**BMC
Plant Biology**

Molecular and phylogenetic characterization of the sieve element occlusion gene family in *Fabaceae* and non-*Fabaceae* plants

Boris Rüping[1,2†], Antonia M Ernst[1,2†], Stephan B Jekat[1,2], Steffen Nordzieke[1], Anna R Reineke[3], Boje Müller[1,2], Erich Bornberg-Bauer[3], Dirk Prüfer[1,2*], Gundula A Noll[1]

Abstract

Background: The phloem of dicotyledonous plants contains specialized P-proteins (phloem proteins) that accumulate during sieve element differentiation and remain parietally associated with the cisternae of the endoplasmic reticulum in mature sieve elements. Wounding causes P-protein filaments to accumulate at the sieve plates and block the translocation of photosynthate. Specialized, spindle-shaped P-proteins known as forisomes that undergo reversible calcium-dependent conformational changes have evolved exclusively in the *Fabaceae*. Recently, the molecular characterization of three genes encoding forisome components in the model legume *Medicago truncatula* (*MtSEO1*, *MtSEO2* and *MtSEO3*; SEO = sieve element occlusion) was reported, but little is known about the molecular characteristics of P-proteins in non-*Fabaceae*.

Results: We performed a comprehensive genome-wide comparative analysis by screening the *M. truncatula*, *Glycine max*, *Arabidopsis thaliana*, *Vitis vinifera* and *Solanum phureja* genomes, and a *Malus domestica* EST library for homologs of *MtSEO1*, *MtSEO2* and *MtSEO3* and identified numerous novel *SEO* genes in *Fabaceae* and even non-*Fabaceae* plants, which do not possess forisomes. Even in *Fabaceae* some *SEO* genes appear to not encode forisome components. All *SEO* genes have a similar exon-intron structure and are expressed predominantly in the phloem. Phylogenetic analysis revealed the presence of several subgroups with *Fabaceae*-specific subgroups containing all of the known as well as newly identified forisome component proteins. We constructed Hidden Markov Models that identified three conserved protein domains, which characterize SEO proteins when present in combination. In addition, one common and three subgroup specific protein motifs were found in the amino acid sequences of SEO proteins. *SEO* genes are organized in genomic clusters and the conserved synteny allowed us to identify several *M. truncatula* vs *G. max* orthologs as well as paralogs within the *G. max* genome.

Conclusions: The unexpected occurrence of forisome-like genes in non-*Fabaceae* plants may indicate that these proteins encode species-specific P-proteins, which is backed up by the phloem-specific expression profiles. The conservation of gene structure, the presence of specific motifs and domains and the genomic synteny argue for a common phylogenetic origin of forisomes and other P-proteins.

Background

In vascular plants, photoassimilates are transported through differentiated sieve elements (SEs) in the phloem forming a network of sieve tubes throughout the plant [1]. The pressure-driven mass flow [2] requires

a high degree of functional specialization of the phloem during development. In order to enable efficient translocation of photoassimilates, SEs loose most of their organelles and thus the ability to perform protein biosynthesis [3]. Mature SEs are therefore dependent on adjacent, metabolically-active companion cells, which are connected to SEs by so-called pore-plasmodesm units [4]. The pressure within the sieve tubes can reach 30 bar [5], so rapid and efficient protection against wounding is essential and favoured the evolution of a

* Correspondence: dpruefer@uni-muenster.de
† Contributed equally
[1]Institut für Biochemie und Biotechnologie der Pflanzen, Westfälische Wilhelms-Universität Münster, Hindenburgplatz 55, D-48143 Münster, Germany
Full list of author information is available at the end of the article

plugging mechanism based on specialized phloem proteins (P-proteins) [6]. These structural proteins accumulate in the cytoplasm of metabolically-active, undifferentiated SEs, but are anchored to the plasma membrane when SEs mature [7]. After wounding, they detach from their parietal location and plug downstream sieve plates by forming a gel-like mass, thereby preventing the loss of photoassimilates [8]. This occurs in all the dicotyledonous plant families that have been studied, and P-proteins have also been identified in certain monocotyledonous plants [9]. There is currently no standardized classification for P-proteins although different types were often distinguished by their tubular, fibrillar, granular or crystalline ultrastructure, which may represent different developmental or conformational states of the same protein subunits rather than evolutionarily-distinct families [10]. Nevertheless, *Fabaceae* plants possess a special type of elongated crystalline P-protein bodies [11], which show a unique type of reactivity. The spindle shaped protein bodies, also known as forisomes ("gate bodies") [12], are able to undergo a reversible, calcium-induced conformational change and can consequently plug and open the sieve elements after wounding and regeneration. Three SEO (sieve element occlusion) proteins named MtSEO1, MtSEO2 and MtSEO3 have been identified in the model legume *Medicago truncatula*, and their role in forisome structure and assembly has been confirmed by immunological and GFP-fusion studies [13-15]. Comprehensive promoter analyses in *M. truncatula* roots and *Nicotiana tabacum* plants demonstrated a restricted expression of the corresponding *MtSEO* genes in immature sieve elements [14,16], indicating a highly conserved regulation of promoter activities among diverse plant species, including non-*Fabaceae* lacking forisomes.

Although genes encoding forisome components of *Fabaceae* have recently been isolated and characterized, little is known about the genetic basis of structural P-proteins in other plant families. The only P-protein to be characterized thus far is phloem protein 1 (PP1) from *Cucurbita maxima* [17]. Immunological studies identified this filamentous protein in SE slime plugs and P-protein bodies, although the corresponding mRNA was shown to accumulate in companion cells [18].

In this study, we report the identification of several new *SEO* genes in *Fabaceae* and, most interestingly, also in non-*Fabaceae* plants (which do not possess forisomes). The unexpected occurrence of *SEO* genes in plant families lacking forisomes may argue that these genes encode other structural P-proteins, which implies a common phylogenetic origin. To further characterize the newly-identified *SEO* genes, we analyzed their gene structure and genomic synteny using bioinformatics and studied their expression by RT-PCR. Selected *SEO* genes were also studied by promoter analysis in transgenic plants.

Results

The *SEO* gene family in *Fabaceae*

BLAST searches were carried out using the nucleotide sequences (and derived amino acid sequences) of the three known *M. truncatula SEO* genes [14,15], identifying six further candidate *SEO* genes in the *M. truncatula* genome and 26 in *Glycine max*. In order to determine whether or not these genes are expressed, we amplified a 1-kbp cDNA fragment from each gene by RT-PCR using total seedling RNA as the template. This generated products for five of the six newly-identified *M. truncatula* genes and 21 of the 26 *G. max* genes, the remaining six genes being identified as potential pseudogenes ('pot. ψ' in Figure 1; Additional file 1). Full-length cDNAs were produced for the expressed genes by PCR using gene-specific primers. The potential pseudogenes were tested by RT-PCR using a collection of different primer combinations and total RNA from young and old leaves, shoots, roots, buds and flowers, with no expression detected (data not shown). Two of the expressed genes from *G. max* were reclassified as potential expressed pseudogenes ('pot. ψe' in Figure 1) because sequencing revealed the presence of frameshift mutations in their open reading frames, causing premature termination of protein synthesis. In order to include all pseudogenes in subsequent phylogenetic analyses, full-length cDNA sequences were generated *in silico* using the procedures described in the Methods section.

Next, we set out to establish whether any of the newly-identified SEO proteins were present in the forisomes of either *M. truncatula* or *G. max*. Purified forisomes from each species were separated by SDS-PAGE and peptide sequences were generated for the 75-kDa protein band by ESI-MS/MS following established protocols [12-14]. Peptide sequences derived from the *M. truncatula* forisomes (Additional file 2) indicated the presence of one further SEO protein in addition to the known components MtSEO1-3 [13-15]. Similarly, four of the 26 *G. max* SEO proteins were identified in *G. max* forisomes, where no components had been previously identified (Additional file 2). These results suggest that many of the *SEO* genes do not encode forisome component proteins, although it is possible that they are present but at levels below our detection threshold. In order to distinguish SEO proteins known to encode forisome components from other SEO proteins, we propose that MtSEO1-3 should be renamed MtSEO-F1, MtSEO-F2 and MtSEO-F3, and that the newly identified component should be named MtSEO-F4 (**SEO-F**, **S**ieve **E**lement **O**cclusion by **F**orisomes).
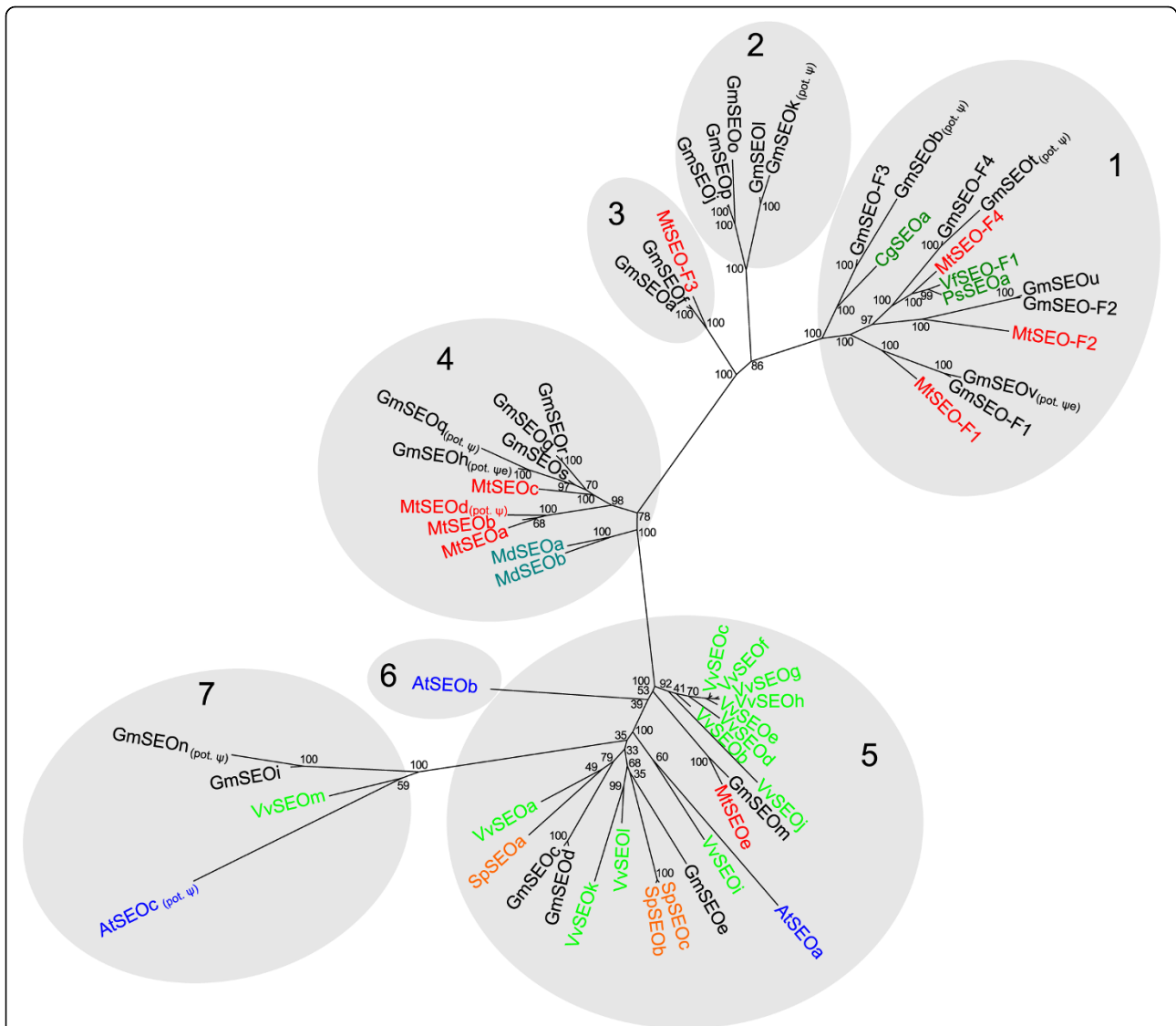
Figure 1 Maximum likelihood phylogenetic tree of SEO proteins from different plants. The phylogenetic tree was constructed with RAxML from a T-Coffee protein sequence alignment and with a bootstrap support of 1000 replicates. Bootstrap percentages are shown on the nodes. Branch lengths are proportional to the number of amino acid substitutions. The shaded parts of the tree represent the subgroups identified with OrthoMCL. Mt = *Medicago truncatula*, Gm = *Glycine max*, Vf = *Vicia faba*, Cg = *Canavalia gladiata*, Ps = *Pisum sativum*, Md = *Malus domestica*, At = *Arabidopsis thaliana*, Vv = *Vitis vinifera*, Sp = *Solanum phureja*, pot. ψ = potential pseudogene, pot. ψe = expressed potential pseudogene.

The four SEO proteins found in purified *G. max* forisomes should similarly be designated GmSEO-F1-F4. For all other *Fabaceae* SEO proteins (and SEO proteins from other plant families, see below) whose function is currently unknown, we recommend the temporary assignment of a lower case letter (SEOa, SEOb, etc) until their presence in the forisome can be confirmed (in which case they will be assigned an SEO-F number) or another function is determined (in which case additional functional categories can be introduced).

The *SEO* gene family in non-*Fabaceae*

Using the *Fabaceae SEO* sequences described above, we screened the genomes of several dicotyledonous non-*Fabaceae* plants, i.e. *Arabidopsis thaliana* (*Brassicaceae*), *Vitis vinifera* (*Vitaceae*) and *Solanum phureja* (*Solanaceae*) as well as an EST collection for *Malus domestica* (*Rosaceae*). This identified three *A. thaliana* genes (designated *AtSEOa-c*), 13 *V. vinifera* genes (designated *VvSEOa-m*) and three *S. phureja* genes (designated *SpSEOa-c*) (Additional file 1). Two full-length *SEO*

genes, designated *MdSEOa* and *b* (Additional file 1), and further partial fragments were found in the *M. domestica* EST collection. Potential *SEO* gene fragments of several other angiosperm plants could also be identified by BLAST search in NCBI GenBank (data not shown). In contrast, no *SEO* genes were identified in the yet sequenced genomes of the monocotyledons *Oryza sativa*, *Brachypodium distachyon*, *Zea mays* and *Sorghum bicolor* nor in the moss *Physcomitrella patens*. RT-PCR confirmed that all the *A. thaliana*, *S. phureja* and *M. domestica* genes were expressed with the exception of *AtSEOc*, which appears to be a pseudogene. The expression profiles of *VvSEOa-m* were not determined due to the lack of the sequenced genotype in our laboratories.

Phylogenetic relationships among *SEO* genes from *Fabaceae* and non-*Fabaceae* plants

The phylogenetic relationships among the *SEO* genes were calculated by creating a maximum likelihood tree from an alignment of all SEO protein sequences. To provide further support for the tree topology we clustered the proteins into subgroups with OrthoMCL (Figure 1). The recently reported forisome protein VfSEO-F1 (formerly known as VfFOR1) from *Vicia faba* [14] and the potential forisome proteins CgSEOa (formerly known as CgFOR1) from *Canavalia gladiata* [14] and PsSEOa (formerly known as PsSEO1) from *Pisum sativum* [19] were also included in the phylogenetic tree. With the exception of MtSEO-F3, all SEO-F proteins clustered in subgroup 1. In addition, several potential pseudogenes as well as one GmSEO protein (GmSEOu) clustered within this group. It should be noted that GmSEOu and GmSEO-F2 share several identical forisome-specific peptide sequences, so GmSEOu is likely to be involved in the formation of forisomes. Subgroups 2 and 3, found on the same branch of the tree, contain predominantly *G. max* SEO proteins of yet unknown function (the only exception is MtSEO-F3). Thus, subgroups 1-3 appear to be *Fabaceae*-specific. Subgroup 4 contains nine *Fabaceae* SEO proteins and both SEO proteins (MdSEOa and b) from the closely related *Rosaceae*, whereas subgroup 5 contains SEO proteins from all plant families included in the study that have sequenced genomes. AtSEOa is the only member of subgroup 6, which is closely related to subgroup 5. Subgroup 7 contains SEO proteins from *G. max, V. vinifera* and *A. thaliana*.

In addition to their high degree of amino acid similarity (>30%), most *SEO* genes have a conserved exon-intron structure as shown in Figure 2, the exceptions being *GmSEOa* and *GmSEOf* (one intron missing) and *MtSEO-F2, AtSEOb, SpSEOb* and *SpSEOc* (additional introns).
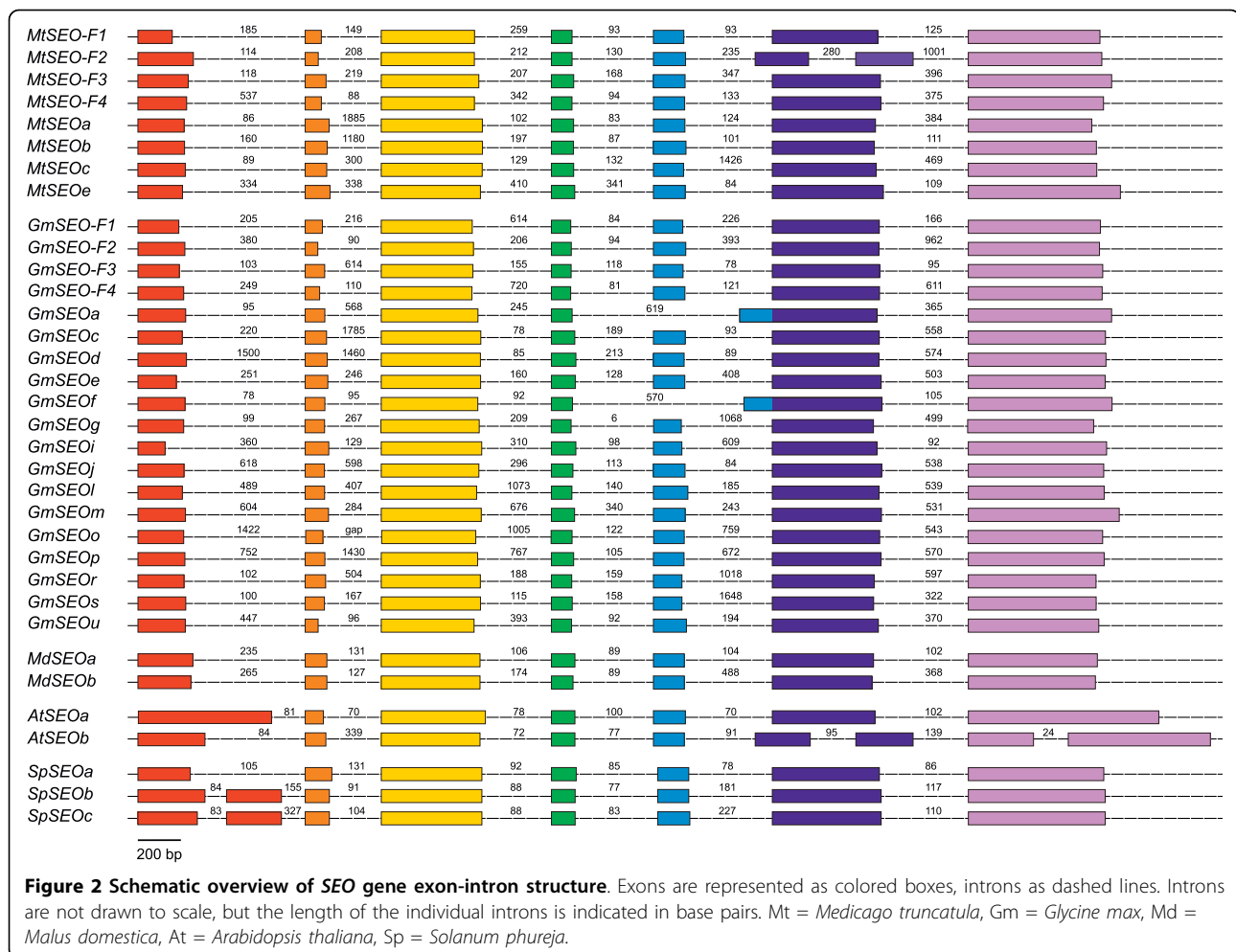
Expression profiles and promoter activities of the *SEO* genes

The expression of *MtSEO-F1-3* was recently shown to be restricted to immature sieve elements [14,16]. To gain an initial impression of whether the *MtSEO, GmSEO, AtSEO* and *SpSEO* genes were also expressed in the phloem we performed RT-PCRs using total RNA from phloem-enriched and phloem-deficient tissue (see Methods). With the exception of gene *GmSEOs*, mRNA levels for all the *SEO* genes were significantly higher in the phloem-enriched tissues (Figure 3). As expected, no mRNA was detected for the potential pseudogenes *MtSEOd, AtSEOc, GmSEOb, GmSEOk, GmSEOn, GmSEOq* and *GmSEOt*. Transcripts for the potential expressed pseudogenes *GmSEOh* and *GmSEOv* harboring frameshift mutations were also detected, although it should be noted that *GmSEOh* mRNA was only found in roots (data not shown). The constitutively expressed *ACT2* (for *A. thaliana*), *GAPDH* (for *M. truncatula* and *S. phureja*) and *F-box* (for *G. max*) genes served as positive controls, as they have been found to be the most appropriate control genes for the plant species included in this study (see Methods).

The preliminary analysis above provided evidence for phloem-specific expression, but for conclusive proof we set out to analyze the activity of two *SEO* promoters in transgenic plants. We chose promoters from the forisome gene *GmSEO-F1* and the none forisome gene *AtSEOa*. A *green fluorescent protein* (*GFP*) gene tagged for retention of the product in the endoplasmic reticulum (ER) was placed under the control of each promoter, producing constructs P*AtSEOa*-GFP$_{ER}$ and P*GmSEO-F1*-GFP$_{ER}$. Ten independent transgenic *A. thaliana* plants expressing P*AtSEOa*-GFP$_{ER}$ were regenerated and analyzed by confocal laser scanning microscopy (CLSM). In stem sections, GFP$_{ER}$ fluorescence was detected in the phloem of the vascular bundle (Figure 4A, B), which appears to be restricted to a 'pipe-line-like' assembly of cells within the phloem. These cells display the typical end-to-end connection of sieve elements and the presence of sieve plates was verified by aniline-blue staining (Figure 4C). Within the cytoplasm, one or more vacuoles were clearly visible indicating an immature nature of these cells (Figure 4D). Detailed CLSM analysis of five independent transgenic *G. max* roots expressing P*GmSEO-F1*-GFP$_{ER}$ also indicated spatially-restricted promoter activity in sieve elements of the vascular cylinder (Figure 4E) showing the same morphology and characteristics (Figure 4F) as described for P*AtSEOa*-GFP$_{ER}$ (Figure 4C, D).
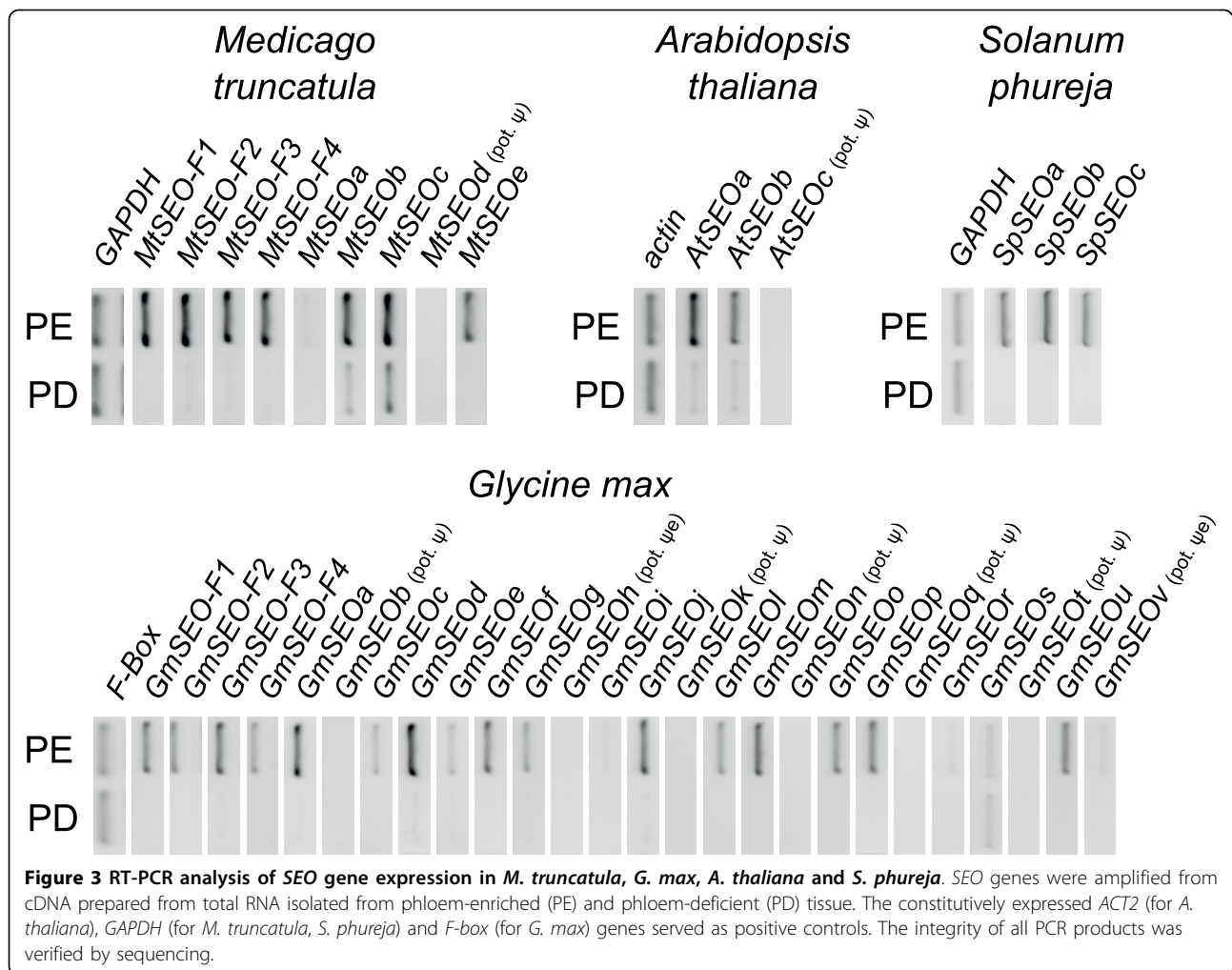
Domain analysis of SEO proteins

The SEO protein sequences were analyzed to identify any conserved protein domains and gain insights into

Figure 2 Schematic overview of *SEO* gene exon-intron structure. Exons are represented as colored boxes, introns as dashed lines. Introns are not drawn to scale, but the length of the individual introns is indicated in base pairs. Mt = *Medicago truncatula*, Gm = *Glycine max*, Md = *Malus domestica*, At = *Arabidopsis thaliana*, Sp = *Solanum phureja*.

their relationship with known protein functions, such as the ability of forisomes to respond to calcium. The deduced amino acid sequences of all SEO proteins were screened against the PfamA and Conserved Domain databases [20,21]. No significant hits were found in the PfamA database, but the Conserved Domain database identified "thioredoxin-like" domains in MtSEO-F2, MtSEOa, GmSEOi, GmSEOl and MdSEOa, which cover several subgroups of the SEO family. Based on this result we constructed a specific Hidden Markov Model by aligning all the identified thioredoxin-like domains, and using this to screen the remaining SEO proteins for similar domains. We obtained significant hits for all the SEO proteins and designated the resulting domain as 'potential thioredoxin fold'. Further analysis of the *M. truncatula* and *A. thaliana* 'potential thioredoxin fold' domains with the protein structure prediction server I-TASSER, and alignment with known structures in the Protein Data Bank (PDB) with TM-align, revealed that SEO proteins are structurally related to tryparedoxin II

(Figure 5), a thioredoxin-like protein [22]. Complete results from I-TASSER and TM-align are provided in Additional file 3. Scanning against predicted proteins from all the plants included in our analysis, the 'potential thioredoxin fold' was also present in several further proteins, including some from monocotyledonous plants, but was not found in *P. patens*. All of these non-SEO proteins contain a PfamA domain belonging to the Pfam-Clan "Thioredoxin-like" indicating that our Hidden Markov Model indeed predicts thioredoxin folds.

We also searched for predicted domains using the PfamB database [20], revealing three domains in all SEO proteins. One domain (PB104124) was non-specific and overlapped the two other predicted PfamB domains, and was therefore rejected from further analysis. The second PfamB domain (PB013523) was predicted in the N-terminal part of the proteins and was therefore named SEO-NTD (SEO N-Terminal Domain), whereas the third (PB006891) spanned the C-terminus. Because

Figure 3 RT-PCR analysis of *SEO* gene expression in *M. truncatula*, *G. max*, *A. thaliana* and *S. phureja*. *SEO* genes were amplified from cDNA prepared from total RNA isolated from phloem-enriched (PE) and phloem-deficient (PD) tissue. The constitutively expressed *ACT2* (for *A. thaliana*), *GAPDH* (for *M. truncatula*, *S. phureja*) and *F-box* (for *G. max*) genes served as positive controls. The integrity of all PCR products was verified by sequencing.

PB006891 partially overlaps with the 'potential thioredoxin fold' we adjusted the domain by building a new Hidden Markov Model that did not interfere with the fold and subsequently renamed it SEO-CTD (SEO C-Terminal Domain). The domain arrangement of the SEO proteins is shown in Figure 6A. It should be noted that the combination of the two PfamB domains together with the 'potential thioredoxin fold' in a single protein could not be identified in any predicted non-SEO proteins from the analyzed plants and appears unique to the SEO family. However, none of these domains alone is specific for SEO proteins - all three domains are present individually in other proteins, although the PfamB domains were identified in only a few proteins from *M. truncatula* and *G. max* (Additional file 4) and in no proteins from *A. thaliana*, *V. vinifera*, the monocotyledonous plants we analyzed or in *P. patens*. *S. phureja* and *M. domestica* were excluded from this search because gene models were not available.

Motif search in SEO proteins

Next, we set out to identify motifs (shorter in length than the domains described above) that are specific for the SEO protein family. Therefore, highly-conserved regions were chosen from an alignment of SEO proteins and were used to construct Hidden Markov Model profiles which characterize the motifs. We identified numerous motifs that appeared to be unique to SEO proteins (Figure 6B) as they could not be detected in any other proteins from any of the other plants included in the analysis. The C-terminal M1 motif, containing several conserved cysteine residues, was representative of the entire SEO family, whereas M2 and M3 appeared to be specific for SEO subgroups, perhaps indicating structural and/or functional specialization. The M4 motif, which was 20-35 amino acids in length and matched an in-house database of disordered regions, was found in the N-terminal portion of SEO proteins in subgroups 5 and 6 (Figure 6B) and could reflect intrinsic
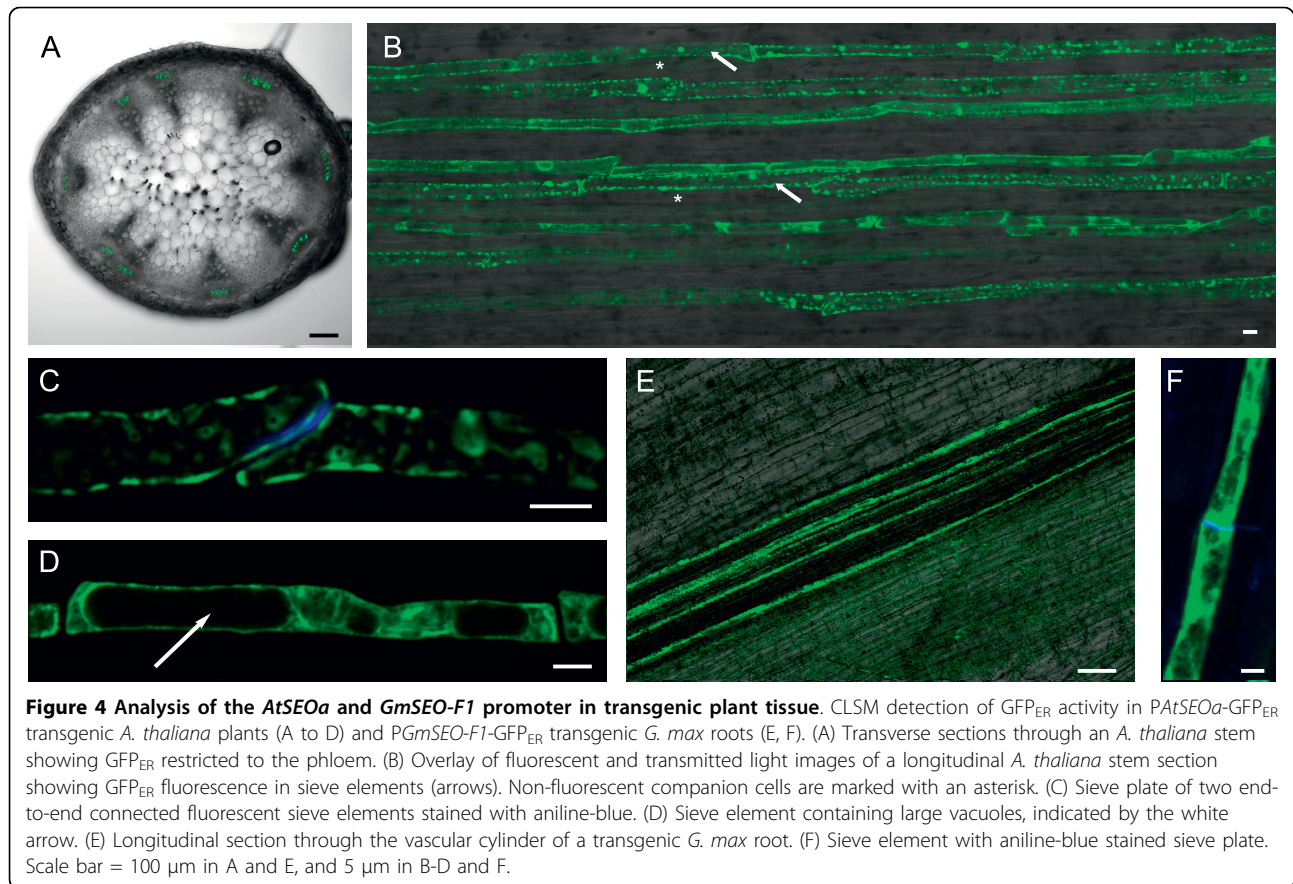
Figure 4 Analysis of the *AtSEOa* and *GmSEO-F1* promoter in transgenic plant tissue. CLSM detection of GFP$_{ER}$ activity in P*AtSEOa*-GFP$_{ER}$ transgenic *A. thaliana* plants (A to D) and P*GmSEO-F1*-GFP$_{ER}$ transgenic *G. max* roots (E, F). (A) Transverse sections through an *A. thaliana* stem showing GFP$_{ER}$ restricted to the phloem. (B) Overlay of fluorescent and transmitted light images of a longitudinal *A. thaliana* stem section showing GFP$_{ER}$ fluorescence in sieve elements (arrows). Non-fluorescent companion cells are marked with an asterisk. (C) Sieve plate of two end-to-end connected fluorescent sieve elements stained with aniline-blue. (D) Sieve element containing large vacuoles, indicated by the white arrow. (E) Longitudinal section through the vascular cylinder of a transgenic *G. max* root. (F) Sieve element with aniline-blue stained sieve plate. Scale bar = 100 μm in A and E, and 5 μm in B-D and F.

disorder within this part of the protein. The individual position of the motifs is indicated in Figure 6A.
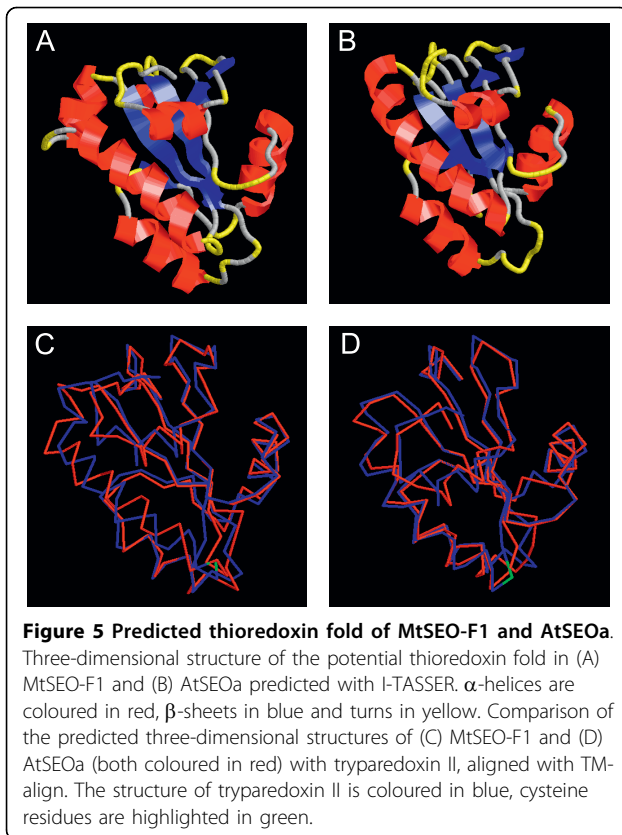
Genomic synteny

Finally, we investigated the organization of *SEO* genes in the *M. truncatula*, *G. max*, *A. thaliana* and *V. vinifera* genomes (*M. domestica* and *S. phureja* were excluded due to the lack of annotated genomic data). As shown in Additional file 5, most *SEO* genes appear to be organized in clusters. Seven of nine *M. truncatula SEO* genes (*MtSEO-F1, 2, 4* and *MtSEOa-d*) are clustered in a 150-kbp segment of chromosome 1, in *G. max*, five *SEO* genes are found in a 50-kbp cluster on chromosome 10, while another six are found in a 55-kbp cluster on chromosome 20. Eight grapevine *SEO* genes are clustered in a 170-kbp segment of chromosome 14 and two of the three *A. thaliana SEO* genes are adjacent on chromosome 3.

The organization of the *G. max* genes is worthy of special attention. *GmSEO-F1, 2* and *4* - the orthologs of the *M. truncatula* forisome-encoding genes *MtSEO-F1, 2* and *4* - are clustered together on a 20-kbp segment of chromosome 10 (Figure 7), but chromosome 20 contains a similarly-arranged cluster comprising the

expressed and non-expressed paralogs *GmSEOv*$_{(potψe)}$, *GmSEOu* and *GmSEOt*$_{(potψ)}$. This similar genomic arrangement could also be verified by analyzing the neighboring non-*SEO* related genes. Out of the 26 *GmSEO* genes 11 paralog pairs could be identified (Figure 7 and Additional file 5). Although we isolated *MtSEO-F3* from cDNA and genomic DNA, we could not identify the sequence in the published *M. truncatula* genome, suggesting a gap in genome coverage. For five out of nine *MtSEO* genes we could identify orthologs in *G. max* (Figure 7 and Additional file 5).

Discussion

Forisomes are specialized P-proteins found solely in the *Fabaceae*. Genes encoding forisome components were recently identified in *M. truncatula*, and are the founder members of the sieve element occlusion (*SEO*) gene family [13-16]. In this study, we identified many additional *SEO* genes in the genomes of two *Fabaceae* and, interestingly, also in several non-*Fabaceae* plant families (which do not possess forisomes). Even within the *Fabaceae*, it appears that only some of the newly-identified *SEO* genes encode forisome components. Phylogenetic analysis showed that all the known and novel *SEO* genes

Figure 5 Predicted thioredoxin fold of MtSEO-F1 and AtSEOa.
Three-dimensional structure of the potential thioredoxin fold in (A)
MtSEO-F1 and (B) AtSEOa predicted with I-TASSER. α-helices are
coloured in red, β-sheets in blue and turns in yellow. Comparison of
the predicted three-dimensional structures of (C) MtSEO-F1 and (D)
AtSEOa (both coloured in red) with tryparedoxin II, aligned with TM-
align. The structure of tryparedoxin II is coloured in blue, cysteine
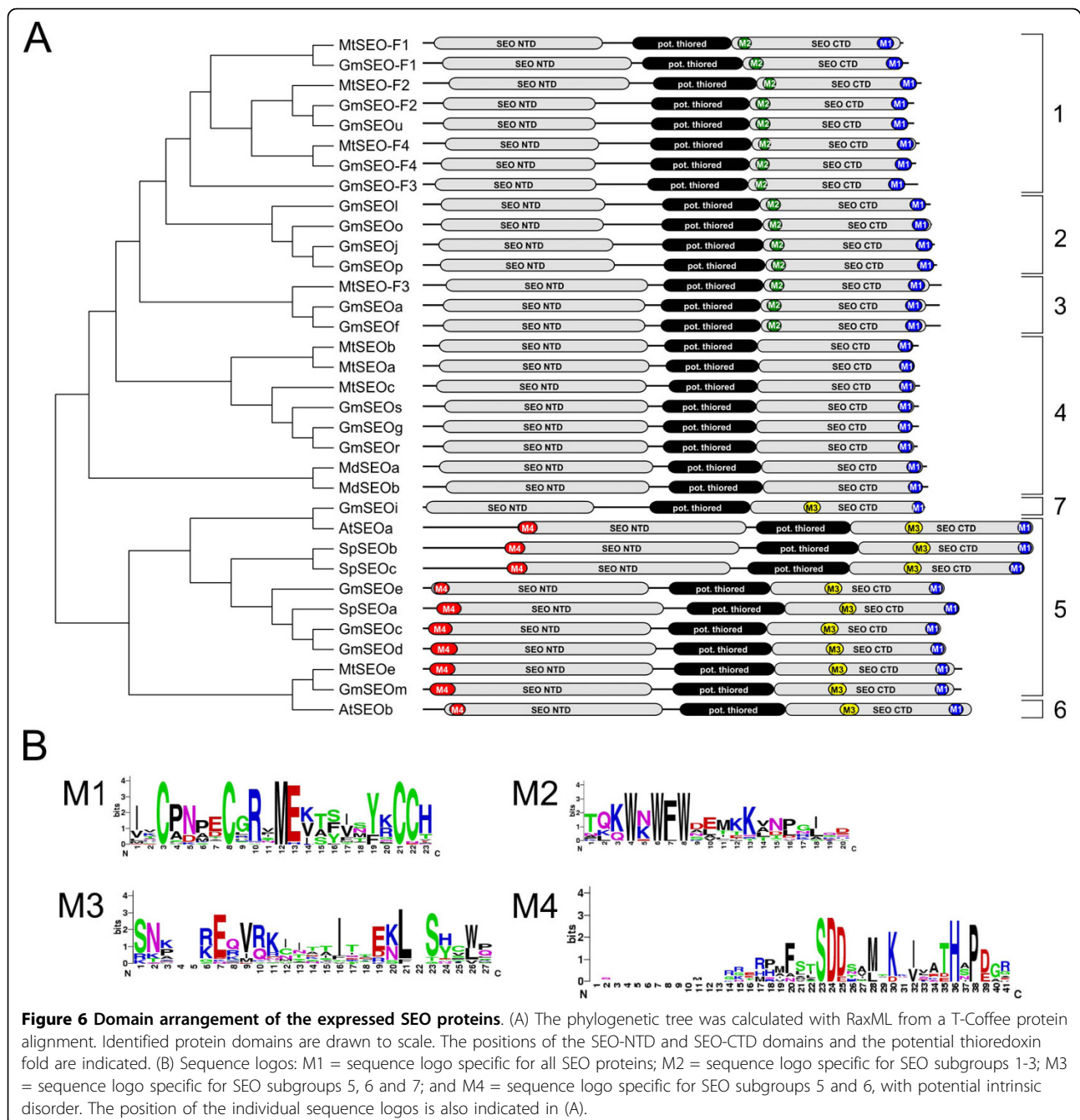residues are highlighted in green.

can be clustered in seven subgroups (Figure 1) and they
have a similar exon-intron structure (Figure 2). Their
expression is likely to be phloem-specific (Figure 3)
most probably restricted to immature sieve elements
(SEs) as shown for one *Fabaceae* and one non-*Fabaceae*
SEO gene (Figure 4) and the proteins contain highly
conserved domains and motifs (Figures 5 and 6).

Forisomes have been described as a special type of P-
protein mainly based on their morphological character-
istics. Like forisomes, P-proteins of various species were
reported to accumulate in immature sieve elements [23]
and in their differentiated state both types of P-protein
share the same function of blocking sieve elements after
phloem injury [8]. Furthermore, forisomes and P-
proteins show very similar ultrastructural characteristics
in both the condensed and dispersed states [24,25]. A
calcium-induced reactivity, as known for forisomes, is
also discussed for P-proteins [26,27]. Although the dis-
persion process is similar in each case, the complete
reversibility of the conformational switch is unique to
forisomes [12]. It therefore seems likely that the non-
forisome *SEO* genes in the *Fabaceae*, and all the *SEO*
genes in non-*Fabaceae* plant families, encode other
(non-forisome) P-proteins. It should be noted that the
Cucurbita maxima PP1 protein, which is the only non-
forisome P-protein to be characterized thus far, shares

neither significant sequence similarities nor any con-
served domains with the SEO proteins described herein.
For this reason, *PP1* should not be assigned to the *SEO*
gene family, despite its potential functional similarity.
Additionally, we and others [18] have not identified any
PP1 orthologs in the genomes of non-*Cucurbitaceae*
plants, suggesting PP1 may play a unique role in the
phloem of the *Cucurbitaceae* family.

We used a number of bioinformatics approaches to
identify functional motifs and domains conserved in all
SEO proteins or in particular phylogenetic clades. We
identified a 'potential thioredoxin fold' domain common
to all SEO proteins, which is also found in enzymes that
catalyze disulfide bond formation [28]. However, the
canonical thioredoxin fold contains two central cysteine
residues that are not present in the SEO proteins, indi-
cating some functional divergence. Interestingly, thiore-
doxin folds lacking cysteines have also been found in
other calcium-binding proteins such as calsequestrin
[29]. Calsequestrin has three such domains that con-
dense to form an acidic platform for high-capacity but
low-affinity calcium adsorption that is most likely non-
specific [30]. Calcium binding in forisomes is also weak
and probably non-specific given that other divalent
cations can also induce the typical conformational
change [31,12]. The core of the calsequestrin thiore-
doxin fold domain is a five-strand β-sheet sandwiched
by four α-helices [29]. A similar arrangement of α-
helices and β-sheets is predicted within the thioredoxin
domain of all SEO proteins analyzed with I-TASSER,
which suggests that the modified thioredoxin fold could
also be involved in calcium binding and the subsequent
dispersion of forisomes and other P-proteins. Although
thioredoxin folds are found in many different proteins,
the presence of this domain together with the two
PfamB domains we identified seems to be a unique
characteristic of SEO proteins. Single PfamB domains
were also identified in *Fabaceae* non-SEO proteins, but
not in non-*Fabaceae* plants, which suggests the transfer
of these domains from *SEO* to non-*SEO* genes by
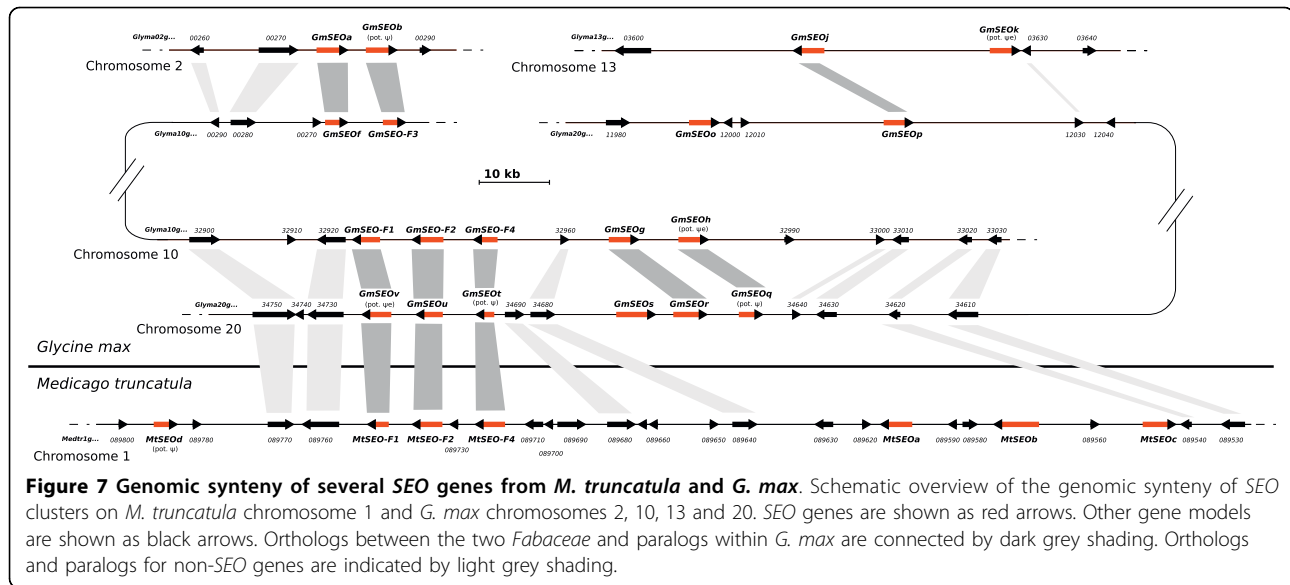domain rearrangement [32,33].

All the SEO proteins contain the conserved C-
terminal motif M1, which is characterized by four spa-
tially conserved cysteine residues (Figure 6B). Although
these residues could form disulfide bridges, this requires
an oxidizing environment and disulfide bridges are gen-
erally not present in cytosolic proteins [34]. However,
given the unique function of P-proteins and forisomes,
it is possible disulfide bridges could form when the
redox state of the cytosol is disrupted by cellular
damage, stabilizing the dispersed state of SEO proteins
and allowing them to seal off the injured sieve element.
Indeed, forisome reactivity declines significantly in the
presence of oxygen [30]. We identified several other

Figure 6 Domain arrangement of the expressed SEO proteins. (A) The phylogenetic tree was calculated with RaxML from a T-Coffee protein alignment. Identified protein domains are drawn to scale. The positions of the SEO-NTD and SEO-CTD domains and the potential thioredoxin fold are indicated. (B) Sequence logos: M1 = sequence logo specific for all SEO proteins; M2 = sequence logo specific for SEO subgroups 1-3; M3 = sequence logo specific for SEO subgroups 5, 6 and 7; and M4 = sequence logo specific for SEO subgroups 5 and 6, with potential intrinsic disorder. The position of the individual sequence logos is also indicated in (A).

motifs that were representative of SEO protein subgroups, e.g. motif M4 with potential intrinsic disorder, found at the N-terminus of subgroups 5 and 6 (Figure 6B). Disordered regions do not have a fixed three-dimensional structure, but can be involved in a variety of different molecular processes such as DNA/RNA or protein binding [35]. However, their precise role(s) in SEO proteins remains unclear.

To investigate the evolution of the *SEO* gene family, we studied the distribution and organization of *SEO*

genes in *Fabaceae* and non-*Fabaceae* genomes (Figure 7 and Additional file 5) and their phylogenetic division into seven subgroups (Figures 1 and 8). Because subgroup 5 contains *SEO* genes from all the dicotyledonous plants with sequenced genomes included in our investigation, it is likely a similar ancestral *SEO* gene pre-dated the split between the rosids and asterids. Subgroup 4 contains only *SEO* genes from the closely-related plant families *Fabaceae* and *Rosaceae*, which suggests the subgroup 4 genes evolved by duplication and mutation

Figure 7 Genomic synteny of several *SEO* genes from *M. truncatula* and *G. max*. Schematic overview of the genomic synteny of *SEO* clusters on *M. truncatula* chromosome 1 and *G. max* chromosomes 2, 10, 13 and 20. *SEO* genes are shown as red arrows. Other gene models are shown as black arrows. Orthologs between the two *Fabaceae* and paralogs within *G. max* are connected by dark grey shading. Orthologs and paralogs for non-*SEO* genes are indicated by light grey shading.

prior to the divergence of the two plant families. With the exception of *MtSEO-F3*, all *SEO-F* genes cluster in the *Fabaceae*-specific subgroup 1 indicating they are unique within the plant kingdom. The function of the *GmSEO* genes clustering in subgroups 2 and 3 remains unclear. Although showing significant similarity to *SEO-F* genes, the corresponding proteins were not detected in forisomes. However, their detection might be affected by a low abundance in forisomes.

The distribution and organization of *SEO* genes in *M. truncatula* and *G. max* genomes indicates that several

Figure 8 Phylogenetic tree of the plants included in this investigation. Phylogenetic tree of the plants included in this investigation according to the Angiosperm Phylogeny Group [75]. Clades and families are shown within the tree, and the numbers of known *SEO* genes are indicated.

gene duplication events have occurred during the evolution of the *Fabaceae SEO-F* genes. With one known exception, *M. truncatula* SEO genes are clustered on chromosome 1 (the location of *MtSEO-F3* is unknown), suggesting proliferation through tandem duplication. There is evidence that most *Fabaceae* share a common whole genome duplication event, which occurred approximately 50-60 million years ago, before *G. max* and *M. truncatula* diverged from a common ancestor [36,37]. Although this event cannot be verified by analyzing the position of *SEO* genes in the *M. truncatula* genome, it is possible that the evidence has been obscured by incidental gene loss events, since orthologs of several of the *SEO* genes present in *G. max* and other plants are missing in *M. truncatula*. For *G. max*, an additional whole genome duplication event is thought to have occurred ~15 million years ago [38,39]. This is supported by the observed synteny in the *SEO* gene clusters on *M. truncatula* chromosome 1, *G. max* chromosome 10, and the paralogs on *G. max* chromosome 20 (Figure 7). Definitive *G. max* orthologs could be identified for five of the nine *M. truncatula SEO* genes, and most of them exist as paralogs in *G. max*, leading to the conclusion that they were present prior to the split of the two *Fabaceae* and therefore the duplication event in *G. max*. No equivalent to the *SEO* gene cluster in *A. thaliana* or the arrangement of *SEO* genes in *V. vinifera* could be identified in *G. max* or *M. truncatula*, as their orthologs are not organized in clusters in both *Fabaceae*. Therefore it seems likely that additional and independent gene duplication and reorganization events affected the *SEO* genes in *A. thaliana* and *V. vinifera*. We also identified a number of potential *SEO*
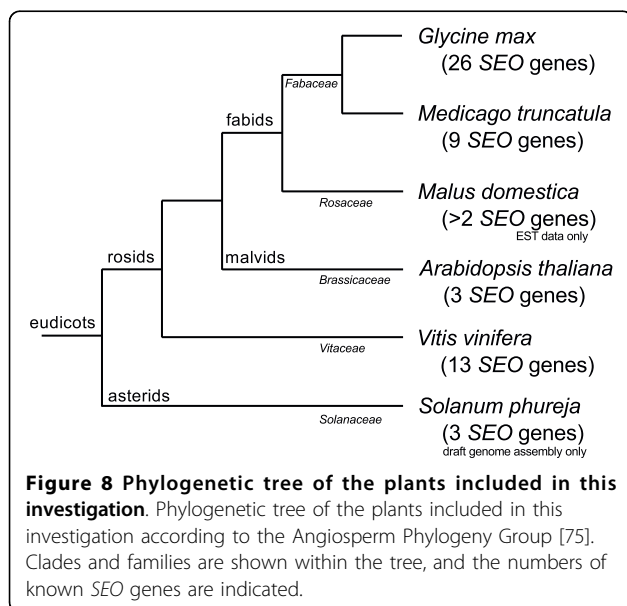
pseudogenes, five of which are found in *G. max*, which probably could be maintained because of the functional redundancy created by gene duplication. Nevertheless, we cannot exclude the possibility that at least the expressed pseudogenes ('pot. ψe') have evolved to take on novel functions, such as the regulation of gene expression [40,41]. *SEO* genes appear to be widespread in dicotyledonous plants and may therefore provide a powerful new tool to study the evolution of gene families in dicotyledons.

Conclusions

We provide evidence that *SEO* genes are widely distributed in non-*Fabaceae* species and most probably encode P-proteins. The strong conservation of the gene structure, protein motifs and domains, the phylogenetic profile and the genomic synteny indicate a common phylogenetic origin for all *SEO* genes. Numerous tandem gene and whole genome duplication events appear to have contributed to the evolution of forisome genes in *Fabaceae*. We identified a fourth *M. truncatula* gene encoding a forisome component and presented the first analysis of forisome genes in *G. max*.

Methods

Identification of gene family members

Protein models were obtained from the sequenced genomes of *Medicago truncatula* [42], *Glycine max* [43], *Arabidopsis thaliana* [44], *Vitis vinifera* [45], *Oryza sativa* [46], *Sorghum bicolor* [47], *Zea mays* [48], *Brachypodium distachyon* [49] and *Physcomitrella patens* [50]. In addition, we used a *Malus domestica* EST collection from the "National Center for Biotechnology Information" and the not yet annotated genome sequence from *Solanum phureja* [51]. With the previously published forisome proteins MtSEO1, MtSEO2 and MtSEO3 [14,15] a BLASTP search was carried out against the protein annotations. Hits with significant similarities (E values lower than 1e-10) were analyzed for global similarity by aligning with the MtSEO proteins. Proteins with only local sequence similarities were not added to the SEO family. To identify further members the BLASTP search was repeated with the newly identified proteins. In addition and in case of missing protein annotations (EST data, non-annotated genomes), the full-length cDNA sequences of all identified *SEO* genes were used in a BLASTN search (threshold E < 0.001). Falsely-annotated *SEO* genes indicated e.g. by the presence of shortened open reading frames were re-annotated from genomic sequences with FGENESH or by aligning the genomic sequence with cDNA sequences of other *SEO* genes. For the identification of correctly spliced cDNA sequences, three independent cDNAs from total seedling RNA (see subheading "Expression

analysis") were used for full-length amplification of the corresponding *SEO* genes using the oligonucleotides listed in Additional file 6. Sequence-verified *SEO* genes were deposited in GenBank (accession numbers are listed in Additional file 1). The exon-intron structure of the *SEO* genes was identified by comparing genomic and cDNA sequences. For *MtSEO-F3*, and the *MdSEO* and *SpSEO* genes, for which no or only preliminary genomic sequence data were available, genomic clones were amplified *de novo* by PCR.

Forisome isolation and peptide sequencing

Forisomes were isolated from purified *M. truncatula* and *G. max* phloem tissue (see next section) by density gradient centrifugation according to established protocols [12]. After fractionation by SDS-PAGE the major 75-kDa protein band was excised from the gel matrix, purified and subsequently characterized by ESI-MS/MS as described [13,14]. The resulting peptide masses were screened against a database containing the SEO proteins and only those peptides unique for a single SEO protein were considered for further analysis.

Expression analysis

Phloem-enriched tissue was prepared from *S. phureja* and *G. max* cv. Williams 82 by scraping the inner side of peeled stem rinds with a scalpel. The remaining stem rind was used as phloem-deficient material. *M. truncatula* cv. Jemalong A17 phloem was enriched by cutting stems in half longitudinally, removing the pith and scraping off the cortex with a scalpel. The cortex of the stem was used as phloem-deficient material. For *A. thaliana* cv. Col-0 phloem-enriched tissue was obtained by cutting out midribs from young leaves. For control experiments with phloem-deficient material we used parts of leaves lacking visible veins.

Total RNA was isolated from tissues ground to powder under liquid nitrogen using the NucleoSpin RNA® Plant Kit (Macherey-Nagel, Düren, Germany). Total RNA was reverse transcribed with SuperScript II (Invitrogen, Karlsruhe, Germany) following the manufacturer's instructions. For all PCRs, partial but intron-spanning parts of the *SEO* genes were amplified using the oligonucleotides listed in Additional file 6. The integrity of all PCR products was verified by sequence analysis. If no products were generated, additional PCRs were performed using different combinations of primers on cDNA derived from young and old leaves, shoots, roots, buds and flowers. Only if no product was detected in this additional experiment were the corresponding *SEO* genes designated as potential pseudogenes. Expressed *SEO* genes containing frameshift mutations were designated as potential expressed pseudogenes. The expression of *M. truncatula GAPDH* [52], *A.*

thaliana ACT2 [53], *S. phureja GAPDH* [54] and the *G. max* F-box gene [55] were used as positive controls.

Promoter analysis

The P*AtSEOa*-GFP$_{ER}$ construct was generated by amplifying a 997-bp *AtSEOa* promoter-specific fragment (P*AtSEOa*) from *A. thaliana* genomic DNA, while the P*GmSEO-F1*-GFP$_{ER}$ construct was generated by amplifying a 2500-bp *GmSEO-F1* promoter-specific fragment (P*GmSEO-F1*) from *G. max* genomic DNA (oligonucleotides listed in Additional file 6). Both PCR products were digested with *Kpn*I and *Xho*I and inserted into the corresponding restriction sites of pBSGFP$_{ER}$, containing the downstream ER-targeted GFP coding region [14]. The promoter-GFP$_{ER}$ constructs were then excised and transferred into the *Kpn*I/*Hin*dIII sites of the binary vector pBIN19 [56] to obtain pBP*AtSEOa*-GFP$_{ER}$ and pBP*GmSEO-F1*-GFP$_{ER}$, respectively. The binary vector pBP*AtSEOa*-GFP$_{ER}$ was introduced into *Agrobacterium tumefaciens* LBA4404 [57] and transformation of *A. thaliana* was carried out by floral dip [58]. Seeds were sterilized and germinated on Murashige and Skoog medium [59] supplemented with 50 μg/ml kanamycin for the selection of transgenic plants. GFP$_{ER}$ expression was monitored in transverse and longitudinal stem sections by confocal laser scanning microscopy (CLSM; Leica TCS SP5 X, Wetzlar, Germany; excitation 488 nm, emission 500-600 nm). Sieve plates were stained with a 0.01% aniline-blue solution according to established protocols [60] and visualized by CLSM (excitation 364 nm, emission 470-530 nm). The binary vector pBP*GmSEO-F1*-GFP$_{ER}$ was introduced into *A. rhizogenes* strain NCPPB2659 and transgenic *G. max* roots were obtained following established protocols [61]. Longitudinal sections of the roots were analyzed as above.

Phylogenetic analysis

The OrthoMCL program [62] was used to cluster SEO proteins into subgroups, with an inflation parameter of 3. The protein sequences were aligned with T-Coffee [63] and the alignment was end trimmed to start and end with the domains predicted for all SEO proteins (SEO-NTD and SEO-CTD). The optimal evolutionary model for the family was calculated from this alignment with ProtTest [64]. RAxML [65] was used for tree building with the evolutionary model parameter JTT+F+I+G and a bootstrap of 1000. The best tree was visualized with FigTree [66].

Domain analysis

Domain annotation was achieved by screening the NCBI Conserved Domain Database (v2.17) [21] in combination with CD-Search [67] and the PfamA and PfamB databases v23.0 [20], using a significance threshold of 1e-05.

Domain arrangements were visualized using Jangstd [68]. HMMER 3.0 beta 2 was used to construct Hidden Markov Models (HMMs) and carry out searches [69]. Unique motifs in the SEO family were identified by extracting partial alignments to construct HMMs. Three dimensional protein structure prediction was carried out with I-TASSER [70]. The resulting protein structures were compared to structures in the Protein Data Bank (PDB) using the program TM-align [71]. Sequence logos of alignments were generated with WebLogo [72]. To identify regions of disorder, an in-house database was created by scanning protein sequences from annotated plant genomes with VSL2B [73]. Disordered sequences with a length of at least 20 amino acids were clustered with cd-hit [74], aligned and used to build HMMs. Disordered regions in the SEO family were predicted with this set of HMMs. All models were tested against SEO proteins as well as all available protein predictions for the plants included in our investigation.

Additional material

Additional file 1: Table of all *SEO* genes included in this investigation. Summary of all known *SEO* genes identified in *Medicago truncatula* (Mt), *Glycine max* (Gm), *Malus domestica* (Md), *Arabidopsis thaliana* (At), *Vitis vinifera* (Vv), *Solanum phureja* (Sp), *Vicia faba* (Vf), *Pisum sativum* (Ps) and *Canavalia gladiata* (Cg). The E values for the different protein products result from a BLASTp search with MtSEO-F1 against a protein database containing all identified SEO proteins.

Additional file 2: Table of peptide sequences obtained from forisomes. Assigned peptide sequences generated by ESI-MS/MS from purified forisomes from *Medicago truncatula* (Mt) and *Glycine max* (Gm).

Additional file 3: Table with results for structure prediction of the potential thioredoxin fold with I-TASSER and TM-align. Accuracy of the predicted thioredoxin fold structure of MtSEO and AtSEO proteins with I-TASSER, and of the TM-alignment with tryparedoxin ll.

Additional file 4: Table of non-SEO proteins containing SEO domains. List of predicted non-SEO proteins from *Medicago truncatula* (Medtr) and *Glycine max* (Glyma) carrying the SEO N-terminal domain (NTD) or SEO C-terminal domain (CTD).

Additional file 5: Chromosomal organization of the *SEO* genes. Schematic overview of *SEO* chromosomal loci in the plants included in this investigation. Orthologs between *MtSEO* and *GmSEO* genes as well as *SEO* paralogs in *G. max* are marked with matching symbols. Chromosomes and genes are not drawn to scale.

Additional file 6: List of oligonucleotides.

Author details
[1]Institut für Biochemie und Biotechnologie der Pflanzen, Westfälische Wilhelms-Universität Münster, Hindenburgplatz 55, D-48143 Münster, Germany. [2]Fraunhofer Institute for Molecular Biology and Applied Ecology (IME), Forckenbeckstraße 6, D-52074 Aachen, Germany. [3]Institut für Evolution und Biodiversität, Westfälische Wilhelms-Universität Münster, Hüfferstraße 1, D-48149 Münster, Germany.

Authors' contributions

BR identified the *SEO* gene family members, cloned the *MtSEO* and *SpSEO* genes, performed the computational and domain analyses and drafted the manuscript. AME identified and cloned the *GmSEO* genes, isolated forisomes from *G. max* and *M. truncatula*, performed the P*GmSEO-F1* promoter studies and participated in drafting the manuscript. SBJ cloned the *AtSEO* genes and performed the P*AtSEOa* promoter studies. SN participated in cloning the *GmSEO* genes. ARR supported the computational analyses and built the disordered region database. BM participated in analyzing the potential thioredoxin fold. EBB conceived the computational analyses and helped drafting the manuscript. DP participated in conceiving, design and coordination of this study and revised the manuscript. GAN conceived the study, participated in its design and helped to draft the manuscript. All authors read and approved the final manuscript.

References

1. Esau K: *The Phloem* Stuttgart: Borntraeger 1969.
2. Münch E: *Die Stoffbewegungen in der Pflanze* Jena: Fischer Verlag 1930.
3. Sjolund RD: **The phloem sieve element: a river runs through it.** *Plant Cell* 1997, **9**:1137-1146.
4. van Bel A, Kempers R: **The pore/plasmodesm unit: key element in the interplay between sieve element and companion cell.** *Progress in Botany* 1996, **58**:278-291.
5. Geiger DR, Giaquinta RT, Sovonick SA, Fellows RJ: **Solute Distribution in Sugar Beet Leaves in Relation to Phloem Loading and Translocation.** *Plant Physiol* 1973, **52**:585-589.
6. Esau K, Cronshaw J: **Tubular components in cells of healthy and tobacco mosaic virus-infected *Nicotiana*.** *Virology* 1967, **33**:26-35.
7. Ehlers K, Knoblauch M, van Bel A: **Ultrastructural features of well-preserved and injured sieve elements: Minute clamps keep the phloem transport conduits free for mass flow.** *Protoplasma* 2000, **214**:80-92.
8. Knoblauch M, van Bel AJE: **Sieve Tubes in Action.** *Plant Cell* 1998, **10**:35-50.
9. Cronshaw J: **Phloem structure and function.** *Annu Rev Plant Physiol* 1981, **32**:465-484.
10. Wergin WP, Palevitz BA, Newcomb EH: **Structure and development of P-protein in phloem parenchyma and companion cells of legumes.** *Tissue Cell* 1975, **7**:227-242.
11. Behnke H: **Nondispersive protein bodies in sieve elements: a survey and review of their origin, distribution and taxonomic significance.** *IAWA Bull* 1991, **12**:143-175.
12. Knoblauch M, Noll GA, Müller T, Prüfer D, Schneider-Hüther I, Scharner D, van Bel AJE, Peters WS: **ATP-independent contractile proteins from plants.** *Nat Mater* 2003, **2**:600-603.
13. Noll GA: **Molekularbiologische Charakterisierung der Forisome.** *PhD thesis* Justus-Liebig Universität Gießen 2005.
14. Noll G, Fontanellaz M, Rüping B, Ashoub A, van Bel AJE, Fischer R, Knoblauch M, Prüfer D: **Spatial and temporal regulation of the forisome gene *for1* in the phloem during plant development.** *Plant Mol Biol* 2007, **65**:285-294.
15. Pélissier HC, Peters WS, Collier R, van Bel AJE, Knoblauch M: **GFP Tagging of Sieve Element Occlusion (SEO) Proteins Results in Green Fluorescent Forisomes.** *Plant Cell Physiol* 2008, **49**:1699-1710.
16. Noll G, Rüping B, Ernst A, Bucsenez M, Twyman R, Fischer R, Prüfer D: **The Promoters of Forisome Genes *MtSEO2* and *MtSEO3* Direct Gene Expression to Immature Sieve Elements in *Medicago truncatula* and *Nicotiana tabacum*.** *Plant Mol Biol Rep* 2009, **27**:526-533.
17. Beyenbach J, Weber C, Kleinig H: **Sieve-Tube Proteins from *Cucurbita maxima*.** *Planta* 1974, **119**:113-124.
18. Clark AM, Jacobsen KR, Bostwick DE, Dannenhoffer JM, Skaggs MI, Thompson GA: **Molecular characterization of a phloem-specific gene encoding the filament protein, Phloem Protein 1 (PP1), from *Cucurbita maxima*.** *Plant J* 1997, **12**:49-61.
19. Tuteja N, Umate P, van Bel AJE: **Forisomes: calcium-powered protein complexes with potential as 'smart' biomaterials.** *Trend Biotechnol* 2010, **28**:102-110.
20. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz H, Ceric G, Forslund K, Eddy SR, Sonnhammer ELL, Bateman A: **The Pfam protein families database.** *Nucl Acids Res* 2008, **36**:D281-288.

21. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, He S, Hurwitz Dl, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Tasneem A, Thanki N, Yamashita RA, Zhang D, Zhang N, Bryant SH: **CDD: specific functional annotation with the Conserved Domain Database.** *Nucl Acids Res* 2009, **37**:D205-210.
22. Hofmann B, Budde H, Bruns K, Guerrero SA, Kalisz HM, Menge U, Montemartini M, Nogoceke E, Steinert P, Wissing JB, Flohé L, Hecht HJ: **Structures of tryparedoxins revealing interaction with trypanothione.** *Biol Chem* 2001, **382**:459-471.
23. Cronshaw J, Esau K: **Tubular and fibrillar components of mature and differentiating sieve elements.** *J Cell Biol* 1967, **34**:801-815.
24. Wu J, Hao B: **Ultrastructure of P-protein in *Hevea brasiliensis* during sieve-tube development and after wounding.** *Protoplasma* 1990, **153**:186-192.
25. Knoblauch M, Peters WS: **Forisomes, a novel type of Ca^{2+}-dependent contractile protein motor.** *Cell Mot Cyto* 2004, **58**:137-142.
26. Knoblauch M, Peters WS, Ehlers K, van Bel AJE: **Reversible calcium-regulated stopcocks in legume sieve tubes.** *Plant Cell* 2001, **13**:1221-1230.
27. King RW, Zeevaart JAD: **Enhancement of phloem exudation from cut petioles by chelating agents.** *Plant Physiol* 1974, **53**:96-103.
28. Atkinson HJ, Babbitt PC: **An Atlas of the Thioredoxin Fold Class Reveals the Complexity of Function-Enabling Adaptations.** *PLoS Comput Biol* 2009, **5**:e1000541.
29. Wang S, Trumble WR, Liao H, Wesson CR, Dunker AK, Kang C: **Crystal structure of calsequestrin from rabbit skeletal muscle sarcoplasmic reticulum.** *Nat Struct Mol Biol* 1998, **5**:476-483.
30. MacLennan DH, Reithmeier RAF: **Ion tamers.** *Nat Struct Mol Biol* 1998, **5**:409-411.
31. Schwan S, Fritzsche M, Cismak A, Heilmann A, Spohn U: **In vitro investigation of the geometric contraction behavior of chemo-mechanical P-protein aggregates (forisomes).** *Biophys Chem* 2007, **125**:444-452.
32. Moore AD, Björklund ÅK, Ekman D, Bornberg-Bauer E, Elofsson A: **Arrangements in the modular evolution of proteins.** *Trends Biochem Sci* 2008, **33**:444-451.
33. Weiner J, Beaussart F, Bornberg-Bauer E: **Domain deletions and substitutions in the modular protein evolution.** *FEBS J* 2006, **273**:2037-2047.
34. Freedman RB: **The formation of protein disulphide bonds.** *Curr Opin Struct Biol* 1995, **5**:85-91.
35. Chen JW, Romero P, Uversky VN, Dunker AK: **Conservation of Intrinsic Disorder in Protein Domains and Families: II. Functions of Conserved Disorder.** *J Proteome Res* 2006, **5**:888-898.
36. Pfeil BE, Schlueter JA, Shoemaker RC, Doyle JJ: **Placing Paleopolyploidy in Relation to Taxon Divergence: A Phylogenetic Analysis in Legumes Using 39 Gene Families.** *Syst Biol* 2005, **54**:441-454.
37. Cronk Q, Ojeda I, Pennington RT: **Legume comparative genomics: progress in phylogenetics and phylogenomics.** *Curr Opin Plant Biol* 2006, **9**:99-103.
38. Zhu H, Choi HK, Cook DR, Shoemaker RC: **Bridging model and crop legumes through comparative genomics.** *Plant Physiol* 2005, **137**:1189-1196.
39. Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K: **The flowering world: a tale of duplications.** *Trends Plant Sci* 2009, **14**:680-688.
40. Balakirev ES, Ayala FJ: **Pseudogenes: Are They "Junk" or Functional DNA?** *Annu Rev Genet* 2003, **37**:123-151.
41. Raes J, Van de Peer Y: **Functional divergence of proteins through frameshift mutations.** *Trends Genet* 2005, **21**:428-431.
42. US/EU *Medicago truncatula* genome sequencing project - release 3.0. [http://www.medicago.org/genome/downloads/Mt3/].
43. Soybean Genome Project, DoE Joint Genome Institute - gene set "Glyma1.0". [ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v4.1/Gmax/].
44. The Arabidopsis Information Resource Genome Release 8. [ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR8_genome_release/].
45. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A,

Scarpelli C, Artiguenave F, Pè ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quètier F, Wincker P: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.** *Nature* 2007, **449**:463-467.

46. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, Orvis J, Haas B, Wortman J, Buell CR: **The TIGR Rice Genome Annotation Resource: improvements and new features.** *Nucl Acids Res* 2007, **35**:D883-887.

47. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob-ur-Rahman , Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS: **The *Sorghum bicolor* genome and the diversification of grasses.** *Nature* 2009, **457**:551-556.

48. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock S, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli T, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK: **The B73 Maize Genome: Complexity, Diversity, and Dynamics.** *Science* 2009, **326**:1112-1115.

49. The *Brachypodium distachyon* Information Resource. [http://www.brachypodium.org/].

50. Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud P, Lindquist EA, Kamisugi Y, Tanahashi T, Sakakibara K, Fujita T, Oishi K, Shin-I T, Kuroki Y, Toyoda A, Suzuki Y, Hashimoto S, Yamaguchi K, Sugano S, Kohara Y, Fujiyama A, Anterola A, Aoki S, Ashton N, Barbazuk WB, Barker E, Bennetzen JL, Blankenship R, Cho SH, Dutcher SK, Estelle M, Fawcett JA, Gundlach H, Hanada K, Heyl A, Hicks KA, Hughes J, Lohr M, Mayer K, Melkozernov A, Murata T, Nelson DR, Pils B, Prigge M, Reiss B, Renner T, Rombauts S, Rushton PJ, Sanderfoot A, Schween G, Shiu SH, Stueber K, Theodoulou FL, Tu H, Van de Peer Y, Verrier PJ, Waters E, Wood A, Yang L, Cove D, Cuming AC, Hasebe M, Lucas S, Mishler BD, Reski R, Grigoriev IV, Quatrano RS, Boore JL: **The *Physcomitrella* Genome Reveals Evolutionary Insights into the Conquest of Land by Plants.** *Science* 2008, **319**:64-69.

51. Potato Genome Sequencing Consortium - whole genome assembly scaffolds V1. [http://potatogenomics.plantbiology.msu.edu/].

52. Schiene K, Donath S, Brecht M, Pühler A, Niehaus K: **A Rab-related small GTP binding protein is predominantly expressed in root nodules of *Medicago sativa*.** *Mol Genet Genomics* 2004, **272**:57-66.

53. An YQ, McDowell JM, Huang S, McKinney EC, Chambliss S, Meagher RB: **Strong, constitutive expression of the *Arabidopsis* ACT2/ACT8 actin subclass in vegetative tissues.** *Plant J* 1996, **10**:107-121.

54. Guo J, Yang Q, Liang F, Xing Y, Wang Z: **Molecular cloning and expression analysis of a novel CONSTANS-like gene from potato.** *Biochemistry Mosc* 2007, **72**:1241-1246.

55. Libault M, Thibivilliers S, Bilgin DD, Radwan O, Benitez M, Clough SJ, Stacey G: **Identification of Four Soybean Reference Genes for Gene Expression Normalization.** *Plant Genome* 2008, **1**:44-54.

56. Bevan M: **Binary *Agrobacterium* vectors for plant transformation.** *Nucl Acids Res* 1984, **12**:8711-8721.

57. Hoekema A, Hirsch PR, Hooykaas PJJ, Schilperoort RA: **A binary plant vector strategy based on separation of *vir*- and T-region of the *Agrobacterium tumefaciens* Ti-plasmid.** *Nature* 1983, **303**:179-180.

58. Clough SJ, Bent AF: **Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*.** *Plant J* 1998, **16**:735-743.

59. Murashige T, Skoog F: **A revised medium for rapid growth and bioassays with tobacco tissue cultures.** *Physiol plant* 1962, **15**:473-97.

60. Thompson MV, Wolniak SM: **A Plasma Membrane-Anchored Fluorescent Protein Fusion Illuminates Sieve Element Plasma Membranes in Arabidopsis and Tobacco.** *Plant Physiol* 2008, **146**:1599-1610.

61. Collier R, Fuchs B, Walter N, Lutke WK, Taylor CG: **Ex vitro composite plants: an inexpensive, rapid method for root biology.** *Plant J* 2005, **43**:449-457.

62. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes.** *Genome Res* 2003, **13**:2178-2189.

63. Notredame C, Higgins DG, Heringa J: **T-coffee: a novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302**:205-217.

64. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution.** *Bioinformatics* 2005, **21**:2104-2105.

65. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**:2688-2690.

66. FigTree. [http://tree.bio.ed.ac.uk/software/figtree/].

67. Marchler-Bauer A, Bryant SH: **CD-Search: protein domain annotations on the fly.** *Nucl Acids Res* 2004, **32**:W327-331.

68. Jangstd. [http://iebservices.uni-muenster.de/angstd/download/].

69. HMMER. [http://hmmer.janelia.org/].

70. Zhang Y: **I-TASSER server for protein 3D structure prediction.** *BMC Bioinformatics* 2008, **9**:40.

71. Zhang Y, Skolnick J: **TM-align: a protein structure alignment algorithm based on the TM-score.** *Nucl Acids Res* 2005, **33**:2302-2309.

72. Crooks GE, Hon G, Chandonia J, Brenner SE: **WebLogo: A Sequence Logo Generator.** *Genome Res* 2004, **14**:1188-1190.

73. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z: **Length-dependent prediction of protein intrinsic disorder.** *BMC Bioinformatics* 2006, **7**:208.

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

75. APG III: **An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III.** *Bot J Linn Soc* 2009, **161**:105-121.