein (979 aa)
AT1G60900	AT1G60900.1	Splicing factor U2af large subunit B; Necessary for the splicing of pre-mRNA (589 aa)
AT1G60830	AT1G60830.1	RNA recognition motif-containing protein (111 aa)
MAC3B	AT2G33340.1	MOS4-associated complex 3B; Probable ubiquitin-protein ligase which is mainly involved pre-mRNA splicing and DNA repai...
MAC3A	AT1G04510.1	MOS4-associated complex 3A; Probable ubiquitin-protein ligase which is mainly involved pre-mRNA splicing and DNA repai...
AT1G31870	AT1G31870.1	pre-mRNA-splicing factor CWC26 (561 aa)

Figure 8

Representative interaction network and conserved amino acid sequence analysis of plant SF1s. (A) Interaction network of Arabidopsis (AT5G51300) based on experimental data. Each network node represents all the proteins produced by a single, protein-coding gene locus. Different colored nodes represent query proteins and first shell of interactors. Filled nodes present some 3D structure is known or predicted, while empty nodes present proteins of unknown 3D structure. Edges represent protein-protein associations that meant to proteins jointly contribute to a shared function. (B) Conserved domains of plant SF1s. The 3-D structure of plant SF1 were generated according to the Arabidopsis sequence

(AT5G51300) and represented with their target RNA. The ribbon representation is colored according to ConSurf Grade (1-blue to 9-purple) by using all identified protein sequences of plant SF1s.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigureS4.jpg
- SF1TableS5promotermotifs.pdf
- SF1TableS1S4.xlsx
- SF1figuresS1S3.docx