Plant Omics Journal POJ

POJ 8(2):112-121 (2015) ISSN:1836-3644

A systematic bioinformatics analysis of small proteins in plants

Xinmiao Jia[1,2], Jun Yu[1,2], Jiayan Wu[1,2*], Jingfa Xiao[1,2*]

[1]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, People's Republic of China
[2]University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China

[*]Corresponding author: Jiayan Wu: wujy@big.ac.cn; Jingfa Xiao: xiaojingfa@big.ac.cn

Abstract

Small proteins (SPs, <=100 amino acids in length) are prevalent in all prokaryotes and eukaryotes. They are known to perform varieties of relevant functions and participate in regulation of various biological processes. Although integrated studies of SPs in prokaryotes and animals have been carried out, the systematic investigation on plant SPs still remains an unwritten story. This is mainly because of the lack of sequenced whole genomes in plant, which is improving by the sequence data explosion triggered by next generation sequencing. In this study, we extract 37,003 SPs from 13 whole genome sequenced plants, including 3 green algae, 1 bryophyte, 3 monocots and 6 dicots. We mainly analysed the compositional features, conservative relations, enriched functions in different conserved groups, and the functional domain and evolution characteristics of plant SPs. We observed that the majority of SPs (64.20%) are species specific and 89.31% of these species-specific SPs do not match with any gene ontology (GO) functional annotation. It seems that organisms are likely to enrich SPs to exert specialized functions. By grouping SPs on the basis of sequence conservation within lineages, we noticed that SPs perform lineage-specific functions and many corresponding biological functions emerge with the evolution of SPs. The domains probably evolve independently in SPs while change to other cooperation patterns in the long course of evolution. In addition, gene duplication could be the primary force in the evolution of some plant SPs, for example, small nuclear ribonucleoproteins.

Keywords: Evolution; Function; Plant; Small proteins; Systematic study.
Abbreviations: AAs_Amino acids; GO_Gene ontology; SPs_Small proteins.

Introduction

Small proteins (SPs, <=100 amino acids (AAs) in length) are widespread in all three kingdoms and play important roles in protein synthesis, energy metabolism, lipid transport, metabolism, transcription regulation, stress response, oxidoreduction and so on (Basrai et al., 1997; Hobbs et al., 2011). They are also found to be functionally important in the growth and development of animal and plants. For example, CLE protein family (75-140AAs) regulates the meristem development of *Arabidopsis thaliana* (Fletcher et al., 1999; Oelkers et al., 2008; Trotochaud et al., 2000) and TAL protein (11AAs) influences the growth of *Drosophila melanogaster* (Galindo et al., 2007). Most SPs contain only one domain and the secondary and tertiary structure of them are relatively simple. So, the study of SPs not only helps us to understand their functions and molecular mechanisms in life more clearly but also be convenient for the researches on protein-folding and stability of protein structure (Imperiali and Ottesen, 1999; Mezo et al., 2001; Polticelli et al., 2001). In addition, the studies on the structure characteristics and binding activities of SPs can promote the selection and design of new drugs (Martin and Vita, 2000).

Although SPs are functionally important, studies on SPs are few compared to larger proteins, primarily due to their small sizes and the limitations of routine biochemical assays and bioinformatics methods. Methodologies like high-throughput sequencing, homology searching, expression-based analysis and gene trapping have enabled to analyse SPs (Basrai and Hieter, 2002; Kumar et al., 2002; Samayoa et al., 2011; Su et al., 2013; Yang et al., 2011; Zhao et al., 2012). At the very start, SPs are used as model systems to study the determinants and stability of protein folding due to their simple and typical structures (Hartley, 1989; Kim and Baldwin, 1990; Kuwajima and Schmid, 1984; Polticelli et al., 2001). As time goes on, the function analysis of several SPs began to draw some attention. Kurata et al. (2005) found that CAPRICE (CPC; 94AAs), a transcription factor, is involved in the intercellular signal transduction associated with root epidermal cell differentiation. Gleason et al. (2008) discovered Cg-1 protein (<33AAs) controls the interaction between tomato and nematode. Ma et al. (2006) demonstrated that plants are exceptional among eukaryotes in employing small heat-shock proteins in the peroxisome matrix to prevent unspecific aggregation of partially denatured proteins under both physiological and stress conditions. Recently, researchers began to concern the large-scale functional and evolutionary significances studies on SPs. Wang et al. (2008) conducted a systematic survey of SPs in bacteria and archaea and found SPs play significant roles in various functions and are likely under differential selective pressures that reflect the respective life-styles of the organisms. Zhao et al. (2012) analysed lineage-specific SPs across eight eukaryotes and revealed that some eukaryotic SPs perform lineage-specific functions and they evolve and express in certain unique ways. However, the functional and evolutionary characteristics across plant species are still unkown.

Here, we selected 13 whole genome sequenced plant species, including 3 green algae (*Ostreococcus tauri*, *Micromonas sp.*

RCC299 and *Chlamydomonas reinhardtii*), 1 bryophyte (*Physcomitrella patens*), 3 monocots (*Oryza sativa*, *Zea mays* and *Sorghum bicolor*) and 6 dicots *(Vitis vinifera*, *Theobroma cacao*, *Arabidopsis thaliana*, *Fragaria vesca*, *Glycine max* and *Populus trichocarpa*), and analysed all protein sequences that are <=100 AAs. The compositional features (AAs distrubution, exon composition, and so on), the conservative relations, the enriched functions in different conserved groups, and the domain and evolution characteristics of SPs all were analysed in this systematic investigation. Our results indicate that SPs have important functions. Organisms are likely to enrich SPs to exert specialized functions; many corresponding biological functions emerge with the evolution of SPs; domains tend to evolve independently in SPs while develop new patterns in the long course of evolution. The variation of SPs copies is predicted to be the primary force in the evolution of some SPs, such as small nuclear ribonucleoproteins.

Results and Discussion

SPs properties in 13 plant species

There are about 98 plant species sequenced yet. We selected 13 species with better annotation (*Ostreococcus tauri*, *Micromonas sp. RCC299*, *Chlamydomonas reinhardtii*, *Physcomitrella patens*, *Oryza sativa*, *Zea mays*, *Sorghum bicolor*, *Vitis vinifera*, *Theobroma cacao*, *Arabidopsis thaliana*, *Fragaria vesca*, *Glycine max* and *Populus trichocarpa*) out of all sequenced (98 species). We downloaded the protein data of these 13 plants, and totally, 37,003 protein sequences with no more than 100 AAs were retrieved. As shown in Table 1, the average percentage of SPs among total proteins in plants is 8.73%, which is lower than the percentage in bacteria and archaea (10.99%) (Wang et al., 2008) but higher than that in invertebrates (about 5%) and vertebrates (about 2%) (Zhao et al., 2012). *Theobroma cacao* in dicots had the highest number of SPs in all 13 species, comprising 14.09% or 6,507 sequences, while *Micromonas sp. RCC299* in green algae had the lowest number, 441 SPs, representing 4.34% of its proteins. The SPs amount in dicots varied with the maximum amplitude, ranging from 5.76% in *Glycine max* to 14.09% in *Theobroma cacao*. We also noted that the average SPs content is higher in monocots (11.08%) than in dicots (8.90%), green algae (7.13%) and *Physcomitrella patens* (5.48%) (Fig. 1a).

The AAs composition of proteins can reflect the requirements of protein structure and function. The stabilization of SPs through disulphide bonds and metal ion binding is the most common strategy adopted to achieve a stable fold in the absence of a hydrophobic core (Polticelli et al., 2001). So, we compared the AAs distribution of SPs and total proteins (control) in 13 plants and measured their differences on structure and function. As displayed in Fig. 1b, the polarity positively charged AAs, K (Lysine), R (Arginine) and H (Histidine) (purple), with higher frequency in SPs than control, while polarity negatively charged AAs, D (Aspartic) and E (Glutamic) (green colour), had a lower frequency. After significance testing, we noted two biased groups. M (Methionine), C (Cysteine) and D (Aspartic) had a significant difference between SPs and the control (Wilcoxon p-value<0.001). Meanwhile, E (Glutamic) exhibited a moderate difference (Wilcoxon p-value<0.01). Because M (Methionine) residue may protect proteins from critical oxidative damage (Levine et al., 1999) and is usually the first residue on translational grounds, it is expected to have a higher occurrence in shorter proteins. C (Cysteine) can react with another C (Cysteine), forming a disulfide bond (Marino and Gladyshev, 2010). So, the significant difference of C (Cysteine) between SPs and total proteins can support our understanding of stabilization force of SPs. The usage biases of D (Aspartic) and E (Glutamic) along with the differences of polarity positively charged AAs to demonstrate the specificity of SPs in molecular structures and functions from the larger proteins. This result agrees with our previous investigation of SPs in 8 eukaryotes (Zhao et al., 2012).

An exon is any nucleotide sequence encoded by a gene that remains present within the final mature RNA product, after introns have been removed by RNA splicing. The exon composition can provide some clues for the research on alternative splicing. Here, we examined the exon composition of SPs in Fig 1c. Interestingly, 90% of these SPs in all 13 plants are composed of no more than 3 exons. The majority of SPs in *Fragaria vesca* had 2 exons, which is so different from all other species. According to our following conservation analysis, *Fragaria vesca* had the largest number of species specific SPs, accounting for 89.60%. So, we inferred this large group of *Fragaria vesca*'s specific SPs cause this difference in the exon composition analysis. We also took *Arabidopsis thaliana* and *Oryza sativa*, two well-annotated model plants, as examples to investigate the alternative splicing events by mapping the full-length cDNAs in their genomes. In *Arabidopsis thaliana*, 31.55% genes were alternativly spliced. But this percentage decrease to 17.76% when we only accounted the genes coding SPs. In *Oryza sativa*, the alternative spliced genes percentages were 22.01% and 12.53% for total genes and the genes coding SPs, respectively. It can be concluded that there are less alternative splicing events in the genes coding SPs than the larger ones.

Conservation analysis in SPs

We used inparanoid/multiparanoid (Alexeyenko et al., 2006; Berglund et al., 2008; Ostlund et al., 2010) to investigate SPs' conservation profile, and to define it across the following parts: conserved in all 13 species (totally 1,017 SPs accounting for 3.78%) and conserved in 9 angiosperms (12.91%), 6 dicots (16.56%), 3 monocots (18.40%) and 3 green algae (10.70%). Then, we defined the SPs failed to classify into any homolog groups, which are unique to a single species, species-specific SPs (Table 2). The average percentage of species-specific SPs was 64.20%, which means the species-specific SPs are far more abundant than the conserved ones in most plants and concurs with the study on prokaryotes (58.79%) (Wang et al., 2008) and 8 eukaryotes mainly on animals (41.06%) (Zhao et al., 2012). We further investigated gene ontology (GO) function classification of these species specific SPs. There were a total of 37 GO terms on GO level 2 (Supplementary Fig. 1), and the GO terms with the maximum SPs involved in biological process is metabolic process (1,212), cellular process (1,116) and so on. But 89.31% of these species specific SPs did not have any GO annotation. These SPs were almost all hypothetical proteins. We predicted that organisms might incline to enrich SPs to exert specialized functions because SPs are easy to generate according to the hypothesis that organisms tend to minimize costs of protein biosynthesis (Seligmann, 2003). The following SPs domain research indicates that most SPs contain only one domain and they can perform functions simply and directly through protein-protein interaction or binding DNA/RNA sequences (Wang et al., 2008), which is very significant for the stress response to assist organisms adapt to environment better.

Table 1. Genome and protein properties of 13 plant species.

Lineage	Species	Genome size(Mb)	Ploidy	Proteins	SPs	Percent of SPs
Green algae	*Ostreococcus tauri*	12.6	Haploid	7,987	736	9.21%
	Micromonas sp. RCC299	21	Haploid	10,154	441	4.34%
	Chlamydomonas reinhardtii	120	Haploid	14,489	1,138	7.85%
Bryophyte	*Physcomitrella patens*	511	Haploid	35,991	1,971	5.48%
Monocot	*Oryza sativa*	389	Diploid	28,555	3,050	10.68%
	Zea mays	2,355	Diploid	43,497	5,749	13.22%
	Sorghum bicolor	730	Diploid	33,005	3,080	9.33%
Dicot	*Vitis vinifera*	487	Diploid	24,289	1,478	6.09%
	Theobroma cacao	430	Diploid	46,178	6,507	14.09%
	Arabidopsis thaliana	125	Diploid	35,375	2,596	7.34%
	Fragaria vesca	240	Diploid	34,754	2,941	8.46%
	Glycine max	1,115	Diploidized tetraploid	44,887	2,586	5.76%
	Populus trichocarpa	485	Diploid	40,521	4,730	11.67%
				Average		8.73%

Function analysis of SPs in different conserved groups

We conduct a function annotation and enrichment of the SPs conserved in all 13 species and those conserved only in 9 angiosperms, 6 dicots and 3 monocots based on DAVID (Huang et al., 2009a; Huang et al., 2009b) (DAVID Bioinformatics Resources 6.7). The SPs conserved only in 3 green algae were escaped from analysing because the number of SPs in this group was very few and the function of most SPs unknown.

We totally generated 179 functional clusters, including 42 shared by 13 species, 57 shared by 9 angiosperms, 15 shared by 6 dicots and 65 shared by 3 monocots. Then we summed up the top 6 largest functional clusters for every conserved group (Fig. 2a-d). SPs conserved in all 13 species were ribosomal protein, small nuclear ribonucleoprotein, RING-type zinc-finger protein, heavy metal transport protein, 10 kDa chaperonin and ubiquitin. These SPs were all related to the most basic genetic information processing metal ion transport (Fig. 2a). The SPs conserved in 9 angiosperms included DVL family protein, wound induced protein, metallothionein-like protein, gibberellin-regulated protein, phytosulfokine and arabinogalactan protein. They are known to play essential roles in plant development, defense response and homeostasis (Fig. 2b). The SPs shared by 6 dicots were LEA protein, rapid alkalinization factor, major latex-related protein, LOB domain-containing protein, pistil-specific extensin-like protein and phloem protein 2. They are related to the development of flower, fruit, seed, leaf and root, and also help to transport the organic matter (Fig. 2c). The SPs conserved in 3 monocots included cyclin-dependent kinase inhibitor, abscisic stress ripening protein, basal endosperm transfer layer4 precursor, root cap-specific glycine-rich protein, egg apparatus 1 and EARLY flowering 4 protein. These SPs are mainly involved in circadian rhythm regulation and embryo development (Fig. 2d). It is very interesting that SPs in these 4 conserved groups perform lineage-specific functions and are related to the new emerged properties (Fig. 2e). For example, DVL family proteins play an important role in the development of shoot system, which is a trait to higher plants different from algae and bryophytes. The endosperm generally persists to the mature seed stage as a storage tissue in monocots, while absorbed during embryo development in most dicots. Basal endosperm transfer layer4 precursor conserved only in 3 monocots is thought to be relative to the development of endosperm (Becraft, 2001), which reflects the phenotypic difference between dicot and monocot. These results demonstrate that SPs are functionally important, and they are selected during the evolution process (Zhao et al., 2012).

In addition, we conducted a thorough search for the 1,017 SPs conserved in 13 plants, against all archaea, bacteria, fungi, protozoa and animals SPs downloaded from NCBI. The blast result shows that 347 out of 1,017 SPs are conserved in all groups. These SPs are ribosomal protein, small nuclear ribonucleoprotein and RNA-binding protein, all concerned in basic genetic information processing. The SPs only conserved in plants were chlorophyll a/b binding protein, photosystem I reaction center subunit VI, photosystem I subunit X and auxin transport protein, which are related to photosynthesis and plant development, suggesting the specific functions that plant SPs perform.

The SPs domain characteristics

We searched the domains (or motifs and conserved regions) within the conserved SPs in Table 2 and found four patterns in the evolution process (Fig. 3). The SPs with independent domains were much more abundant than the other three patterns with a percentage of 80.93%. The other three patterns were all cooperation patterns, named as co-occurring with other domains, chimera with other domains and self-tandem. We examined all the independent domains in Interpro database (Apweiler et al., 2001), and noticed that most of them will evolve with other domains, when protein length increases. The exception could be divided into two groups, one is never related with other domains, including PSI_PSAK, Spt4, DUF1903, Toxin_2, and Toxin_3; one is not related with other domains only in plants, including zf-Apc11 and L51_S25_CI-B8. Pattern 2, co-occurring with other domains, could be the dominant way in protein evolution that attains functional integration by binding with different domains. This tendency is consistent with the SPs function evolution trail displayed in Fig 2e. We analyzed the domains of SPs in archaea, bacteria, fungi, protozoa and animals as well and found 19 unique domains in plant SPs. Most of these unique domains had unknown function except for CK1gamma_C which is casein kinase 1 gamma C-terminal with protein serine/threonine kinase activity and Auxin_inducible participates in plant development.

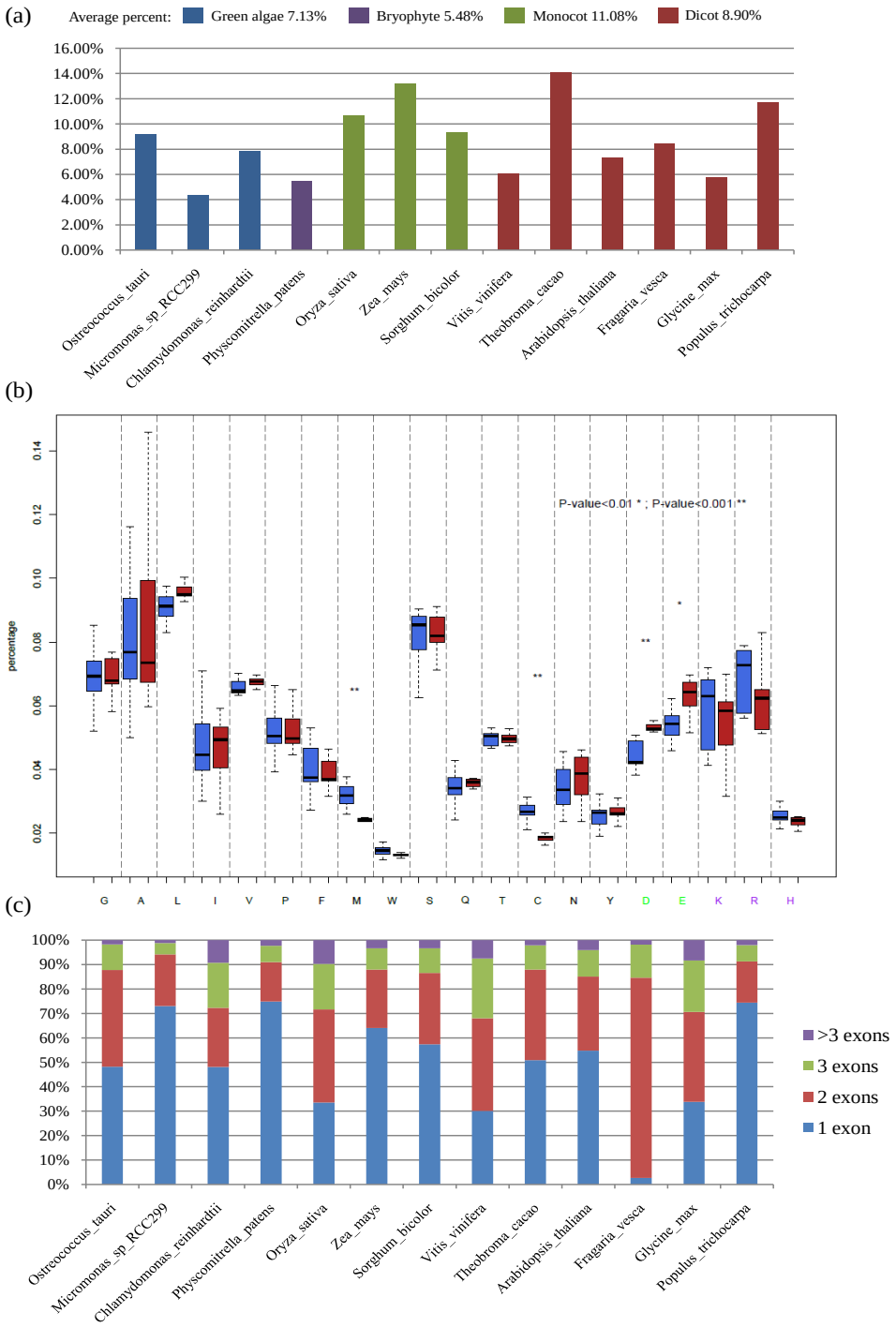

Fig 1. SPs properties in 13 plant species. (a)The percentages of SPs in total proteins of 13 plant species. The species are color-coded with lineage: blue for green algae, purple for bryophyte, green for monocots and red for dicots. The average percentage of each lineage is shown above. (b) Comparison of the amino acids distribution in SPs (blue box) and total proteins (red box). Polarity positively charged amino acids were coloured with purple and polarity negatively charged amino acids coloured with green. (c) Exon composition of SPs in 13 plant species.

Table 2. Conservation analysis of SPs.

Species	Total SPs	Species specific	Conserved in 13 species	Conserved in 9 angiosperms	Conserved in 6 dicots	Conserved in 3 monocots	Conserved in 3 green algae
Ostreococcus tauri	736	633(86.01%)	24	-	-	-	50(4)
Micromonas sp. RCC299	441	222(50.34%)	46	-	-	-	79(3)
Chlamydomonas reinhardtii	1,138	898(78.91%)	53	-	-	-	84(3)
Physcomitrella patens	1,971	735(37.29%)	151	-	-	-	-
Oryza sativa	3,050	2,268(74.34%)	60	299(68)	-	545(61)	-
Zea mays	5,749	4,477(77.87%)	121	549(144)	-	957(90)	-
Sorghum bicolor	3,080	2,088(67.79%)	60	313(66)	-	637(66)	-
Vitis vinifera	1,478	331(22.40%)	71	352(92)	402(17)	-	-
Theobroma cacao	6,507	5,197(79.87%)	80	310(83)	348(11)	-	-
Arabidopsis thaliana	2,596	1,713(65.99%)	79	381(109)	450(31)	-	-
Fragaria vesca	2,941	2,636(89.60%)	21	127(44)	150(5)	-	-
Glycine max	2,586	967(37.39%)	114	680(185)	759(29)	-	-
Populus trichocarpa	4,730	3,161(66.83%)	137	607(144)	711(35)	-	-
Total	37,003	25,326	1,017	3618(935)	2820(128)	2139(217)	213(10)

The number in parentheses from column 5 to column 8 indicates the SPs conserved only in this group.

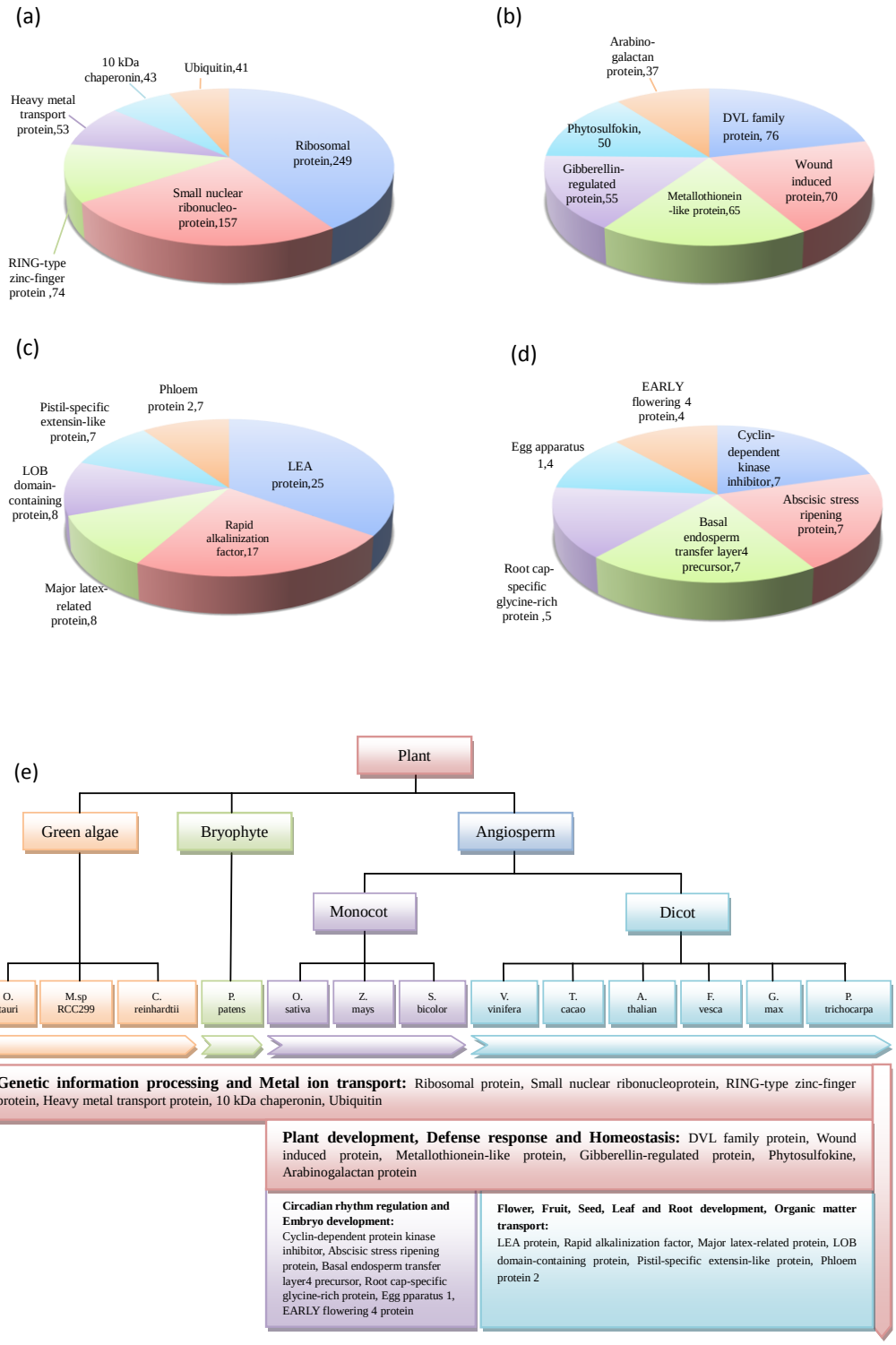

Fig 2. Functional analysis of species conserved SPs. The total counts of the functional clusters of related SPs were summed up to 100% in the pie chart. Top 6 functional SPs clusters conserved in all 13 plant species (a), in 9 angiosperms (b), 6 dicots (c) and 3 monocots (d) are shown. Texts are the function of each cluster followed by the counts of SPs. (e) The schema displays SPs evolving from simple to complex (left to right) with time (arrows). The red downward arrow (right) indicates increasing functional complexity and new functions. At each step, novel SPs are related to lineage-specific functions and new emerged functions (as shown in the frames).

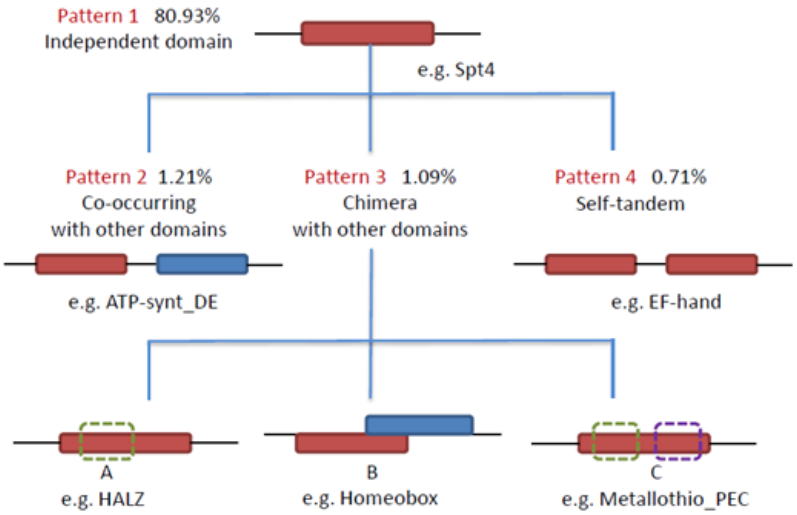

Fig 3. SPs domain patterns. Red box and blue box mean different domains and red box domains are exampled below. Dashed lines boxes represent the domain have been evolved to a part of other domain or conserved region of protein family.

Evolution significances of SPs

Different lineages have different variation features in gene constitution. To our best knowledge, for prokaryotes the primary variation force is gene gain and loss (Treangen and Rocha, 2011). For animals, the pattern of gene constitution variation is mainly gene duplication and gene gain and loss (Friedman and Hughes, 2001). For many plants, the main variation feature is whole genome duplication (Severns et al., 2013) along with horizontal gene transfer in some early plants evolution (Huang and Gogarten, 2008). As for the major driving force for the evolution of small proteins in plants, we further analyze the copy number of all SPs in plants as well as the copy number of SPs in animal, fungi, protozoa, bacteria and archaea. We found the percent of multi-copy SPs in plants is 40.08% higher than animal (29.02%), protozoa (15.23%), fungi (9.64%), bacteria (10.68%) and archaea (10.46%). This percentage improved to 65.74% when we only concerned the SPs conserved in all species. We also noted that not every copy of SPs in plants are gathered together. So, we suggest gene duplication and mutations in different copies may promote the evolution of SPs in plants. This phenomenon also reflects the whole genome duplication in plants.

Here, we choose small nuclear ribonucleoproteins that conserved in the archaea-bacteria-fungi-protozoa-animals-plants group as an example and conducted a phylogenetic analysis. In order to make the result clearer, we only selected 3 to 5 species from prokaryotes, fungi, protozoa and animals, respectively (Supplementary Fig. 2). This analysis showed 8 clusters consistent with separate small nuclear ribonucleoproteins subgroups, which are small nuclear ribonucleoprotein E (cluster 1), U6 snRNA-associated Sm-like protein LSm7 (cluster 2), small nuclear ribonucleoprotein G (cluster 3), U6 snRNA-associated Sm-like protein LSm8 (cluster 4), U6 snRNA-associated Sm-like protein LSm6 (cluster 5), Like-Sm ribonucleoprotein core (cluster 6), U6 snRNA-associated Sm-like protein LSm5 (cluster 7) and U6 snRNA-associated Sm-like protein LSm3 (cluster 8). SPs in cluster 2 only originated from plants, while SPs in cluster 5 and 6 originated from other lineages except plants. We further investigated the SPs phylogeny in each cluster and discovered that there are more duplications of small nuclear ribonucleoproteins in plants than those in other species (Supplementary Figs. 3-10), which reflect the higher amount of genome duplication in plants. Taking small nuclear ribonucleoprotein E (cluster 1, Supplementary Fig. 3) as an example, almost all plant species have at least 2 copies. The copies of the same species are not always gathered together such as *Zea mays* and *Oryza sativa*. It seems that gene duplication could be the primary force in the evolution of small nuclear ribonucleoprotein E. At the same time, the phylogenetic tree of small nuclear ribonucleoprotein E in our analysis is not consistent with the phylogeny of plants. For instance, *Physcomitrella patens* and *Chlamydomonas reinhardtii* are present among angiosperm. We predicted that small nuclear ribonucleoprotein E in plants may derive from green algae after a horizontal gene transferring between green algae and fungi or protozoa. The similar phenomena also occur in U6 snRNA-associated Sm-like protein LSm5 (Supplementary Fig. 9) and U6 snRNA-associated Sm-like protein LSm3 (Supplementary Fig. 10).

Materials and Methods

Data retrieve and pre-processing

We downloaded the proteins and genome sequences of *Ostreococcus tauri*, *Micromonas sp. RCC299*, *Chlamydomonas reinhardtii*, *Physcomitrella patens*, *Oryza sativa*, *Sorghum bicolour*, *Vitis vinifera*, *Arabidopsis thaliana*, *Glycine max* and *Populus trichocarpa* from RefSeq (release54, ftp://ftp.ncbi.nlm.nih.gov/refseq/relese/plant/). The proteins of *Zea mays* are from NCBI (http://www.ncbi.nlm.nih.gov/protein) up to date 17thAugust, 2012 and the genome sequences from GenBank (accession nos. GK000031.2, GK000032.2, GK000033.2, CM000780.2, CM000781.2, CM000782.2, CM000783.2, CM000784.2, CM000785.2 and CM000786.2). The proteins and genome sequences of *Theobroma cacao* were retrieved from http://cocoagendb.cirad.fr/gbrowse/download.html, Version 1.0. The proteins and genome sequences of *Fragaria vesca* are from http://www.rosaceae.org/species/fragaria/fragaria_vesca/, Version 1.0. The proteins in archaea,

bacteria, fungi, protozoa and animals were downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/). The full-length cDNAs of *Arabidopsis thaliana* were retrieved from GenBank (query "*Arabidopsis thaliana* [ORAN] AND FLI_CDNA[KYWD]") for a total of 44,854 sequences. For *Oryza sativa*, sets of 40,816 cDNAs were obtained from GenBank by using the similar query.

We used cd-hit (Li and Godzik, 2006) to reduce redundancy of the proteins for *Zea mays*, *Theobroma cacao* and *Fragaria vesca* with identity equals 100%. For *Theobroma cacao*, we did not find any SPs less than 50 AAs in it. So we carried out a BLAT (Kent, 2002) search based on the SPs less than 50 AAs in the other 12 species with identity and coverage higher than 60%. The percentages of SPs in *Micromonas sp. RCC299*, *Physcomitrella patens*, *Vitis vinifera* and *Glycine max* were relatively low compared to the other species. To reduce the inaccuracy from annotation error, we conducted a re-annotation of SPs in these four species based on the SPs in other species using BLAT (Kent, 2002) with identity and coverage higher than 60%.

Conservation analysis

We explored the conservation of SPs using inparanoid (Berglund et al., 2008; O'Brien et al., 2005; Ostlund et al., 2010) and multiparanoid (Alexeyenko et al., 2006). Each cluster is a homology group, if the SPs in a homology group cover all 13 species. We considered these SPs in this homology group were conserved in 13 species. The same with the SPs conserved in 9 angiosperms, 6 dicots, 3 monocots and 3 green algae. If the SPs in a homology group only cover 9 angiosperms, we considered the SPs in this group are conserved only in 9 angiosperms. The same with the SPs conserved only in 6 dicots, 3 monocots and 3 green algae. If the SPs were not belonging to any group, these SPs identified as species specific SPs.

GO terms of the species specific SPs were identified by InterProScan Perl-based version 4.8 (Zdobnov and Apweiler, 2001) with iprscan_DATA_43.1, iprscan_MATCH_DATA_43.1 and iprscan_PTHR_DATA_38.0. We got the figure of GO function classification of these SPs by using WEGO (Ye et al., 2006) (Web Gene Ontology Annotation Plot).

Function annotation and enrichment

Before function analysis, we carried out a functional annotation of these SPs. The homology groups were coming from the above conservation analysis. All SPs in each homology group were considered to have similar functions. The functions of each homology group were derived from DAVID (Huang et al., 2009a; Huang et al., 2009b) (DAVID Bioinformatics Resources 6.7). We chose the InterPro (Apweiler et al., 2000; Apweiler et al., 2001; Hunter et al., 2009; Hunter et al., 2012; Mulder and Apweiler, 2007) annotation as the final annotation. We used Gene Functional Classification on DAVID to conduct the function enrichment of the SPs conserved in all 13 species, conserved only in 9 angiosperms, 6 dicots, and 3 monocots. If two SPs not existed in the same homology group are classified in one function cluster, we merged these two SPs into one function cluster.

To figure out the functions of the SPs conserved in the archaea-bacteria-fungi-protozoa-animals-plants group and the SPs only present in plants, we performed a BLAST (Altschul et al., 1998) alignment between 1,017 SPs conserved in all 13 plant species and 923,163 SPs in other species (48,024 in archaea, 827,091 in bacteria, 9,880 in fungi, 2,883 in protozoa and 35,285 in animals) with E-value < 10^{-5} and coverage > 50%.

Domain and phylogenetic analysis

We carried out the domain analysis by using Pfam (Finn et al., 2006; Sonnhammer et al., 1998; Sonnhammer et al., 1997) and Interpro (Apweiler et al., 2000; Apweiler et al., 2001; Hunter et al., 2009; Hunter et al., 2012; Mulder and Apweiler, 2007). Moreover, we used Mega 5.2 software (Tamura et al., 2011) (bootstrapped Maximum Likelihood method) for the phylogenetic tree construction.

Conclusions

SPs (<=100 AAs in length) are ubiquitous in all prokaryotes and eukaryotes, and play important roles in various biological processes. In this study, we extract a total of 37,003 SPs from 13 whole genome sequenced plants, including 3 green algae, 1 bryophyte, 3 monocots and 6 dicots. The compositional features (AAs distrubution, exon composition, and so on), the conservative relations, the enriched functions in different conserved groups, and the domain and evolution characteristics of SPs were analysed in this systematic investigation. Our results indicated that SPs have important functions. Organisms are likely to enrich SPs to exert specialized functions. Many corresponding biological functions emerge with the evolution of SPs and domains tend to evolve independently in SPs while develop new patterns in the long course of evolution. The variation of SPs copies is predicted to be the primary force in the evolution of some SPs, such as small nuclear ribonucleoproteins.

Conflict of interest

The authors declare that they have no conflict of interests

Authors' contributions

JX and JW designed the project. XJ analyzed and interpreted of data. XJ and JW wrote the paper. JX and JY revised the paper.

Acknowledgements

This work is supported by a grant (No. 2010CB126604) from the National Basic Research and Development Program (973 Program), the Ministry of Science and Technology of the People's Republic of China, and a grant from the National Science Foundation of China (No. 31271386, No. 31471248 and No. 31101063).

References

Alexeyenko A, Tamas I, Liu G, Sonnhammer EL (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics. 22(14): e9-15.

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25(17):3389-3402.

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJA, Zdobnov EM (2000) InterPro - an integrated documentation resource for protein families, domains and functional sites. Bioinformatics. 16(12):1145-1150.

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti T, Corpet F, Croning MDR,

Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJA, Zdobnov EM (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. 29(1):37-40.

Basrai MA, Hieter P (2002) Transcriptome analysis of Saccharomyces cerevisiae using serial analysis of gene expression. Methods Enzymol. 350:414-444.

Basrai MA, Hieter P, Boeke JD (1997) Small open reading frames: Beautiful needles in the haystack. Genome Res. 7(8):768-771.

Becraft PW (2001) Cell fate specification in the cereal endosperm. Semin Cell Dev Biol. 12(5):387-394.

Berglund AC, Sjolund E, Ostlund G, Sonnhammer EL (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res. 36(suppl 1):D263-D266.

Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer ELL, Bateman A (2006) Pfam: clans, web tools and services. Nucleic Acids Res. 34 (suppl 1):D247-D251.

Fletcher LC, Brand U, Running MP, Simon R, Meyerowitz EM (1999) Signaling of cell fate decisions by CLAVATA3 in Arabidopsis shoot meristems. Science. 283(5409):1911-1914.

Friedman R, Hughes AL (2001) Pattern and timing of gene duplication in animal genomes. Genome Res. 11(11):1842-1847.

Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP (2007) Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol. 5(5):e106.

Gleason CA, Liu QL, Williamson VM (2008) Silencing a candidate nematode effector gene corresponding to the tomato resistance gene Mi-1 leads to acquisition of virulence. Mol Plant Microbe Interact. 21(5):576-585.

Hartley RW (1989) Barnase and Barstar - 2 Small Proteins to Fold and Fit Together. Trends Biochem Sci. 14(11):450-454.

Hobbs EC, Fontaine F, Yin X, Storz G (2011) An expanding universe of small proteins. Curr Opin Microbiol. 14(2):167-173.

Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37(1):1-13.

Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 4(1):44-57.

Huang JL, Gogarten JP (2008) Concerted gene recruitment in early plant evolution. Genome Biol. 9(7):R109.

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJA, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2009) InterPro: the integrative protein signature database. Nucleic Acids Res. 37 (suppl 1):D211-D215.

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, McMenamin C, Mi HY, Mutowo-Muellenet P, Mulder N, Natale D, Orengo C, Pesseat S, Punta M, Quinn AF, Rivoire C, Sangrador-Vegas A, Selengut JD, Sigrist CJA, Scheremetjew M, Tate J, Thimmajanarthanan M, Thomas PD, Wu CH, Yeats C, Yong SY (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. 40 (D1):D306-D312.

Imperiali B, Ottesen JJ (1999) Uniquely folded mini-protein motifs. J Peptide Res. 54(3):177-184.

Kent WJ (2002) BLAT - The BLAST-like alignment tool. Genome Res. 12(4):656-664.

Kim PS, Baldwin RL (1990) Intermediates in the folding reactions of small proteins. Annu Rev Biochem. 59:631-660.

Kumar A, Harrison PM, Cheung KH, Lan N, Echols N, Bertone P, Miller P, Gerstein MB, Snyder M (2002) An integrated approach for finding overlooked genes in yeast. Nat Biotechnol. 20(1):58-63.

Kurata T, Ishida T, Kawabata-Awai C, Noguchi M, Hattori S, Sano R, Nagasaka R, Tominaga R, Koshino-Kimura Y, Kato T, Sato S, Tabata S, Okada K, Wada T (2005) Cell-to-cell movement of the CAPRICE protein in Arabidopsis root epidermal cell differentiation. Development. 132(24):5387-5398.

Kuwajima K, Schmid FX (1984) Experimental studies of folding kinetics and structural dynamics of small proteins. Adv Biophys. 18:43-74.

Levine RL, Berlett BS, Moskovitz J, Mosoni L, Stadtman ER (1999) Methionine residues may protect proteins from critical oxidative damage. Mech Ageing Dev. 107(3):323-332.

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22(13):1658-1659.

Ma CL, Haslbeck M, Babujee L, Jahn O, Reumann S (2006) Identification and characterization of a stress-inducible and a constitutive small heat-shock protein targeted to the matrix of plant peroxisomes. Plant Physiol. 141(1):47-60.

Marino SM, Gladyshev VN (2010) Cysteine function governs its conservation and degeneration and restricts its utilization on protein surfaces. J Mol Biol. 404(5):902-916.

Martin L, Vita C (2000) Engineering novel bioactive mini-proteins from small size natural and de novo designed scaffolds. Curr Protein Pept Sci. 1(4):403-430.

Mezo AR, Cheng RP, Imperiali B (2001) Oligomerization of uniquely folded mini-protein motifs: Development of a homotrimeric beta beta alpha peptide. J Am Chem Soc. 123(17):3885-3891.

Mulder N, Apweiler R (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol Biol. 396:59-70.

O'Brien KP, Remm M, Sonnhammer EL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 33(suppl 1):D476-D480.

Oelkers K, Goffard N, Weiller GF, Gresshoff PM, Mathesius U, Frickey T (2008) Bioinformatic analysis of the CLE signaling peptide family. BMC Plant Biol. 8(1):1.

Ostlund G, Schmitt T, Forslund K, Kostler T, Messina DN, Roopra S, Frings O, Sonnhammer EL (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38(suppl 1):D196-D203.

Polticelli F, Raybaudi-Massilia G, Ascenzi P (2001) Structural determinants of mini-protein stability. Biochem Mol Biol Edu. 29(1):16-20.

Samayoa J, Yildiz FH, Karplus K (2011) Identification of prokaryotic small proteins using a comparative genomic approach. Bioinformatics. 27(13):1765-1771.

Seligmann H (2003) Cost-minimization of amino acid usage. J Mol Evol. 56(2):151-161.

Severns PM, Bradford E, Liston A (2013) Whole genome duplication in a threatened grassland plant and the efficacy of seed transfer zones. Divers Distrib. 19(4):455-464.

Sonnhammer ELL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res. 26(1):320-322.

Sonnhammer ELL, Eddy SR, Durbin R (1997) Pfam: A comprehensive database of protein domain families based on seed alignments. Proteins. 28(3):405-420.

Su M, Ling Y, Yu J, Wu J, Xiao J (2013) Small proteins: untapped area of potential biological importance. Front Genet. 4:286.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28(10):2731-2739.

Treangen TJ, Rocha EP (2011) Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. PLoS Genet. 7(1):e1001284.

Trotochaud AE, Jeong S, Clark SE (2000) CLAVATA3, a multimeric ligand for the CLAVATA1 receptor-kinase. Science. 289(5479):613-617.

Wang FY, Xiao JF, Pan LL, Yang M, Zhang G, Jin SG, Yu J (2008) A systematic survey of mini-proteins in bacteria and archaea. PLoS One. 3(12):e4027.

Yang X, Tschaplinski TJ, Hurst GB, Jawdy S, Abraham PE, Lankford PK, Adams RM, Shah MB, Hettich RL, Lindquist E, Kalluri UC, Gunter LE, Pennacchio C, Tuskan GA (2011) Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. Genome Res. 21(4):634-641.

Ye J, Fang L, Zheng HK, Zhang Y, Chen J, Zhang ZJ, Wang J, Li ST, Li RQ, Bolund L, Wang J (2006) WEGO: a web tool for plotting GO annotations. Nucleic Acids Res. 34(suppl 2):W293-W297.

Zdobnov EM, Apweiler R (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 17(9):847-848.

Zhao Q, Xiao JF, Yu J (2012) An integrated analysis of lineage-specific small proteins across eight eukaryotes reveals functional and evolutionary significance. Prog Biochem Biophys. 39(4):359-367.