Title: Synopsis of The SOFL Plant-Specific Gene Family

Reuben Tayengwa*†1, Jianfei Zhao‡, Courtney F. Pierce†2, Breanna E. Werner†3, Michael M. Neff*†

*Program in Molecular Plant Sciences, Washington State University, Pullman, WA, USA.
†Department Crop and Soil Sciences, Washington State University, Pullman, WA, USA.
‡Department of Biology, University of Pennsylvania, Philadelphia, PA, USA.

Present addresses

1University of Maryland, Plant Sciences and Horticultural Landscape Department, College Park, MD 20742.
2Colorado State University, Department of Animal Sciences, Fort Collins, CO 80523.
3Washington State University College of Nursing, Spokane, WA 99202.
Running Title: The *SOFL* gene family

Keywords


For correspondence (e-mail mmneff@wsu.edu)

Address: 387 Johnson Hall, PO Box 646420, Pullman WA 99164-6420 USA

Phone: 509-335-7705, Fax: 509-335-8674
ABSTRACT

SUPPRESSOR OF PHYB-4\#5 DOMINANT (sob5-D) was previously identified as a suppressor of the phyB-4 long-hypocotyl phenotype in Arabidopsis thaliana. Overexpression of SOB5 conferred dwarf phenotypes similar to those observed in plants containing elevated levels of cytokinin (CK) nucleotides and nucleosides. Two SOB-FIVE- LIKE (SOFL) proteins, AtSOFL1 and AtSOFL2, which are more similar at the protein level to each other than they are to SOB5, conferred similar phenotypes to the sob5-D mutant when overexpressed. We used founding SOFL gene family members to perform database searches and identified a total of 289 SOFL homologues in sequenced genomes of 89 angiosperm species. Phylogenetic analysis results implied that the SOFL gene family emerged during the expansion of angiosperms and later evolved into four distinct clades. Among the newly identified gene family members are four previously unreported Arabidopsis SOFLs. Multiple sequence alignment of the 289 SOFL protein sequences revealed two highly conserved domains; SOFL-A and SOFL-B. Overexpression and site-directed mutagenesis studies demonstrated that the SOFL domains are necessary for SOB5 and AtSOFL1’s overexpression phenotypes. Examination of the subcellular localization patterns of the founding Arabidopsis thaliana SOFLs suggested they may be localized in the cytoplasm and the nucleus. We have discovered that SOFLs are a plant-specific gene family characterized by two conserved domains that are important for function.

INTRODUCTION

Many genes can be grouped into specific families based on nucleotide and protein sequence similarity. Such gene families have arisen as a result of duplication and expansion of individual members (Wang et al. 2012). Gene family sizes vary across different lineages and may have
important functional outcomes related to adaptation or speciation (Qu and Zhu 2006; Vollrath et al. 1998). Functional and evolutionary studies of genes belonging to various families have greatly enhanced our understanding of components involved in plant growth and development (Kim et al. 2007; Higuchi et al. 2004; Xie et al. 1999; Kim 2006; Kondou et al. 2010; Zhao et al. 2014). Despite significant progress made towards discovery and curations of many gene families, it is estimated that the majority of plant gene families remain uncharacterized (Guo 2012). Therefore, the continued identification and study of hitherto uncharacterized gene families remains crucial for our continued endeavor to eventually achieve the goal of functionally characterizing most of the genes.

The *Arabidopsis thaliana* activation-tagged mutant *sob5-D* (*SUPPRESSOR OF PHYB-4#5 DOMINANT*) was previously identified as a suppressor of the long-hypocotyl phenotype conferred by a *phyB-4* mutation (Zhang et al. 2006). Database searches using SOB5 (At5g08150) protein sequence as a query revealed two SOB-FIVE- LIKE (SOFL) proteins in *Arabidopsis*, AtSOFL1 (At1g26210) and AtSOFL2 (At1g68870) (Zhang et al. 2006). AtSOFL1 and AtSOFL2 are more similar to each other than they are to SOB5 at the protein level (Zhang et al. 2006, 2009). When individually overexpressed, AtSOFL1 and AtSOFL2 conferred phenotypes similar to the *sob5-D* activation-tagged mutant (Zhang et al. 2006, 2009). Transgenic lines individually overexpressing *SOB5, AtSOFL1* and *AtSOFL2* genes displayed reduced apical dominance, and conferred smaller rosette leaves, retarded root growth and delayed senescence phenotypes (Zhang et al. 2006, 2009). In addition, these transgenic lines also contained elevated levels of cytokinin (CK) nucleotides and nucleosides; *trans*-zeatin riboside (tZR), *trans*-zeatin riboside monophosphate (tZRMP), and $N^6$-($\Delta^2$-isopentenyl)adenosine monophosphate (iPRMP) (Zhang et
al. 2006, 2009). These results suggested that the founding SOFL gene family members may have similar or overlapping CK-related functions. However, the specific details of what roles, if any, intermediate CK nucleotide/nucleoside species play during plant growth and development are not fully understood and will have to be further investigated in the future.

Zhang et al., (2006), reported that the three founding SOFLs constituted a small, novel and intron-less Arabidopsis thaliana gene family. In addition, SOFL homologues and expressed sequence tags were also identified in a few monocotyledonous and dicotyledonous species; Malus x domestica, Brassica napus, Oryza sativa and Populus trichocarpa (Zhang et al. 2006). Due to limited sequenced genomes at the time it was difficult to obtain a comprehensive phylogenetic overview of the SOFL gene family. Nonetheless, based on these preliminary results we speculated that SOFL homologues also existed in genomes of other eukaryotic, prokaryotic, and or mammalian species (Zhang et al. 2006).

Since Arabidopsis SOFLs were first identified, more genome sequence information has been released (Fang et al. 2013; Wheeler et al. 2008; Kersey et al. 2016). We have taken a phylogenetic approach to update our understanding of the evolution and diversification of the SOFL gene family. Database searches using the three founding SOFL members’ proteins sequences as queries revealed at least 289 SOFL homologues in 89 angiosperm species. During this process, we also discovered that the Arabidopsis thaliana genome contains four additional and more divergent SOFLs that were previously unreported. In addition, phylogenetic analysis showed that the 289 SOFL homologues evolved into four main clades. Finally, to generate preliminary data as well as gain some insights into this gene family, we performed subcellular
localization, tissue expression and domain mutational studies using the three founding SOFL members. This updated synopsis of the SOFL gene family provides basic background information needed to design future studies to eventually fully characterize this novel gene family.

MATERIALS AND METHODS

Plant materials and growth conditions

All Arabidopsis thaliana lines used in this manuscript are in the Columbia (Col-0) background. For phenotypic analysis, all plants were grown on soil in pots in growth chambers set at 21°C with white light (200 μmol m$^{-2}$ sec$^{-1}$) and 60-70 % humidity, and under 16 hours light and 8 hours darkness.

Sequence alignment and phylogenetic analysis

SOB5, AtSOFL1 and AtSOFL2 protein sequences were downloaded from NCBI (Lamesch et al. 2012; Goodstein et al. 2011; Wheeler et al. 2008; Sayers et al. 2009). The sequences were used as queries to search NCBI and Phytozome databases for SOFL homologues using BLASTP option using a < 1 X 10-2 cut-off value. The amino acid sequences were aligned using Probalign V1.3 (Roshan 2014) on CIPRES Science Gateway Server (Miller et al. 2015). Bayesian inference analysis was performed with the Mr. Bayes 3.2.1 on XSEDE tool on CIPRES Science Gateway for 30 million generations with 8 chains to run. Generations were sampled every 10,000 generations with the first 25% generations were used as a burn-in.

Sequence logo analysis
Sequence logo analysis of conserved SOFL-A and SOFL-B domains was performed using the online WebLogo tool (Crooks et al. 2004). 289 SOFL sequences retrieved from NCBI (Sayers et al. 2009) and Phytozome (Goodstein et al. 2011) were used to generate the sequence logo using default settings.

**Overexpression and point-mutation analysis of SOFL domains in SOB5 and AtSOFL1**

Site-directed point mutations were generated using the QuikChange® Lightning mutagenesis kit according to manufacturer’s instructions (Agilent Technologies). Mutagenesis primers were designed using Agilent’s web-based QuikChange® Primer Design Program [http://www.genomics.agilent.com/primerDesignProgram.jsp](http://www.genomics.agilent.com/primerDesignProgram.jsp) (Table S1). Gateway® compatible entry vectors, pENTR223 (ABRC), carrying SOB5 and AtSOFL1 coding sequences, were used as templates in mutagenesis PCR reactions. Each point mutation was confirmed via sequencing.

Once mutagenesis was confirmed, SOB5 or AtSOFL1 coding sequences carrying point mutations were cloned via LR® reactions into pCHF3 (Zhang et al. 2006) and pEarlyGate100 (Earley et al. 2006) binary vectors, respectively. Destination binary vectors carrying mutated SOB5 and AtSOFL1 genes under control of the constitutive 35S promoter were transformed into Arabidopsis thaliana Col-0 wild-type plants using the floral dip method and the Agrobacterium tumafaciens strain GV3101 (Clough and Bent 1998). Transformants were screened on 1/2× Linsmaier and Skoog media plates containing 25 mg L⁻¹ Basta (pEarleyGate 100) and 30 mg L⁻¹ kanamycin (pCHF3). Homozygous lines were identified in the T₃ generation. Three independent transgenic lines representing each construct were selected for analysis. All plants were grown and analyzed in a growth chamber.
**RNA extraction and real-time PCR**

Total RNA was extracted from tissues pooled from three independent plants (n=3 for all tissues) of six-day old seedlings, 10-day old juvenile plants, rosette leaves from a 20-day old plant, floral tissue, siliques, a 25-day old adult plant, and roots from a 25-day old plant using the Qiagen® RNeasy Mini Kit (Qiagen). On-column DNase digestion was performed using the RNase-Free DNase Set (Qiagen, Valencia, CA). Complementary DNA (cDNA) was further generated using the iScript Reverse Transcription Supermix for RT-qPCR (Bio Rad, Hercules, CA). Identical amounts of cDNA input were used in PCR reactions using either ubiquitin10 (UBQ10) or gene-specific primers (Table S1). Amplification using the UBQ10 primers was done in 30 cycles while for gene-specific primers 45 cycles were used. Negative control, no reverse transcriptase (No RT), samples were prepared using the same reaction conditions and reagents (minus reverse transcriptase enzyme) used to make cDNA.

**CFP-SOFL fusion constructs and onion bombardment**

To generate pSAT4-CFP-SOB5, pSAT4-CFP-AtSOFL1, and pSAT4-CFP-AtSOFL2 N-terminal fusion constructs, respective full-length coding sequences, each contained in the entry vector pENTR223 (ABRC), were cloned into pSAT4-CFP (ABRC) destination vector via Gateway® LR reactions (Invitrogen, Carlsbad, CA). Onion epidermal layers were co-bombarded with pSAT6-mRFP (ABRC) plasmid and pSAT4-CFP constructs carrying SOFLs fused to CFP using a PDS-1000/He Biolistic transformation system (Bio Rad, Hercules, CA). Bombarded onion epidermal layers were incubated in the dark for 40 hours. To identify successfully transformed cells and observe fluorescent signals, the onion epidermal layer was examined on a Leica TCS SP8 X (Leica Microsystems, Mannheim, Germany) confocal microscope.
**Data availability**

Genetic material used in this manuscript is available upon request. Primers used in the study are listed in Table S1. Total numbers of SOFL homologues identified in various species are listed in Table S2. Table S3 shows results from protein localization prediction analysis. Alternatively-spliced AtSOFL1 product sequences (AtSOFL1.1 and AtSOFL1.2) are provided in the supplemental file (File S2). Figure S1 shows a cartoon and gel image of AtSOFL1’s alternative splicing products. Sequences extracted from our database searches using the founding SOFL protein sequences as queries are provided in Supplemental file.

**RESULTS**

**SOFLs are a plant-specific gene family**

To identify SOFL homologues, we performed BLASTP searches against the NCBI database (Sayers et al. 2009) using SOB5, AtSOFL1 and AtSOFL2 sequences as queries. We extracted 289 SOFL homologues from 89 flowering plant species, of which 15 were monocotyledons and 74 were dicotyledons (Table S2). To gain insight into the evolutionary history of the SOFL gene family, we next performed phylogenetic analysis using the 289 SOFL protein sequences extracted via database searches. Our analysis suggested that SOFLs likely evolved into four major clades: Clade-I, II, III and IV (Figure 1a and 1b). We hypothesize that SOFLs within each clade have greatly expanded through evolution, with Clades-I and -IV showing the largest expansion (Figure 1a and 1b). However, since the tree was constructed only with the existing SOFL genes in the plant species examined and no outgroup used, we cannot rule out gene loss events. Annotation data from NCBI (Sayers et al. 2009), Phytozome (Goodstein et al. 2011) and
individual species genome databases revealed that 60 SOFL homologues contained introns and another 226 did not (Table S2). There was no annotation data for *Arabis alpina* and *Erythranthe guttata* (*Mimulas guttatus*) SOFL homologues.

**Arabidopsis thaliana genome contains seven SOFL members.**

It was previously reported that the *Arabidopsis* genome contained three SOFL members (Zhang et al. 2006, 2009). However, latest database search results revealed that the *Arabidopsis thaliana* genome contained a total of seven SOFL members (File S1). In addition, we also identified seven SOFLs in *Arabidopsis lyrata* (File S1). We have designated the newly identified *Arabidopsis thaliana* SOFL genes as *AtSOFL3* (AT3G30580), *AtSOFL4* (AT5G38790), *AtSOFL5* (AT4G33800) and *AtSOFL6* (AT1G58460). *SOB5* is in Clade VI of the phylogenetic tree (Figure 1a and 1b), compared to *AtSOFL1* and *AtSOFL2* which were in Clade I. Interestingly, the newly identified SOFL members, *AtSOFL3*, *AtSOFL4*, *AtSOFL5* and *AtSOFL6* are in Clade IV (Figure 1b), suggesting that genetic redundancy may exist among these genes and SOB5. All *Arabidopsis* SOFLs are intron-less except *AtSOFL1* (recently classified by TAIR as intron-containing), *AtSOFL5* and *AtSOFL6*.

We next assessed the expression patterns of *Arabidopsis thaliana* SOFL homologues. RNA was extracted from seedlings, juvenile plants, adult rosette leaf, floral structures, siliques, an entire flowering plant and adult plant roots. *SOB5*, *AtSOFL1* and *AtSOFL2* transcripts were detected in all samples tested except roots (Figure 2). *AtSOFL3*, *AtSOFL4*, *AtSOFL5* and *AtSOFL6* were expressed at varying levels in seedlings, juvenile plants, flowers, siliques and whole plant (Figure 2). *AtSOFL3*, *AtSOFL4*, *AtSOFL5* were the only ones detected in roots. *AtSOFL6*
showed little to no expression in rosette leaves and roots, and similarly, *AtSOFL5* showed little
to no expression in seedlings and rosette leaves (Figure 2). *AtSOFL4* was expressed in all
samples tested but had the lowest transcript levels overall. The varying transcript accumulation
levels and differential tissue expression patterns may suggest unique functional roles among the
seven *Arabidopsis thaliana* SOFLs during plant development (Figure 2). Surprisingly, during
efforts to amplify a full-length *AtSOFL1* (At1g26210) cDNA we identified a previously
unreported splice variant. We have designated the original 447 bp transcript *AtSOFL1*
(At1g26210.1), and the newly identified 771bp second transcript as (At1g26210.2) (Figure S1a,
b).

**SOFL homologues are characterized by two conserved domains**

Multiple sequence alignment analysis is a useful tool used to infer relationships among gene
family members. Alignment data can provide functional information through the identification of
key conserved domains and other important features. We aligned SOFL homologues protein
sequences and revealed two conserved domains in the N-terminal region (Figure 3a), which were
previously reported by Zhang et al. (2006), albeit based on fewer protein sequences. We have
designated the two conserved domains, SOFL-A and SOFL-B (Figure 3a). When only
*Arabidopsis* SOFL sequences were aligned (Figure 3b) and compared to an alignment of all 289
SOFL homologues (Figure 3c, 3d) we observed slight differences in the SOFL-B domain. In an
alignment of *Arabidopsis* SOFL sequences only, SOFL-B domain contains eight 100%
conserved residues SM×SDASS×P (Figure 3b). However, the same domain only contains four
100% conserved residues (S××SDA) when all 289 SOFLs homologous sequences were aligned
(Figure 3d). In general, based on all SOFL orthologue sequences identified so far, SOFL-A
domain contains an SGWT\(\times\)Y motif and SOFL-B domain contains an S\(\times\)SDA motif (\(\times\) = amino acid residue that is not 100% conserved) (Figure 3c, d). There were no areas of high sequence similarity in the C-terminal region, except for an increased presence of basic amino acid residues in most SOFL homologues (Table S3).

Conserved amino acid residues in SOFL domains are required for the manifestation of

**SOB5 and AtSOFL1 overexpression phenotypes.**

Previously, Zhang et al., (2009), investigated the requirement of some of the 100% conserved residues in SOFL-A and SOFL-B domains for the manifestation of AtSOFL2’s overexpression phenotype via site-directed mutagenesis. Constructs carrying AtSOFL2 gene harboring individual point mutations in each of the two conserved domains were used to generate plants overexpressing an aberrant protein. Resultant transgenic mutant plants overexpressing AtSOFL2 gene harboring T21I and D80N point mutations lost the phenotype that is typically associated with the overexpression of the wild type AtSOFL2 gene (Zhang et al. 2009). These results implied that the conserved amino acid residues were important for the manifestation of AtSOFL2 overexpression phenotype.

To further investigate the biological importance of some of the 100% conserved residues in the other two founding SOFLs, SOB5 and AtSOFL1, we generated a series of site-directed point mutations in their respective SOFL-A and SOFL-B domains. Wild-type Arabidopsis thaliana plants were separately transformed with constructs carrying mutations in SOB5 (T21I and P61R) and AtSOFL1 (T23I and P84R) coding sequences. Transgenic plants overexpressing SOB5 (T21I, P61R) and AtSOFL1 (T23I, P84R) mutated genes lost the dwarf/semi-dwarf phenotypes
typically observed when wild-type versions were overexpressed (Figure 4a, b). We also performed site-directed mutagenesis on non-conserved amino acid residues; D33H and D53H for SOB5 and AtSOFL1, respectively, which are not located in the two conserved domains (Figure 3b). Transgenic plants overexpressing mutated SOB5 (D33H) and AtSOFL1(D53H) exhibited phenotypes similar to those observed when wild-type genes are constitutively expressed (Figure 4a, b) (Zhang et al. 2006, 2009). These data suggest that less conserved amino acid residues are not required for the observed phenotypes.

**SOFL subcellular localization**

We used two publicly available online prediction programs, Wolf PSORT (Horton et al. 2007) and SeqNLS (Lin and Hu 2013), to gain insights into subcellular localization patterns of the 289 SOFL homologues. Wolf PSORT converts protein amino acid sequences into numerical localization features based on sorting signals, amino acid composition and functional motifs such as DNA-binding motifs (Horton et al. 2007). SeqNLS uses a sequential pattern mining algorithm to effectively identify potential nuclear localization signals (NLS) in protein sequences (Lin and Hu 2013). Wolf PSORT predicted 274 SOFL homologues to localize to the nucleus, with the remainder predicted to localize to chloroplast, cytoplasm, mitochondria, peroxisome, Golgi and extracellular space (Table S3). In contrast, the SeqNLS program detected nuclear localization signals in only 156 SOFLs with a statistically significant prediction score of at least 0.89 (Table S2) (Lin and Hu 2013). Interestingly SeqNLS did not detect nuclear localization signals in 13 SOFLs which were predicted to localize to the nucleus by Wolf PSORT (Table S3). 119 SOFLs scored below the default cutoff threshold to be classified as containing a NLS. Overall, at least
156 SOFL homologues were predicted to localize to the nucleus by both Wolf PSORT and SeqNLS algorithms (Table S3).

To further assess some of the subcellular prediction data, we examined the intracellular distribution of the founding Arabidopsis thaliana SOFL members. We fused SOB5, AtSOFL1 and AtSOFL2 to the carboxyl end of a cyan fluorescent protein (CFP) and transiently expressed the fusion proteins in onion epidermal cells. CFP-SOB5, CFP-AtSOFL1 and CFP-AtSOFL2 fluorescent signals were observed in both the cytoplasm and nucleus of the onion epidermal cells (Figure 5). We also examined whether the fluorescent signals of all three fusion proteins were localized to the cytosol and not the cell wall. This was achieved by inducing plasmolysis through exposure of onion epidermal peels to 0.8 M mannitol. Ensuing the mannitol treatment, CFP signal remained restricted to the edges of the plasmolyzed and plasma membrane and or cytoplasm, suggesting that CFP-SOB5, CFP-AtSOFL1 and CFP-AtSOFL2 fusion proteins were not localized to the cell wall.

DISCUSSION

SOFL homologues emerged in angiosperms

The founding members of the SOFL gene family were first identified via an activation tagging screen in Arabidopsis thaliana (Zhang et al. 2006). Based on the limited number of sequenced genomes available at the time, it was initially concluded that SOB5, AtSOFL1 and AtSOFL2 were a small three-member intron-less Arabidopsis thaliana gene family. To expand on earlier studies and acquire a basic level understanding of this poorly characterized gene family, we took advantage of an increasing number of sequenced genomes to perform database searches using
the founding *Arabidopsis thaliana* founding SOFL members as queries. We have retrieved a total of 289 SOFL sequences from 89 angiosperm species (Table S2). However, we cannot rule out the possibility that SOFL homologues are present in unsequenced organisms or were lost in other genomes. In addition, we expect additional SOFL homologues to be identified in other species as more genomes are sequenced. Furthermore, as genome sequence annotation improves, it is possible that some of the SOFL homologues currently classified as intron-less or intron-containing may be re-categorized in the future.

No SOFL homologues were identified in *Volvox carteri* (green algae), *Chlamydomonas reinhardtii* (green algae), *Ostreococcus lucimarinus* (single-celled water algae), *Micromonas pusilla* (water algae), *Physcomitrella patens* (non-vascular bryophyte, moss) and *Selaginella moellendorffii* (member of an ancient vascular plant lineage) (Lamesch et al. 2012), strongly suggesting that SOFLs may have emerged after the separation of lycophytes, pterophytes and seed plants. Out of the four main plant groups: bryophytes, seedless vascular (pteridophytes/lycophytes), gymnosperms and angiosperms, SOFLs have so far only been identified in angiosperms. We hypothesize that the SOFL gene family emerged and expanded during the evolution of flowering plant species.

**SOFL gene family is comprised of intron-containing and intron-less genes**

Approximately 20% of SOFL homologues identified so far contain introns and 80% are intron-less (Table S2). These results are inconsistent with overall data from rice and *Arabidopsis* genomes which revealed the presence of only 19.9% and 21.7% intron-lacking genes, respectively (Jain et al. 2008). Our latest database search results also showed that *Arabidopsis*
*thaliana* contains seven SOFL genes, in contrast to previous reports of three intron-less family members (Zhang et al. 2006, 2009). Out of the four newly discovered *Arabidopsis* SOFLs, only *AtSOFL5* and *AtSOFL6* contain introns, a departure from the previous designation of the *Arabidopsis* SOFL family as being comprised of members lacking introns (Zhang et al. 2006, 2009). Genes lacking introns are a characteristic of prokaryotes and are a useful resource for studying the evolution of gene architecture in eukaryotes, but information on their biological significance remains limited (Yan et al. 2014). Considering that SOFLs seem to have emerged during the evolution of angiosperms, it is surprising that the majority of the genes in this family are intron-less, a trait expected in early species (Zou et al. 2011). On the other hand, this result can be explained by the fact that majority of the intron-less SOFLs could have potentially arisen because of gene duplication events (Yan et al. 2016; Lecharny et al. 2003). Gene prediction and gene functional studies data suggests that intron-less genes may play unique roles in growth and development, including translation and energy metabolism in maize, rice and *Arabidopsis* as well as cell envelope and amino acid biosynthesis in rice and *Arabidopsis* (Yan et al. 2014; Jain et al. 2008). In addition, gain-of-function studies from Zhang et al. (2006, 2009) suggested that the three-founding intron-lacking SOFLs may be involved in CK-related functions. According to Zhao et al. (2014) the presence of introns enhances the transcription of associated genes. Therefore, to further explore the biological significance of intron-less genes, studies that include gain-of-function, loss of function, gene expression pattern, gene expression level analysis and subcellular localization will need to be performed in the future.

**SOFL-A and SOFL-B domains are important for SOB5 and AtSOFL1’s overexpression phenotypes**
Previously, Zhang et al. (2009) demonstrated via site-directed mutagenesis experiments that certain amino residues in the SOFL domains were necessary for the manifestation of \textit{AtSOFL2}'s overexpression phenotype. In our study, we have similarly shown that specific conserved amino acid residues in SOFL-A and SOFL-B domains are also necessary for both \textit{SOB5}'s and \textit{AtSOFL1}'s overexpression phenotypes (Figure 4). These results, together with sequence logo analysis showing that certain amino acid residues in SOFL-A and SOFL-B domains are 100% conserved in all 289 SOFLs, further strengthen the hypothesis that they are important for function. These results should, however, be interpreted with caution because point mutations can potentially cause proteins to fold incorrectly. Nonetheless, our results demonstrated, at least, that not all point mutations lead to a loss of protein function. Future experiments in which all conserved residues are replaced with alanines or entire domains are deleted may provide answers to whether SOFL-A and SOFL-B domains are important for function. It is not yet clear what biological role these highly conserved domains play. However, conserved domains typically play crucial roles in protein-protein interactions, DNA binding, and other important cellular processes.

**Founding Arabidopsis thaliana SOFL members localize to the nucleus and cytoplasm**

To begin to examine the intracellular distribution of SOFLs homologues we first used nuclear localization signal (NLS) detection and sub-cellular localization programs, SeqNLS (Lin and Hu 2013) and Wolf PSORT (Horton et al. 2007). The two programs predicted that majority of SOFLs contained NLSs and may localize to the nucleus among other organelles (Table S3). This hypothesis was further supported by CFP-SOFL protein fusion subcellular localization experimental data, which showed that SOB5, AtSOFL1 and AtSOFL2 localize to the nucleus.
and the cytosol (Figure 5). However, SOB5 subcellular localization results were at odds with findings by Zhang et al. (2006) who reported that SOB5 only localized to the cytoplasm and/or plasma membrane, but not the nucleus. These contrasting conclusions could be due to a different experimental design and fluorescence detection methods used. The SOB5-GFP C-terminal fusion protein in Zhang et al. (2006) was missing five C-terminal amino acids from SOB5, whereas in our case we used a full-length SOB5 protein fused to the carboxyl end of the CFP tag. Secondly, we used confocal microscopy to detect CFP fluorescence, which is more reliable compared to fluorescence microscopy used by Zhang et al. (2006). In addition, we generated and analyzed N-terminal fusions in contrast to the C-terminal fusion described in Zhang et al., (2006). C-terminal and N-terminal tagged proteins have been shown to show opposite localization patterns (Palmer and Freeman 2004), a possibility that may also explain SOB5-GFP and CFP-SOB5 localization pattern. In the future, both C-and N-terminal fusions of each protein should be examined to avoid such potential problems.

Even though we used a plasmolysis assay to show that the three founding SOFLs were not localized to the cell wall, we could not distinguish whether they were localized in the plasma membrane, cytosol or both. To try and answer this question, we used a bioinformatics approach by running the 289 SOFL homologue sequences against web-based transmembrane protein topology prediction algorithms. TMHMM (Kahsay et al. 2005) and TMMOD (Kahsay et al. 2005), both hidden Markov prediction models, did not predict any transmembrane helices in all 289 SOFL orthologue sequences. This outcome could be verified experimentally by discriminating between cytosolic and cell membrane protein using osmotic disruption of the protoplast vacuole in hypotonic solution (Serna 2005). This method results in the diffusion of the
GFP signal from the cell periphery to the central part of the cell volume, an outcome that will not occur when the protein under study is attached to the cell membrane (Serna 2005). Overall, these data suggest that at least, the three founding Arabidopsis thaliana SOFLs and possibly several other SOFL homologues may localize to the nucleus or cytoplasm.

**CONCLUSION**

The identification of at least 289 SOFL homologues creates an opportunity to study and characterize this novel gene family in at least 89 flowering plant species. A combination of gain-of-function, loss-of-function, protein-protein interaction and CK quantitation studies will go a long way towards answering various questions raised by Zhang et al. (2006, 2009) about the SOFL gene family. One critical question is whether CK nucleosides and nucleotides have biological activity in plant growth and development, as suggested by Zhang et al., (2006, 2009). Gain-of-function and CK-quantitation studies in selected species may be used to test the hypothesis that overexpression of SOB5, AtSOFL1 and AtSOFL2 homologues can cause similar CK-related phenotypes reported by Zhang et al., (2006, 2009). Results from such studies can then be compared to higher order null mutants in which putative redundant SOFLs from the same phylogenetic clades are knocked out. In addition, our site-directed mutagenesis studies suggested that conserved SOFL-A and SOFL-B domains were important for function. Similar studies involving mutagenesis of additional conserved amino acid residues in both Arabidopsis and other species will likely provide new functional details regarding these poorly characterized domains. Finally, even though our latest database searches suggest that SOFLs are a plant specific gene family, the ever-increasing number of sequenced genomes will continue to test this hypothesis.
ACKNOWLEDGEMENTS

We thank Dr. Jingyu Zhang of the Institute of Botany, Chinese Academy of Sciences, Beijing, China for assistance in providing background information regarding prior work. We also thank Dr. Amit Dhingra and his lab members at Washington State University for help using the Biolistic PDS-1000/He Particle Delivery System during transient expression of CFP-protein fusion in onion epidermal layers. We are grateful to Dr. Daniel Mullendore of the WSU Franceschi Microscopy and Imaging Center for the guidance and assistance in using the confocal microscope during sub-cellular localization analyses. This research was supported by the USDA National Institute of Food and Agriculture, HATCH project 1007178 (to M.M.N.). This work was also supported in part by the funds provided by the Program in WSU Molecular Plant Sciences GPSI Fellowship (to R.T.).

Conflict of interest

The authors have no conflict of interest to declare.
Short legends for Supporting Information

Table S1. Primers used in this manuscript. List of primers that were used for genotyping, construct development and measuring transcript accumulation.

Table S2. A total of 289 SOFL homologues were identified in at least 89 flowering plant species. Founding Arabidopsis thaliana SOFL family members were used as queries to blast the NCBI (Sayers et al. 2009) and Phytozome databases. SOFL gene family homologues were retrieved from genome sequences of 89 monocotyledon and dicotyledon species. The total number of SOFLs included in this list is not final and is likely to change over time as the curation, annotation and sequencing of more genome sequences continue to improve and increase.

Table S3. Subcellular localization prediction analysis. All 289 SOFL orthologue protein sequences were analyzed through Wolf PSORT (Horton et al. 2007) and SeqNLS (Lin and Hu 2013) online databases. SeqNLS did not detect NLS in 14 SOFL homologues and classified an additional 119 SOFLs as scoring below a cut-off threshold score to be confidently predicted as nuclear localizing (containing residues in yellow, green, cyan, blue and navy-blue shades). Wolf PSORT predicted 274 SOFLs out of 289 homologues to localize to the nucleus. Underlined orange and red colored parts of the peptide sequence represent predicted NLSs.

Figure S1. AtSOFL1 is alternatively spliced into AtSOFL1.1 and AtSOFL1.2 splice variants. (a) AtSOFL1.1 is the first splicing variant and represents At1g26210, previously thought to be an intron-less gene, and includes a partially retained intron (dashed box and red solid line).
AtSOFL1.2 represents the second splicing variant and includes two predicted exons. The blue box [a] and the green box [b] both represent the two predicted exons. The red solid line represents an intron. The brown box represents 3’ UTR region. Small orange and purple regions represent the SOFL-A and SOFL-B domains, respectively. (b) Agarose gel image showing a AtSOFL1.1 PCR product amplified from genomic DNA and an alternatively spliced AtSOFL1.2 semi-quantitative PCR product. M = marker.

File S1. SOFL homologue protein sequences. List of 289 SOFL homologues retrieved from NCBI and Phytozome database searches.

File S2. Annotation of AtSOFL1 and the two alternative splicing variants. Our BLASTP search revealed a, (a) 168 amino acid hypothetical protein, GenBank accession number AAG50669.1 with a predicted (b) 507 nucleotide long coding sequence (CDS), GenBank accession number AC079829.6 from which we designed PCR primers (Table S1). To further investigate this putative SOFL gene, we performed PCR using genomic DNA and complementary DNA as templates and amplified products of (c) 986 and (d) 771 base pairs sizes (which includes the 3’ UTR sequence), respectively. Based on the predicted CDS length we had expected a PCR product of 507 base pairs if we used cDNA as template. To investigate this anomaly, we sequenced the two amplified products. Sequence alignment of the genomic DNA and cDNA PCR amplified products revealed that AtSOFL1 contains a 215-base pair intron (highlighted in grey). Sequence alignment analysis showed that (e) AtSOFL1.1’s first 319 nucleotides out of 986 are 100% identical to (f) AtSOFL1 (At1G26210) gene’s CDS. Furthermore, the terminal 128 nucleotides of AtSOFL1.1’s CDS are 100% identical to the first
segment of AtSOFL1’s 215 nucleotide long intron segment. This suggests that AtSOFL1’s 986 nucleotide-long gene produce two alternatively spliced variants; AtSOFL1.1 and AtSOFL1.2. AtSOFL1.1 CDS is translated into a 148-amino acid long peptide and (g) AtSOFL1.2 CDS is translated into a predicted 118 amino acid products. *represents a stop codon.

FIGURE LEGENDS

Figure 1. Phylogenetic analysis of 289 SOFL homologue protein sequences showed that the gene family evolved into four main clades. (A) and (B) Annotated rectangular layout mid-point root Mr. Bayes phylogenetic trees. Bayesian inference analysis was performed on 289 SOFL homologues amino acid sequences with the Mr. Bayes 3.2.1 on XSEDE tool on CIPRES Science Gateway for 30 million generations with 8 chains to run with convergence at 0.0099 (Miller et al. 2015).

Figure 2. Expression levels of Arabidopsis thaliana SOFL genes in various tissues. Total RNA was isolated from; 6-day old seedlings (S), juvenile plants (J), rosette leaf (L), floral structure (F), siliques (SQ), whole adult plant (WP) and roots (R) and used for RT-PCR analysis. UBIQUITIN10 was used as an internal control for PCR. No reverse transcriptase (No RT) RNA samples were used as negative controls.

Figure 3. Multiple sequence alignment and analysis of SOFL proteins. (a) Illustration showing the topology of SOFL proteins. Purple rectangular blocks represent the two conserved domains, SOFL-A and SOFL-B. (b) N-terminal amino acid sequence alignment of Arabidopsis thaliana SOFL proteins. The alignment was obtained using JalView Probcons with Defaults program (Malek 2001; Zhang 2003). Semi-conserved amino acids are indicated with a light
purple shade. Conserved amino acids are indicated by a darker purple shade. Red squares indicate amino acid residues that were selected for site-directed mutagenesis. WebLogo© analysis of conserved, (c) SOFL-A domain, and (d) SOFL-B domain, using all 289 SOFL protein sequences (Lin and Hu 2013). Red asterisk denotes 100% conserved amino acid residues among all 289 SOFL sequences.

**Figure 4. Phenotypes of transgenic plants overexpressing full-length SOB5 and AtSOFL1 genes harboring point mutations generated via site-directed mutagenesis.** (a) Wild-type and sob5-D control plants compared to transgenic plants overexpressing the following SOB5 point mutations: T21I, P61R and D33H. D33H mutation is in a non-conserved portion of SOB5 gene. (b) Wild-type and 35S:AtSOFL1 control plants compared to transgenic plants overexpressing the following AtSOFL1 point mutations: T23I, P84R and D53H. D53H is in a non-conserved portion of AtSOFL1 gene. All plants were grown together in a growth chamber for two weeks. Three independent transgenic lines were analyzed. This experiment was repeated three times with similar outcomes.

**Figure 5. Subcellular localization of CFP-SOB5 (a), CFP-AtSOFL1 (b), CFP-AtSOFL2 (c) fusion proteins in onion epidermal cells.** Most biochemical functions carried out in plant cells are performed by proteins in specific cellular locations. Protein subcellular localization studies, via fluorescent protein fusions, are a useful tool to narrow down cellular functions of novel or unknown proteins. Onion epidermal peels were biolistically bombarded with CFP-protein fusion constructs and incubated in the dark for 40 hours, then visualized under a confocal microscope. Each panel shows, in a clockwise direction, CFP-protein fusion signal, free RFP signal, merge of CFP and RFP signal and bright-field view showing outline of cells. White arrows indicate outline of
plasmolyzed membrane. Free RFP construct was used as a control for successful bombardment as well as a localization marker. Plasmolysis was induced by treatment with 0.8 M mannitol.
**LITERATURE CITED**


570 Earley, K.W., J.R. Haag, O. Pontes, K. Opper, T. Juehne et al., 2006 Gateway-compatible vectors for


574 Goodstein, D.M., S. Shu, R. Howson, R. Neupane, R.D. Hayes et al., 2011 Phytozome: a comparative

576 Guo, Y.L., 2012 Gene family evolution in green plants with emphasis on the origination and evolution of

578 Higuchi, M., M.S. Pischke, A.P. Mahonen, K. Miyawaki, Y. Hashimoto et al., 2004 In planta functions of
579 the Arabidopsis cytokinin receptor family. *Proceedings of the National Academy of Sciences of

581 Horton, P., K.J. Park, T. Obayashi, N. Fujita, H. Harada et al., 2007 WoLF PSORT: protein localization

583 Jain, M., P. Khurana, A.K. Tyagi, and J.P. Khurana, 2008 Genome-wide analysis of intronless genes in

586 detection and topology prediction and its applications to complete genomes. *Bioinformatics* 21
587 (9):1853-1858.

588 Kersey, P.J., J.E. Allen, I. Armean, S. Boddu, B.J. Bolt et al., 2016 Ensembl Genomes 2016: more


Zhao, J., D.S. Favero, J. Qiu, E.H. Roalson, and M.M. Neff, 2014 Insights into the evolution and
diversification of the AT-hook Motif Nuclear Localized gene family in land plants. *BMC Plant
Biol* 14 (1):266.

Zou, M., B. Guo, and S. He, 2011 The roles and evolutionary patterns of intronless genes in
Clade III

Clade VI