Stepwise evolution of a buried inhibitor peptide over 45 million years

Achala S. Jayasena¹, Mark F. Fisher¹, Jose L. Panero², David Secco¹, Kalia Bernath-Levin¹, Oliver Berkowitz³, Nicolas L. Taylor¹, Edward E. Schilling⁴, James Whelan³, and Joshua S. Mylne^{1, *}

¹ The University of Western Australia, School of Molecular Sciences & ARC Centre of Excellence in Plant Energy Biology, 35 Stirling Highway, Crawley, Perth 6009, Australia.

² University of Texas, Section of Integrative Biology, 1 University Station C0930, Austin, TX 78712, USA.

³ La Trobe University, Department of Botany, School of Life Sciences & ARC Centre of Excellence in Plant Energy Biology, AgriBio, the Centre for AgriBioscience, 5 Ring Road, Bundoora Victoria 3086, Australia.

⁴ University of Tennessee, Department of Ecology & Evolutionary Biology, Knoxville, TN 37996, USA.

* Corresponding author: E-mail: joshua.mylne@uwa.edu.au

Abstract

The *de novo* evolution of genes and the novel proteins they encode has stimulated much interest in the contribution such innovations make to the diversity of life. Most research on this *de novo* evolution focuses on transcripts, so studies on the biochemical steps that can enable completely new proteins to evolve and the time required to do so have been lacking. Sunflower PawS1 is an unusual albumin precursor because in addition to producing albumin it also yields a potent, bicyclic protease-inhibitor called SFTI-1. Here we show how this inhibitor peptide evolved stepwise over tens of millions of years. To trace the origin of the inhibitor peptide SFTI-1 we assembled seed transcriptomes for 110 sunflower relatives whose evolution could be resolved by a chronogram, which allowed dates to be estimated for the various stages of molecular evolution. A genetic insertion event in an albumin precursor gene ~45 million years ago introduced two additional cleavage sites for protein maturation and conferred duality upon *PawS1-Like* genes such that they also encode a small buried macrocycle. Expansion of this region, including two Cys residues, enlarged the peptide ~34 million years ago and made the buried peptides bicyclic. Functional specialization into a protease inhibitor occurred ~23 million years ago. These findings document the evolution of a novel peptide inside a benign region of a pre-existing protein. We illustrate how a novel peptide can evolve without *de novo* gene evolution and, critically, without affecting the function of what becomes the protein host.

One sentence summary: A peptide is born within another protein and evolves stepwise into a protease inhibitor

Introduction

Proteins often evolve by duplication followed by divergence of their encoding genes, but it is also apparent that proteins can evolve more suddenly by *de novo* gene evolution (Bornberg-Bauer, et al. 2010; Tautz and Domazet-Lošo 2011; McLysaght and Guerzoni 2015; McLysaght and Hurst 2016). The rise of next generation sequencing and comparative genomics have revealed many examples of transcripts coming from ancestrally non-genic regions. Pioneering studies that provided the first evidence of *de novo* evolved genes (Levine, et al. 2006; Begun, et al. 2007; Zhou, et al. 2008) have led to more examples that include hundreds in *Drosophila melanogaster* (Zhao, et al. 2014), *Saccharomyces cerevisiae* (Carvunis, et al. 2012), novel genes in rodents (Heinen, et al. 2009; Murphy and McLysaght 2012) and 60 new protein-coding genes in humans (Wu, et al. 2011). The functionality of genes evolved *de novo* has been demonstrated in *Drosophila* (Chen, et al. 2010; Reinhardt, et al. 2013), mouse (Heinen, et al. 2009), and human (Wu, et al. 2011).

Direct evidence of proteins encoded by *de novo* evolved genes using antibodies, proteomics or by separation science is mostly lacking, although a tagged version of the protein encoded by the *de novo* *MDF1* gene from yeast was detectable by western blot once over-expressed (Li et al., 2010). Proteins may also evolve *de novo* without *de novo* gene evolution. These instances include frameshifts to an open reading frame (ORF) (Ohno 1984) or major genetic rearrangements (Ross, et al. 2013), both of which dramatically alter the originally-encoded protein. Proteins encoded by different reading frames of the same transcript are common in viruses and termed overprinting (Carter, et al. 2013; Pavesi, et al. 2013).

DNA sequence insertions within an existing ORF could also lead to the creation of a novel peptide if the DNA insertion was in a benign region of the encoded protein and proteases acted to excise the new peptide region. In this way, an existing ORF and the protein it encodes could expand such that it produced the original protein as well as a new and unrelated peptide. This would represent the evolution of a new peptide in the absence of any change in coding frame, new transcription or dramatic changes to the protein encoded. For a new mature peptide product to evolve in this way would be subtle as it does not perturb the protein encoded by the ancestral gene, making it distinct from genetic changes that alter the function of an existing ORF (Chen, et al. 1997; Carter, et al. 2013; Ross, et al. 2013).

To determine whether this subtle type of new protein evolution can occur, we focused on an unusual protease inhibitor in the common sunflower (*Helianthus annuus*). What makes this protease inhibitor unusual is that the sequence for this 14-residue, bicyclic trypsin-inhibitor SFTI-1 (SunElower Trypsin Inhibitor-1) is embedded adjacent to an albumin storage protein within a non-intronic open reading frame (ORF) of an albumin precursor protein called PawS1 (Preproalbumin with SFTI-1) (Luckett, et al. 1999; Mylne, et al. 2011) (**Fig. 1**). Albumin is a water-soluble class of seed storage protein sometimes specified as napin-type albumins after their prototypic member from *Brassica napus*. Albumins, like most seed storage proteins, are not highly conserved in sequence, but are structurally stable proteins that are hyper-accumulated during seed development and degraded upon germination as a source of nitrogen and sulphur to the developing seedling. Napin-type albumins have been studied in detail (reviewed recently in (Mylne, et al. 2014)). Most plants contain multiple copies of albumin precursor genes and the encoded proteins are especially rich in glutamine. Albumin precursors are subjected to post-translational processing that includes folding of its conserved disulfide bond network followed by proteolytic cleavages featuring the enzyme asparaginyl endopeptidase, which targets Asn and Asp residues. During this proteolytic maturation, albumins are often converted from linear proteins into a heterodimer of a small and large subunit. Some matured albumins remain monomers after cleavages or trimming of regions flanking the albumin. The sunflower albumin precursor PawS1 is an unusual albumin precursor as it is simultaneously matured by asparaginyl endopeptidase into two proteins; the trypsin inhibitory macrocycle SFTI-1 plus a heterodimeric seed storage albumin (Mylne, et al. 2011), making it represent a test case for a subtle type of evolution that forms a new peptide.

Proteins that are matured by proteases into multiple peptides are common, but three features distinguish PawS1 from most of these; namely differing size, topologies and functions of the mature products. The mature PawS1 seed storage albumin is 10.5 kDa and much larger than SFTI-1, which is 1.5 kDa. The mature PawS1 albumin is cleaved at several points and its topology becomes a heterodimer of linear proteins held together by disulfide bonds whereas SFTI-1 is a disulfide-braced macrocyclic peptide. Finally, seed storage albumin has a nutritive function for the developing seedling whereas SFTI-1 is more likely to be involved in defense (Mylne, et al. 2011).

Plants do not contain trypsins, but have evolved a panoply of trypsin inhibitors. SFTI-1 was initially discovered for its strong affinity for bovine trypsin (Luckett, et al. 1999), but later work showed that SFTI-1 can also inhibit insect trypsins (Elliott, et al. 2014). Protease inhibitors stored in seeds are thought to provide protection from graminivorous insects (Shewry 1999) and some serine protease inhibitors

have been demonstrated to protect plants from insects (Dunse, et al. 2010; Hartl, et al. 2010). It is the Lys5 residue in sunflower SFTI-1 that is the target of trypsin and the surrounding residues create the beta-strand structure required for recognition by proteases (Luckett, et al. 1999; Tyndall, et al. 2005).

SFTI-1 was discovered in 1999 by Luckett *et al.* and not until 2011 was its precursor protein revealed (Mylne, et al. 2011). Several years later it was realised that SFTI-1 was just the first member of a large class of seed peptides similarly buried within preproalbumin. PCR-based discovery of *PawS1* genes from relatives of sunflower revealed a diverse class of PawS-Derived Peptides (PDPs) whose distribution suggests the peptide class is at least 20 million years old (Elliott, et al. 2014). PCR also revealed a gene that might have been an ancestral form, called *PawL1* (*PawS-Like 1*) that had many similarities to *PawS1*, but whose corresponding peptide region lacked Cys residues and appeared to make no stable PDP (Elliott, et al. 2014). The proteins encoded by *PawL1* in *Arnica montana* and sunflower *PawS1* were shown to produce mature, napin-type albumins by purifying and sequencing them by mass spectrometry (Mylne, et al. 2011; Elliott, et al. 2014).

The sunflower family (Asteraceae) is the second largest family of vascular plants, containing over 25,000 species, which represents >8% of all flowering plants (Panero and Crozier 2016). This extensive and recent radiation makes the Asteraceae an ideal family in which to monitor the origins and evolution of a protein. Here we studied the evolution of SFTI-1 by a survey of 110 Asteraceae seed transcriptomes. By coupling the gene and peptide distributions to a chronogram we could retrodict and date several stable biochemical steps that enabled a potent inhibitory peptide to evolve within a latent region of a gene encoding a seed storage albumin.

Results

To trace the origins of SFTI-1, we extracted RNA from the seeds of 110 Asteraceae species, performed RNA-seq and assembled a transcriptome for each species (**Table S1**). The species chosen were mostly from subfamily Asteroideae where *PawS1* genes were discovered (Mylne, et al. 2011; Elliott, et al. 2014), but included more distantly related subfamilies Mutisioideae, Carduoideae and Cichorioideae. We evaluated the best assembly of each transcriptome using six transcripts chosen for their ubiquity (Jayasena, et al. 2014) and because they cover a wide range of mRNA expression levels (**Fig. S1, Dataset 1**). To find *PawS1* and *PawL1* transcripts we used tBLASTn to search each transcriptome using the 151-residue sunflower *PawS1* as the query (**Dataset 2**). Ninety-seven percent of *PawS1*/*PawL1* transcript levels fell within the range of the six core transcripts (**Fig. S1**), which suggested *PawS1*/*PawL1* genes

were absent if a species had no detectable transcript. To confirm assembly quality, we designed primers against sequences from 22 transcriptomes and cloned 39 genomic DNA sequences with complete ORFs that were 99-100% identical to the assembled transcriptome (**Table S2, Dataset 3**). In addition, we profiled peptide extracts from seeds of the same 110 species by liquid chromatography-mass spectrometry (LC-MS) for corresponding peptide support (**Fig. S2**).

To illustrate the distribution of the gene types and estimate the timing of their origins, we generated a chronogram based on nuclear ribosomal internal transcribed spacer (ITS) sequences and chloroplast genes that included the species for which we obtained transcriptomes. Four nodes were calibrated using fossil evidence and a major tectonic event (**Fig. 2**). The gene types were mapped to the chronogram and the timing for steps between gene types could then be inferred based on dates of the divergence points in our chronogram.

The hypothesis that a single-product, seed storage albumin precursor gene is the ancestor of both *PawL1* and *PawS1* was supported by the observations that transcripts for *SEED STORAGE ALBUMIN* (*SESA*) genes were present in all 110 sampled Asteraceae species and many similar sequences are found in GenBank sourced more widely from dicots. Most plant species have multiple copies of *SESA* genes. The common sunflower has 12 *SESA* genes (Jayasena, et al. 2016), *Arabidopsis* contains five (Krebbbers, et al. 1988; van der Klei, et al. 1993), Brazil nut has at least ten (Moreno, et al. 2004) whereas *Brassica napus* (rapeseed) has over 16 (Scofield and Crouch 1987).

PawS1 or *PawL1* transcripts were absent from the transcriptomes of all 29 sampled species in subfamilies Mutisioideae, Carduoideae, and Cichorioideae, suggesting the evolution of PDPs was confined to the subfamily Asteroideae (**Fig. 2**). We found *PawL1* transcripts more widely in 51 of 81 Asteroideae transcriptomes that spanned twelve of the 17 tribes sampled. The 51 species with *PawL1* transcripts contained from 1 to 13 copies (median 3, mean 3.3). We interpret the broad distribution of *PawL1* genes across the subfamily Asteroideae as evidence for a single origin for this stable gene type that, based on our chronogram (**Fig. 2**), we estimate evolved approximately 45 million years ago (Mya).

The sequences of seven *PawL1* genes were previously known and protein evidence was sought only in *Arnica montana* (Elliott, et al. 2014). The 51 new sequences we report exhibit strong conservation that, despite lacking a Cys (C) pair, have parallels with processing of PDPs (**Fig. 3, Datasets 4-6**). These include an absolutely conserved asparaginyl endopeptidase target sequence Asn (N), a highly conserved central Pro (P), a strongly conserved Asp (D) also targeted by asparaginyl endopeptidase. The trailing residues

are also conserved following the Asp, notably a small P1' residue (Gly/Ser, G/S) and a highly conserved P2' Leu (L) shared in precursors of peptides macrocyclized by asparaginyl endopeptidase (Mylne, et al. 2012). This conservation implied *PawL1* genes might produce small peptide macrocycles. Using a less polar seed extraction method we discovered and sequenced cyclic octapeptides encoded by *PawL1* genes in *Zinnia haageana* (Heliantheae) and *Senecio pinnatifolius* (Senecioneae) (**Fig. 4, Dataset 2**), but did not find any peptide to support *Arnica montana* PawL1. These PawL1-Derived Peptides (PLPs) represent a new biosynthetic route for this type of peptide macrocycle (aka orbitides). Previously described orbitide biosynthetic routes are asparaginyl endopeptidase-independent and encoded by single-product precursors (Arnison, et al. 2013). These data suggest the dual-product *PawL1* genes with their buried, Cys-less macrocycles represent the first stable evolutionary step away from a typical, single-product SESA precursor.

Within a sub-set of the *PawL1*-containing tribes we found *PawS1* genes and peptide evidence for the Cys-containing PDPs, specifically Helenieae, Coreopsideae, Heliantheae, and Millerieae tribes. No evidence of *PawS1* genes was found in the *PawL1*-containing tribes Inuleae or Doroniceae or anywhere else in the Asteroideae (**Fig. 2**). The distribution of *PawS1* genes and their buried PDPs implies that some 4-10 million years (My) passed before the Cys-less PLP macrocycles evolved to Cys-stabilised bicyclic PDPs approximately 41-34 Mya. Fifteen of 17 species containing *PawS1* genes also contain *PawL1* genes, thus the evolution of PDP from PLP was interspersed by *PawL1* gene duplication. Whole genome duplication (WGD) is common in plants (Wendel, et al. 2016) and was hypothesized to have occurred within the Heliantheae ~30-35 Ma (Huang, et al. 2016). This was preceded by an earlier hexaploid WGD event at the origin of the Asteroideae, although this was thought to have become diploidized by the time of the origin of Asteroideae (Panero and Crozier 2016). Either event could have provided the multiple duplications that have led to at least 26 preproalbumin genes in sunflower (21 *SESA*, 3 *PawL1*, 2 *PawS1*), even if only a sub-set is detectable at the mRNA and protein level (Jayasena, et al. 2016).

Although *PawS1* transcripts and PDPs exist in several tribes, transcripts for potent trypsin-inhibitory PDPs like SFTI-1 ("PawS-TI") are confined in distribution to one subtribe, the Helianthinae. Our transcriptomes detected *PawS-TI* transcripts in three species (*Tithonia rotundifolia*, *H. annuus* and *Sanvitalia procumbens*). Furthermore, *PawS-TI* genes have previously been cloned by PCR from *Iostephane heterophylla*, *Aldama phenax* and *Viguiera phenax*, as well as eight different *Helianthus* species (Mylne, et al. 2011; Elliott, et al. 2014). The result of searching 267 Asteraceae seed peptide LC-MS profiles for the masses of trypsin-inhibitory PDPs was consistent with them being confined to

subtribe Helianthinae within the tribe Heliantheae (Elliott, et al. 2014). Also, previous in-gel trypsin inhibitory assays with extracts from a wide range of the Asteraceae (Konarev, et al. 2002) did not detect low-MW trypsin inhibitors outside tribe Heliantheae. Combined, these data suggest that PDPs with a trypsin-inhibitory function evolved in a series of steps. We estimate transition to the lineage-specific functional peptide found throughout members of Helianthinae occurred ~23 Mya, some 20 My after the initial event(s) that created the PLPs (**Fig. 2**).

Discussion

We have elucidated the evolutionary pathway of a trypsin-inhibitory, bicyclic peptide (**Fig. 5**). Three biochemical steps were observable due to the steps being interspersed by gene duplication, which is a major source of new functions and protein products (Conant and Wolfe 2008; Andersson, et al. 2015). The stable biochemical steps observable were represented by three gene types: (i) an insertion event in preproalbumin created two additional asparaginyl endopeptidase target sites (Asn, Asp) and made SESA into a dual product *PawL1* gene type with embedded Cys-less peptide macrocycles; their distribution implies they appeared ~45 Mya, during the Eocene epoch; (ii) *PawS1* whose distribution implies Cys-containing bicyclic PDPs evolved ~34 Mya, during the Oligocene epoch; (iii) the trypsin inhibitory type *PawS1-TI* evolved ~23 Mya, during the Miocene epoch. It is interesting to note that the evolution of *PawS1* is coincident in phylogenetic position with several evolutionary changes that affect the reproductive effort of sunflowers, notably the appearance of large receptacular pales and appearance of phytomelanin in the fruit wall (Panero and Crozier 2016). These properties and PDPs would all appear to provide protection for the developing seed, suggesting *PawS1* and their PDPs might be part of a syndrome associated with divergence of the Heliantheae alliance clade, perhaps allowing success in the relatively dry habitats of western North America/Mexico where this group is so abundant.

What is arguably missing is a gene sequence that represents a link between *PawL1* and *PawS1*; the conservation of a central Pro in both *PawL1* and *PawS1* type genes implies each Cys appeared independently either side of the Pro residue. Sequences with a single Cys either side of the Pro were extremely rare with only two sequences found in *Z. elegans* and *E. peristenia* (**Fig. 2, Dataset 2**). Seed storage albumins with an odd number of Cys residues are likely to be unstable or prone to aggregation and, consistent with this, in the literature and our transcriptomes preproalbumins almost exclusively have even numbers of Cys residues; only five of 121 (4.1%) full length preproalbumins taken at random

from 48 transcriptomes (86 sequences) and GenBank (35 sequences) had an odd number of Cys residues.

The different gene types have been subject to significant gene duplication and, as an example, we observed that within the 51 species with a *PawL1* transcript, some contained up to 13 copies and the average was three copies of *PawL1* (**Fig. 2**). *Zinnia haageana*, *Sanvitalia procumbens* and *Gaillardia pulchella* had twelve, eight and five copies of *PawS1* respectively (**Fig. 2**). These copy numbers were determined from the assembled transcriptome and it is worth noting that without having a genome sequence of these species we cannot exclude that some species might contain recent and near identical gene duplicates. Raw RNA-seq reads for very similar gene copies could be merged into a single transcript during assembly. However, using the same assembly methods, we have previously been able to distinguish highly similar *PawS1* genes differing by several nucleotides over ~450 bp and for which supporting peptide evidence could be found (Jayasena, et al. 2014). Some nearly identical gene copies seen in assembled transcriptomes have been confirmed by PCR cloning using the genomic DNA in this work (**Dataset 3**) and previously (Jayasena, et al. 2014). Also, in this work with the core transcripts used to monitor the quality of each assembly, we could detect differences between the potentially synonymous species *Doronicum orientale* and *D. caucasicum* that shared one identical core transcript, but whose other four core transcripts varied by as few as one or as many as 63 SNPs.

For a new, stable peptide to evolve within an existing ORF differs from other known examples of novel proteins evolving from existing ORFs. A single frameshift was the first documented example (Ohno 1984), but several examples of multi-step, novel protein evolution have been documented. *Umbrea* is a neogene with a key role in chromosome segregation in a number of *Drosophila* species. The steps leading to *Umbrea* required the *Heterochromatin Protein 1B* sequence to integrate within the intron of the *dumpy* gene and its functionality was acquired after a series of genetic events including gene duplication 12-15 Mya, loss of an ancestral domain 10-12 Mya, modification of the chromoshadow domain 5-7 Mya and some species-specific changes (Ross, et al. 2013). The recent characterization of the non-essential overprinted ALTO protein showed its evolution required several stop codons to be erased for it to acquire ALTO-coding capability (Carter, et al. 2013). The genes for antifreeze glycoproteins in Antarctic fish evolved 5-14 Mya by a deletion within a trypsinogen gene and then internal expansion of a peptide sequence to encode tandemly repeating Ala-Thr rich peptides (Chen, et al. 1997). All these examples involve stepwise changes to evolve a radically new protein, but what

differentiates PawS1 is that its new peptide is an addition and did not destroy its ancestral function as a precursor for a seed storage albumin.

In conclusion, we provide evidence for how a novel seed peptide evolved within an existing protein and, by being able to date the stable steps, showed each step was interspersed by 4-10 My. Further research will determine whether this novel example of the evolution of a peptide within an existing protein is a more widespread mechanism.

Materials and Methods

Plant material

Seeds for Asteraceae species (**Table S1**) were sourced directly from the field, herbarium collections, or purchased from commercial seed suppliers. These included Nindethana Seed Company, Mr Fothergill's Seeds & Bulbs and Royston Petrie Seeds (Australia); Chiltern Seeds (UK); Sheffield's Seed Company (USA); B & T World Seeds (France) and Silverhill Seeds (South Africa). Attempts were made to have representative species throughout the Asteraceae, but we were limited to species permitted entry into Western Australia.

RNA extraction and library preparation

RNA was extracted from mature dry Asteraceae seeds using the mini-hot phenol method (Box, et al. 2011). This extraction method is based on a protocol described by Botella *et al.* (1992). Total RNA was DNase treated and further purified using the NucleoSpin RNA Clean-up kit (Macherey-Nagel). RNA was visualized on a 1% agarose gel and the A260/280 and A260/230 ratios were measured using a NanoDrop spectrophotometer (Thermo Scientific) to assess its quality. Sequencing libraries were generated using the TruSeq® Stranded Total RNA LT with Ribo-Zero Plant kit (Illumina) with 300-1000 ng of purified total seed RNA according to the manufacturer's instructions. Sequencing was then performed on an Illumina HiSeq 1500 instrument as 100 bp single read runs or 2 x 100 bp paired-end read runs.

De novo transcriptome assembly

De novo transcriptome assemblies were performed using the second method described in Jayasena *et al.* (2014). Quality trimming and filtering parameters were maintained as Q30, I50 and q22 and p90 respectively. However, a few datasets gave better assemblies when reads were filtered to achieve q30, instead of q22 (**Table S1**). CLC Genomics Workbench (CLC bio) software was used for *de novo* assemblies. If the reads were from a paired-end run, forward and reverse reads were uploaded to CLC

indicating that they were paired-end reads. If the reads were from a single-end run, then trimmed, filtered reads were collapsed prior to uploading to CLC to remove duplicates.

Default CLC *de novo* assembly settings were used, except that each dataset was assembled with at least four different word sizes / *k-mers* (i.e. 23, 30, 40, and 60). For paired reads, assembly parameters were set as auto-detect paired distances: yes, and perform scaffolding: yes. In both single and paired datasets, reads were mapped back to contigs maintaining the mapping parameters as mismatch cost: 2, insertion cost: 3, deletion cost: 3, length fraction: 0.8, similarity fraction: 0.8, and update contigs: yes.

Eight of the 110 RNA samples (i.e. *Coreopsis tinctoria*, *Helianthus annuus-1*, *Ambrosia trifida*, *Lindheimera texana*, *Arnica montana*, *Chromolaena odorata*, *Tetragonotheca texana*, and *Melampodium leucanthum*) were sequenced and assembled *de novo* by the Beijing Genomics Institute using Trinity (**Table S1**). Clean reads provided by BGI were re-assembled with CLC as described above and compared with the Trinity assemblies. The best assembly for each species was selected based on either the number of *PawS1* transcripts identified in an assembly or by assessing *de novo* transcriptome quality statistics (described below).

Searching for *PawS1* genes and *PawS1* intermediates

Assemblies from each species were queried for *PawS1* and *PawS1* intermediates by tBLASTn using the sunflower *PawS1* protein sequence (151-residues). Novel transcripts were examined for the characteristic *PawS1* features such as the presence of an ER signal, GLDN or GLDN-like tail, and the characteristic Cys residue pattern of the accompanying albumin partner before assigning them as *PawS1* or *PawL1*.

WebLogo of SESA, PawL1 and PawS1 predicted proteins

Alignments of SESA (105 sequences), PawL1 (146 sequences), and PawS1 (49 sequences) were made using CLC Genomics Workbench 8.2, maintaining alignment parameters as gap open cost: 5.0, gap extension cost: 5.0, and end gap cost: free. Alignments were uploaded for display with WebLogo 2.8.2 (Schneider and Stephens 1990; Crooks, et al. 2004).

Quality assessment of the *de novo* transcriptomes

To assess the quality of *de novo* transcriptomes as previously described (Jayasena, et al. 2014), the same six core transcripts, namely *LATE EMBRYOGENESIS ABUNDANT (LEA)*, *OLEOSIN (OLE)*, *ASPARTIC PROTEINASE (AP)*, *PATHOGENESIS RELATED (PR)*, *CYSTEINE PROTEASE-1 (CP-1)*, and *SERINE/THREONINE PROTEIN KINASE (PK)* were used. These core transcripts are conserved in all Asteraceae and represent

different mRNA expression levels from high to low. The presence of a core transcript and coverage in an assembly was a reliable measure of transcriptome quality. Core transcripts in each assembly were identified by sequence similarity searches using tBLASTn. Percentage coverage of each core transcript was calculated by dividing the amino acid length of the assembled core transcript by the amino acid length for the reference core transcript. Abundance of core transcripts in each species was assessed by mapping trimmed, filtered raw reads onto the six core transcript sequences identified. Mapping was done using CLC Genomics Workbench 8.0.1, maintaining default mapping parameters except linear gap cost: 0.95, similarity fraction: 0.95, and non-specific match handling: ignore.

Even versus odd numbers of Cys residues in preproalbumins

Eighty six *SESA* transcripts were selected from transcriptomes to represent the full breadth of the Asteraceae (2 transcripts from Mutisioideae, 4 from Carduoideae, 12 from Cichorioideae, 12 from Senecioneae, 6 from Calenduleae, 1 from Doroniceae, 2 from Gnaphalieae, 1 from Astereae, 8 from Anthemideae, 4 from Helenieae, 2 from Inuleae, 4 from Coreopsideae, 21 from Heliantheae, 1 from Madieae, 3 from Eupatorieae and 3 from Millerieae). Non-Asteraceae *SESA* sequences were downloaded from the GenBank protein depository to include at least one 2S albumin from each species listed. In the five SESAs with an odd number of Cys residues, the 'extra' Cys was always in the small albumin subunit.

Asteraceae ITS and chloroplast sequences

ITS sequences for 33 species (*Anacyclus depressus*, *Arctotis gumbletonii*, *Arctotis scapigera*, *Carthamnus tinctorius*, *Centaurea cyanooides*, *Centaurea pseudophrygia*, *Gazania splendens*, *Gazania x hybrida*, *Guizotia abyssinica*, *Helichrysum bracteatum*, *Helichrysum macranthum*, *Hymenopappus tenuifolius*, *Inula ensifolia*, *Ligularia przewalskii*, *Olearia lyrata*, *Oncosiphon africanus*, *Osteospermum ecklonis*, *Osteospermum jucundum*, *Osteospermum sinuatum*, *Othonna arborescens*, *Parthenium argentatum*, *Sanvitalia procumbens*, *Schoenia filifolia*, *Senecio magnificus*, *Senecio pinnatifolius* ssp. *latilobus*, *Senecio pinnatifolius* var. *maritimus*, *Senecio pinnatifolius* var. *pinnatifolius*, *Streptoglossa macrocephala*, *Tagetes tenuifolia*, *Tanacetum coccineum*, *Tetragonotheca texana*, *Waitzia nitida*, *Xerochrysum viscosum*) were obtained by PCR from genomic DNA template with the primers ITS7.5 (5'-GAG TCA TCA GCT CGC GTT GAC TA-3'), ITS4 (5'-TCC TCC GCT TAT TGA TAT GC-3') and ITS5 (5'-GGA AGT AAA AGT CGT AAC AAG G-3'). The ITS, *rbcl* and *ndhF* sequences for *Zinnia haageana* were acquired from its transcriptome. For a list of GenBank codes used to generate **Fig. 2** see **Dataset 8**.

Asteraceae chronogram

Sequences were edited using the program Sequencher v 4.9 (Gene Codes Corporation, Ann Arbor, USA) and assembled into primer-based matrices. The ITS and chloroplast gene matrices were exported as Nexus interleaved files and concatenated in Paup* (Swofford 2003). For the alignment used to generate the phylogeny in FASTA format see **Dataset 7**. The matrices were exported into Mesquite (Maddison and Maddison 2015) to produce Phylip and nbrf files. Alignment of the nbrf files was performed at the European Bioinformatics Institute web site (<http://www.ebi.ac.uk/Tools/msa/mafft>) using the program MAFFT (Katoh and Standley 2013). The aligned data matrices were imported into Mesquite, edited and subsequently exported as simplified Nexus interleaved files and concatenated in Paup*. The concatenated file was exported into Mesquite and exported as a simplified Nexus sequential file to use in divergence time analyses. The concatenated matrix contained a total of 5929 sites. For the ITS and chloroplast matrices the model of molecular evolution was evaluated under the Akaike information criterion using the program jModeltest 2 (Guindon and Gascuel 2003; Darriba, et al. 2012). The GTR + I + G had the highest likelihood of all models for the ITS matrix and GTR + G for the chloroplast matrix. The GTR + I + G model was used in the divergence time analysis.

Four absolute time constraints were chosen that could be assigned with confidence to specific nodes of the phylogeny: (i) We constrained the root of the tree using the fossil pollen grain *Mutisiapollis telleriae* co-fossilized with the macrofossil *Raiguenrayun cura* and dated to 47.5 Mya (middle Eocene Rio *Pichileufu paleoflora* from central Patagonia; (Barreda, et al. 2012)). We modeled an exponential prior distribution with an offset of 47.5 My and a mean of 2.2 My to constrain the minimum age of the Mutisioideae-Asteroideae clade as specified by Panero *et al.* (2014). The mean value was chosen to be contained within the 95% confidence interval of the distribution assuming a soft maximum constraint of 54 My that corresponds to the age of a pollen grain of *Tubulifloridites* with well-developed spines as in *Mutisiapollis telleriae* collected in Antarctica and magneto- and bio-stratigraphically dated to approximately 53 Mya (Pross, et al. 2012; Contreras, et al. 2013); (ii) *Centaurea* type fossil pollen from the Pliocene (Popescu 2002) was used to model a prior using a lognormal distribution with an offset of 4.6 Mya and mean and standard deviation of 0.5 My to constrain the *Centaurea-Carthamus* (Carduoideae) node; (iii) The oldest *Artemisia* fossil pollen has been recorded from the Eocene-Oligocene boundary of northwestern China (Wang 2004; Hobbs and Baldwin 2013). We modeled this prior using a lognormal distribution with an offset of 31 Mya and a mean (0.2 My) and standard deviation (0.5 My) to constrain the node *Artemisia-Chrysanthemum*; (iv) The high elevation Páramo habitat of northern South America has resulted from orogeny that has been dated to the last 2-4 Mya

(Hughes and Eastwood 2006; Sklenář, et al. 2011). The *Espeletia* clade of the tribe Millerieae is endemic to the Páramo habitat and adjacent high elevation Andean forests. We modeled a lognormal prior on the root node of the *Espeletia* clade (*Carramboa*, *Espeletia*, *Libanothamnus*) with an offset of 3 My, a standard deviation of 0.4 My and a mean of 0.2 My.

BEAST v. 1.8.0 (Drummond, et al. 2012) was used to infer the phylogeny and divergence times for taxa in Asteraceae. The analysis was performed at the CIPRES Science Gateway (<http://www.phylo.org/>, (Miller, et al. 2015)) assuming a Bayesian uncorrelated lognormal clock of evolutionary rates. A general time reversible model with gamma-distributed substitution rates among sites (Lanave, et al. 1984; Tavaré 1986) allowing four single-nucleotide substitution categories was applied to the concatenated data matrix. The Markov chain Monte Carlo algorithm was run for 120 million posterior iterations sampling every 20,000 posterior iterations. Tracer v. 1.6 (Rambaut, et al. 2014) was used to assess effective sample sizes with values above 200 considered for optimal convergence and tree likelihood stationarity. After discarding the initial 10% of samples (burn-in), a maximum clade credibility tree was constructed in TreeAnnotator v. 1.8.0 (Drummond, et al. 2012) depicting the maximum sum of posterior clade probabilities. The maximum clade credibility was visualized in FigTree v. 1.4.2 (<http://tree.bio.ed.ac.uk/software/figtree/>) including mean ages and a time scale and bars for the 95% highest posterior densities for all branches with > 0.95 posterior probabilities.

Relationships among the five subfamilies of Asteraceae included in our study are completely congruent with recent molecular phylogenies of the family, whether inferred from multiple chloroplast markers (Panero and Crozier 2016) or from more than 60 low copy nuclear genes (Huang, et al. 2016). Only three tribal relationships within Asteroideae differ between this study and topologies recovered in those more comprehensive phylogenetic studies. The position of Coreopsideae here is incongruent with previous molecular studies that have sampled the tribe (Baldwin, et al. 2002; Liu, et al. 2015; Panero and Crozier 2016), but these all differ in their placements as well. It has been hypothesized that this is a tribe of hybrid origin (Liu, et al. 2015). The position of Heliantheae in our analysis is incongruent with that in Panero & Crozier (2016) and Baldwin *et al.* (2002). Those earlier studies are similarly incongruent with each other, but the tribe's resolution is only weakly supported here and in Baldwin *et al.* (2002). Lastly, we observed the position here of tribe Bahieae is incongruent in its placement in Panero & Crozier (2016) and with Baldwin *et al.* (2002), the earlier two are also incongruent (Bahieae was only weakly supported in the latter). We don't know if these incongruences are due to variously incomplete sampling, to hybrid origin of some of the species sampled, or different nuclear and chloroplast

evolutionary histories. Luckily, alternate placement in any of these cases would not affect our conclusions about the order of events in this instance of peptide evolution. Ages of major backbone divergences and internal nodes exclusive of the three tribes mentioned are similar with those reported in other recent studies of Asteraceae, with those reported by Huang *et al.* (2016), slightly older than those of Panero & Crozier (2016), but all within the 95% high posterior density of the ages reported here.

Cloning *PawS1* genes and *PawS1* intermediates

To confirm the putative *PawS1* and *PawS1* intermediate transcripts identified from *de novo* transcriptomes were not assembly artefacts, 39 genes were cloned either from seed genomic DNA or cDNA. For this purpose, genomic DNA was extracted from mature, dry Asteraceae seeds using the DNeasy plant mini kit (Qiagen). RACE-ready cDNA was synthesized with 1 µg of the same total RNA used for RNA-seq using the SMARTer RACE cDNA amplification kit (Clontech). Primers used for 3' and 5' RACE and full length gene cloning are listed in **Table S2**.

If the *PawS1/PawS1* intermediate transcript identified from the transcriptome appeared to be full length, forward and reverse primers were designed outside the ORF. Full length genes were PCR amplified using either genomic DNA or cDNA as the template with *Taq* DNA polymerase, using specific primers for each gene (**Table S2**). Amplified fragments were cloned into the pGEM-Teasy vector (Promega) and three independent clones for each transcript were sequenced to avoid errors introduced by *Taq* DNA polymerase.

If the transcript was partial, either 3' or 5' or both 3' and 5' RACE were performed as necessary using specific primers (**Table S2**) with respective RACE-ready cDNA. The amplified fragments were cloned into the pGEM-T Easy vector (Promega) and sequenced to identify the full length gene sequence. The full length genes were then PCR amplified using specific primers and cloned as described above.

Peptide extractions

To extract peptides, dry mature Asteraceae seeds were ground to a fine powder under liquid nitrogen with a pinch of glass beads. A volume of approximately 0.5 mL tissue powder was mixed with 0.4 mL methanol, 0.4 mL dichloromethane, and 0.1 mL 0.05% trifluoroacetic acid in water. After mixing by vortex, the mixture was centrifuged for 3 min at 13,000 rpm. If the phases were not separated at this point, 0.1 mL chloroform or 0.1 mL 0.05% trifluoroacetic acid in water was added alternately, followed

by a short centrifugation after each addition, until phase separation was achieved. The upper (aqueous) layer containing peptides was collected and dried by centrifugation under vacuum.

An alternative seed extraction method was used to search for PLPs in *Z. haageana* and *S. pinnatifolius*. Approximately 0.5 g of dry mature seeds was crushed in 50 mL of 50% methanol 50% dichloromethane (v/v). The crude extract was filtered under a vacuum and the filtrate collected. The crushed seeds were washed twice with 50 mL of 50% methanol 50% dichloromethane. These filtrates were collected, combined with the previous filtrate and dried with excess anhydrous magnesium sulfate. Polar molecules were removed by flash chromatography through a 10 mm silica gel column (60 µm average particle size). The eluate was collected, dried in a rotary evaporator at 40 °C and stored at 4 °C.

LC-MS and LC-MS/MS

Lyophilized crude peptide was dissolved in HPLC-grade 5% acetonitrile (v/v) and 0.1% formic acid (v/v) (Honeywell) prior to loading onto a C18 high capacity nano LC chip trapping column (160 nL) (Agilent Technology) in 95% solvent A (0.1% formic acid in water) and 5% solvent B (0.1% formic acid in acetonitrile) using an 1100 series or 1200 series capillary pump (Agilent Technologies). Then the trapping column was switched in-line with an 1100 series or 1200 series nano pump and a C18 analytical column (75 µm x 150 mm). A gradient of 5% solvent B to 60% solvent B or 95% solvent B in 15 min was used to elute peptides into a 6510 series QTOF mass spectrometer (Agilent Technologies). The QTOF was operated in MS-mode only and the MS data were collected at eight spectra per second.

In cases where a mass corresponding to a predicted PLP was found by LC-MS, tandem LC-MS (LC-MS/MS) experiments were carried out, fragmenting the putative peptide using collision-induced dissociation at voltages between 31 V and 36 V. The resulting MS/MS spectra were manually analyzed to identify any peptide fragment ions produced, and sequenced *de novo* by assignment of amino acid residues to b-ions in the spectra when they were within 0.015 Da of the residue mass.

tBLASTN search for PLP sequences

As further confirmation that each PLP was encoded by a *PawL1* gene as predicted, a tBLASTn search was carried out against all six ORFs of the assembled transcriptome of the corresponding species using CLC Genomics Workbench. The search queries included all possible linear amino acid residue sequences of each cyclic peptide allowing for ambiguity between Ile and Leu, a total of 96 queries for *Senecio pinnatifolius* and 768 queries for *Zinnia haageana*.

Acknowledgements

This work was supported by an Australian Research Council (ARC) Discovery Project (DP130101191 to J.S.M., J.L.P. and E.E.S.), an ARC Super Science Fellowship (FS100100022 to D.S.), ARC Future Fellowships (FT130100123 to N.L.T. and FT120100013 to J.S.M.), an ARC Centre of Excellence grant (CE140100008 to J.W.). K.B.-L. was supported by the ARC (DP120103369). A.S.J. was supported by an Australian Postgraduate Award and an International Postgraduate Research Scholarship. The authors thank H. Walker for IT support and K. Stubbs for assistance with seed peptide extractions. This work was also supported by an Australian Postgraduate Award (A.S.J.) and an Australian Postgraduate Award and a Bruce and Betty Green Postgraduate Research Scholarship (M.F.F.) ITS sequences generated for phylogenetic tree construction are KY397478- KY397511. *Zinnia haageana* ITS, rbcL and ndhF sequences are KY405009-KY405011. The 44 cloned Asteraceae *PawL1* and *PawS1* sequences are deposited in GenBank under accessions KY078350-KY078393. All new RNA-seq data used for *de novo* transcriptomes were deposited in the National Centre for Biotechnology Information under accession number SRP099044.

Author contribution statement

J.S.M. conceived the study and performed RNA extractions; A.S.J. performed all bioinformatics and cloned the genes, A.S.J. and M.F.F. performed seed peptide extractions and analyzed LC-MS data; J.L.P. and E.E.S. provided seeds and constructed the Asteraceae chronogram; D.S., O.B., and J.W. prepared libraries and performed sequencing; K.B.-L. and M.F.F. ran LC-MS; N.L.T. and M.F.F. performed LC-MS/MS; all authors analyzed the data; A.S.J., J.S.M. and M.F.F. wrote the manuscript with contributions from all authors.

References

- Andersson DI, Jerlström-Hultqvist J, Näsvall J. 2015. Evolution of New Functions De Novo and from Preexisting Genes. *Cold Spring Harbor Perspectives in Biology* 7.
- Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, Camarero JA, Campopiano DJ, Challis GL, Clardy J, et al. 2013. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Natural Product Reports* 30:108-160.
- Baldwin BG, Wessa BL, Panero JL. 2002. Nuclear rDNA evidence for major lineages of helenioid Heliantheae (Compositae). *Systematic Botany* 27:161-198.

Barreda VD, Palazzesi L, Katinas L, Crisci JV, Tellería MC, Bremer K, Passala MG, Bechis F, Corsolini R. 2012. An extinct Eocene taxon of the daisy family (Asteraceae): evolutionary, ecological and biogeographical implications. *Annals of Botany* 109:127-134.

Begun DJ, Lindfors HA, Kern AD, Jones CD. 2007. Evidence for de novo evolution of testis-expressed genes in the *Drosophila yakuba*/*Drosophila erecta* clade. *Genetics* 176:1131-1137.

Bornberg-Bauer E, Huylmans AK, Sikosek T. 2010. How do new proteins arise? *Current Opinion in Structural Biology* 20:390-396.

Botella JR, Arteca JM, Schlagnhauser CD, Arteca RN, Phillips AT. 1992. Identification and characterization of a full-length cDNA encoding for an auxin-induced 1-aminocyclopropane-1-carboxylate synthase from etiolated mung bean hypocotyl segments and expression of its mRNA in response to indole-3-acetic acid. *Plant Molecular Biology* 20:425-436.

Box MS, Coustham V, Dean C, Mylne JS. 2011. Protocol: A simple phenol-based method for 96-well extraction of high quality RNA from *Arabidopsis*. *Plant Methods* 7:7.

Carter JJ, Daugherty MD, Qi X, Bheda-Malge A, Wipf GC, Robinson K, Roman A, Malik HS, Galloway DA. 2013. Identification of an overprinting gene in Merkel cell polyomavirus provides evolutionary insight into the birth of viral genes. *Proceedings of the National Academy of Sciences of the United States of America* 110:12744-12749.

Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charlotiaux B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and *de novo* gene birth. *Nature* 487:370-374.

Chen L, DeVries AL, Cheng C-HC. 1997. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proceedings of the National Academy of Sciences of the United States of America* 94:3811-3816.

Chen S, Zhang YE, Long M. 2010. New genes in *Drosophila* quickly become essential. *Science* 330:1682-1685.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics* 9:938-950.

Contreras L, Pross J, Bijl PK, Koutsodendris A, Raine JJ, van de Schootbrugge B, Brinkhuis H. 2013. Early to Middle Eocene vegetation dynamics at the Wilkes Land Margin (Antarctica). *Review of Palaeobotany and Palynology* 197:119-142.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: A sequence logo generator. *Genome Research* 14:1188-1190.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* 9:772-772.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29:1969-1973.

Dunse KM, Stevens JA, Lay FT, Gaspar YM, Heath RL, Anderson MA. 2010. Coexpression of potato type I and II proteinase inhibitors gives cotton plants protection against insect damage in the field. *Proceedings of the National Academy of Sciences of the United States of America* 107:15011-15015.

Elliott AG, Delay C, Liu H, Phua Z, Rosengren KJ, Benfield AH, Panero JL, Colgrave ML, Jayasena AS, Dunse KM, et al. 2014. Evolutionary origins of a bioactive peptide buried within preproalbumin. *Plant Cell* 26:981-995.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52:696-704.

Hartl M, Giri AP, Kaur H, Baldwin IT. 2010. Serine protease inhibitors specifically defend *Solanum nigrum* against generalist herbivores but do not influence plant growth and development. *Plant Cell* 22:4158-4175.

Heinen TJAJ, Staubach F, Häming D, Tautz D. 2009. Emergence of a new gene from an intergenic region. *Current Biology* 19:1527-1531.

Hobbs CR, Baldwin BG. 2013. Asian origin and upslope migration of Hawaiian *Artemisia* (Compositae–Anthemideae). *Journal of Biogeography* 40:442-454.

Huang C-H, Zhang C, Liu M, Hu Y, Gao T, Qi J, Ma H. 2016. Multiple Polyploidization Events across Asteraceae with Two Nested Events in the Early History Revealed by Nuclear Phylogenomics. *Molecular Biology and Evolution* 33:2820-2835.

Hughes C, Eastwood R. 2006. Island radiation on a continental scale: Exceptional rates of plant diversification after uplift of the Andes. *Proceedings of the National Academy of Sciences of the United States of America* 103:10334-10339.

Jayasena AS, Franke B, Rosengren KJ, Mylne JS. 2016. A tripartite 'omic analysis identifies the major sunflower seed albumins. *Theoretical and Applied Genetics* 129:613-629.

Jayasena AS, Secco D, Bernath-Levin K, Berkowitz O, Whelan J, Mylne JS. 2014. Next generation sequencing and *de novo* transcriptomics to study gene evolution. *Plant Methods* 10:34.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30:772-780.

Konarev AV, Anisimova IN, Gavrilova VA, Vachrusheva TE, Konechnaya GY, Lewis M, Shewry PR. 2002. Serine proteinase inhibitors in the Compositae: distribution, polymorphism and properties. *Phytochemistry* 59:279-291.

Krebbes E, Herdies L, De Clercq A, Seurinck J, Leemans J, Van Damme J, Segura M, Gheysen G, Van Montagu M, Vandekerckhove J. 1988. Determination of the processing sites of an *Arabidopsis* 2S albumin and characterization of the complete gene family. *Plant Physiology* 87:859-866.

Lanave C, Preparata G, Saccone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution* 20:86-93.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences of the United States of America* 103:9935-9939.

Liu M, Zhang CF, Huang CH, Ma H. 2015. Phylogenetic reconstruction of tribal relationships in Asteroideae (Asteraceae) with low-copy nuclear genes. *Chinese Bulletin of Botany* 50.

Luckett S, Garcia RS, Barker JJ, Konarev AV, Shewry PR, Clarke AR, Brady RL. (1992-1999 co-authors). 1999. High-resolution structure of a potent, cyclic proteinase inhibitor from sunflower seeds. *Journal of Molecular Biology* 290:525-533.

Mesquite: a modular system for evolutionary analysis [Internet]. 2015. Available from: <http://mesquiteproject.org>

McLysaght A, Guerzoni D. 2015. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.

McLysaght A, Hurst LD. 2016. Open questions in the study of *de novo* genes: what, how and why. *Nature Reviews Genetics* 17:567-578.

Miller MA, Schwartz T, Pickett BE, He S, Klem EB, Scheuermann RH, Passarotti M, Kaufman S, O'Leary MA. 2015. A RESTful API for access to phylogenetic tools via the CIPRES science gateway. *Evolutionary Bioinformatics Online* 11:43-48.

Moreno FJ, Jenkins JA, Mellon FA, Rigby NM, Robertson JA, Wellner N, Clare Mills EN. 2004. Mass spectrometry and structural characterization of 2S albumin isoforms from Brazil nuts (*Bertholletia excelsa*). *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1698:175-186.

Murphy DN, McLysaght A. 2012. De Novo Origin of Protein-Coding Genes in Murine Rodents. *PLoS ONE* 7:e48650.

Mylne JS, Chan LY, Chanson AH, Daly NL, Schaefer H, Bailey TL, Nguyencong P, Cascales L, Craik DJ. 2012. Cyclic peptides arising by evolutionary parallelism via asparaginyl-endoribonuclease-mediated biosynthesis. *Plant Cell* 24:2765-2778.

Mylne JS, Colgrave ML, Daly NL, Chanson AH, Elliott AG, McCallum EJ, Jones A, Craik DJ. 2011. Albumins and their processing machinery are hijacked for cyclic peptides in sunflower. *Nature Chemical Biology* 7:257-259.

Mylne JS, Hara-Nishimura I, Rosengren KJ. 2014. Seed storage albumins: biosynthesis, trafficking and structures. *Functional Plant Biology* 41:671-677.

Ohno S. 1984. Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. *Proceedings of the National Academy of Sciences of the United States of America* 81:2421-2425.

Panero JL, Crozier BS. 2016. Macroevolutionary dynamics in the early diversification of Asteraceae. *Molecular Phylogenetics and Evolution* 99:116-132.

Panero JL, Freire SE, Ariza Espinar L, Crozier BS, Barboza GE, Cantero JJ. 2014. Resolution of deep nodes yields an improved backbone phylogeny and a new basal lineage to study early evolution of Asteraceae. *Molecular Phylogenetics and Evolution* 80:43-53.

Pavesi A, Magiorkinis G, Karlin DG. 2013. Viral proteins originated *de novo* by overprinting can be identified by codon usage: application to the “gene nursery” of *Deltaretroviruses*. *PLoS Computational Biology* 9:e1003162.

Popescu S-M. 2002. Repetitive changes in Early Pliocene vegetation revealed by high-resolution pollen analysis: revised cyclostratigraphy of southwestern Romania. *Review of Palaeobotany and Palynology* 120:181-202.

Pross J, Contreras L, Bijl PK, Greenwood DR, Bohaty SM, Schouten S, Bendle JA, Rohl U, Tauxe L, Raine JJ, et al. 2012. Persistent near-tropical warmth on the Antarctic continent during the early Eocene epoch. *Nature* 488:73-77.

Tracer v1.6, available from <http://beast.bio.ed.ac.uk/Tracer> [Internet]. 2014. Available from: <http://beast.bio.ed.ac.uk/Tracer>

Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. 2013. *De novo* ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genetics* 9:e1003860.

Rico M, Bruix M, González C, Monsalve RI, Rodríguez R. 1996. ¹H NMR assignment and global fold of napin BnIb, a representative 2S albumin seed protein. *Biochemistry* 35:15672-15682.

Ross BD, Rosin L, Thomae AW, Hiatt MA, Vermaak D, de la Cruz AFA, Imhof A, Mellone BG, Malik HS. 2013. Stepwise evolution of essential centromere function in a *Drosophila* neogene. *Science* 340:1211-1214.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* 18:6097-6100.

Scofield SR, Crouch ML. 1987. Nucleotide sequence of a member of the napin storage protein family from *Brassica napus*. *Journal of Biological Chemistry* 262:12202-12208.

Shewry P. 1999. Enzyme Inhibitors of Seeds: Types and Properties. In: Shewry P, Casey R, editors. *Seed Proteins*. Dordrecht: Kluwer. p. 587-615.

Sklenář P, Dušková E, Balslev H. 2011. Tropical and Temperate: Evolutionary History of Páramo Flora. *The Botanical Review* 77:71-108.

PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods) [Internet]. Sinauer Associates.; 2003.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nature Reviews Genetics* 12:692-702.

Tavaré S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures in Mathematics in the Life Sciences* 17:57-86.

Tyndall JDA, Nall T, Fairlie DP. 2005. Proteases universally recognize beta strands in their active sites. *Chemical Reviews* 105:973-1000.

van der Klei H, Damme JV, Casteels P, Krebbers E. 1993. A fifth 2S albumin isoform is present in *Arabidopsis thaliana*. *Plant Physiology* 101:1415-1416.

Wang W-M. 2004. On the origin and development of *Artemisia* (Asteraceae) in the geological past. *Botanical Journal of the Linnean Society* 145:331-336.

Wendel JF, Jackson SA, Meyers BC, Wing RA. (Wendel2016 co-authors). 2016. Evolution of plant genome architecture. *Genome Biology* 17:37.

Wu D-D, Irwin DM, Zhang Y-P. 2011. *De novo* origin of human protein-coding genes. *PLoS Genetics* 7:e1002379.

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of *de novo* genes in *Drosophila melanogaster* populations. *Science* 343:769-772.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in *Drosophila*. *Genome Research* 18:1446-1455.

a

b

Figure 1 | Summary of sunflower PawS1 and its buried protease inhibitor SFTI-1

(A) Sunflower PawS1 is a 151 amino acid protein that is matured by the Asn- and Asp- recognizing asparaginyl endopeptidase (AEP). AEP cleavages at Asn53 and Asn84 release the albumin. Cleavage at Asn35 releases an 18 amino acid peptide that is macrocyclised by a specialist AEP (AEP*) during cleavage of its GLDN tail to form SFTI-1. **(B)** Disulfide bonds connect the small (green) and large (orange) albumin subunits as a heterodimer of linear peptides. During maturation regions of PawS1 are degraded (black) such as the region between small and large subunits in albumin which is often trimmed by an exoprotease. The ER signal (rose) is typically trimmed upon entry to the secretory pathway. The ball-format structure of SFTI-1 displayed is from PDB ID 1SFI, whereas the mature albumin shown to scale with SFTI-1 is not sunflower PawS1 albumin. It is a model of the prototypic, napin-type albumin from *Brassica napus* (PBD 1PNB) (Rico, et al. 1996). Mature, napin-type albumins have a conserved disulfide bond network and five-helix structure (Mylne, et al. 2014).

(See next page)

Figure 2 | Chronogram with distribution of *PawS1*, *PawL1* and ancestral *SESA* transcripts

Presence and absence of *SESA*, *PawL*, *PawS* and *PawS-TI* genes are marked and colored as follows: species not sequenced (white), sequenced and no transcript found (gray), transcript(s) found (dark blue), found and at least one transcript has been confirmed by cloning (light blue). The numbers within boxes indicate the number of transcripts identified. Numbers adjacent to nodes indicate estimated divergence times in Mya, boxed numbers on the phylogram indicate calibration points (see Materials and Methods). The colored shades indicate subfamilies. The *PawL* gene was found throughout the subfamily Asteroideae, whose tribes are delineated with labels on the right. Darker shades of blue within Asteroideae mark the distribution of *PawS1* and *PawS1-TI* genes. Arrows on branch points indicate the key transitions of *PawS* evolution. Confidence intervals are shown as green bars on nodes with $\geq 95\%$ support.

Figure 3 | Summary of the predicted SESA, PawL1, and PawS1 sequences from *de novo* transcriptomes

Weblogo illustrating conserved residues in SESA (105 sequences), PawL1 (146 sequences), and PawS1 (49 sequences). Pale blue bars delineate the sequence of the peptide itself. Inset in the top right hand corner summarizes in which subfamilies and tribes the key transitions exist. For a full list of sequences used in the alignment see **Dataset 4-6**.

Figure 4 | Evidence for PawL-Derived Peptides

Liquid chromatography-mass spectrometry evidence for novel peptide macrocycles from seeds of two Asteraceae species (*Zinnia haageana* in blue; *Senecio pinnatifolius* in green). (A): Total ion chromatogram with the retention time of the peptide marked (B): Extracted ion chromatogram showing predicted cyclic peptide mass and mass spectrum (inset) (C): Tandem mass spectrum fragments, which correspond to the predicted residues. Note: leucine (L) and isoleucine (I) residues cannot be distinguished by mass alone: other techniques are required to confirm them.

Figure 5 | Birth and stepwise evolution of SFTI-1 within a prealbumin gene

We propose a genetic expansion event in a typical albumin precursor protein upstream of the region coding for the small albumin subunit led to the stable *PawL1* gene type that produced albumin and an additional small macrocycle (~47-45 Mya). Contemporary with genome duplication in the lineage, two Cys residues evolved such that the peptide was bicyclic (~41-34 Mya). Fully functional bicyclic trypsin inhibitors evolved ~24-22 Mya based on our taxon sample.

Supplementary Information

Stepwise evolution of a buried inhibitor peptide over 45 million years

Achala S. Jayasena, Mark F. Fisher, Jose L. Panero, David Secco, Kalia Bernath-Levin, Oliver Berkowitz,
Nicolas L. Taylor, Edward E. Schilling, James Whelan, and Joshua S. Mylne

Supplementary Tables

Supplementary Table 1: *De novo* transcriptome assembly and quality statistics

Species are listed by classification in the same order as **Figure 2**. Sub-family is abbreviated (Ast., Asteroideae; Cic., Cichorioideae; Car., Carduoideae; Mut., Mutisiodeae). Quality statistics are shown for only one among four assemblies. 'ws' is the word size/*k-mer*. The first line for each species shows the percentage coverage and the second line the average coverage (the sum of the bases of the aligned part of all the reads divided by the length of the reference sequence). The species pairs *Doronicum caucasicum*/*D. orientale* and *Ageratum mexicanum*/*A. houstonianum* are potential synonyms. We used core transcripts from one of each pair as a query to BLASTn the other transcriptome and found sufficient SNPs to keep them listed as separate species for the purpose of this work. Only one core transcript was identical in the *D. orientale* and *D. caucasicum* transcriptomes. The number of SNPs in the other four core transcripts varied between 1 and 63. Five identical *A. houstonianum* core transcripts were identified in the *A. mexicanum* transcriptome; however the sixth core transcript varied by two SNPs.

Species	Sub-F., Tribe	Raw reads	Clean reads	Run	Assembler	N50	Contigs	LEA	OLE	AP	PR	CP-1	PK
<i>Guizotia abyssinica</i>	Ast., Millereae	92,632,632	84,292,029	Paired	CLC ws23	209	404,438	63% 4,823	86% 12,320	99% 435	100% 13,971	80% 66	75% 23
<i>Galinsoga parviflora</i>	Ast., Millereae	45,125,599	44,125,863	Single	CLC ws60	592	58,907	100% 18,843	99% 13,042	100% 1,515	100% 24,802	59% 31	85% 22
<i>Tetragonotheca texana</i>	Ast., Millereae	29,779,210	25,713,114	Paired	CLC ws60	469	36,979	100% 31,032	100% 19,735	100% 491	100% 29,514	100% 107	100% 58
<i>Melampodium paludosum</i>	Ast., Millereae	46,789,706	30,202,652	Single	CLC ws60	431	78,633	100% 11,506	100% 17,840	100% 412	100% 140	99% 828	100% 71
<i>Melampodium leucanthum</i>	Ast., Millereae	33,267,168	22,094,416	Paired	Trinity	151	216,779	98% 161	48% 10	7% 2	35% 4	19% 13	19% 34
<i>Arnica montana</i>	Ast., Madieae	28,916,086	27,516,042	Paired	Trinity	444	137,471	100% 88,119	83% 21,902	100% 712	100% 4,529	100% 66	100% 39
<i>Arnica chamissonis</i>	Ast., Madieae	25,809,950	24,999,482	Single	CLC ws60	383	71,637	89% 7,789	93% 8,972	92% 47	100% 187	64% 1,816	66% 31
<i>Hymenopappus tenuifolius</i>	Ast., Bahieae	44,461,228	39,454,124	Single	CLC ws40	517	64,929	100% 126	79% 37	100% 168	100% 7,087	89% 249	73% 26
<i>Liatris spicata</i>	Ast., Eupatorieae	155,822,114	134,696,761	Paired	CLC ws60	402	94,536	99% 40	100% 13,552	92% 132	75% 2,845	99% 232	69% 34
<i>Ageratum houstonianum</i>	Ast., Eupatorieae	65,972,650	62,844,045	Single	CLC ws60	380	126,789	72% 11,075	100% 17,455	43% 1,908	100% 33	75% 48	100% 67
<i>Ageratum mexicanum</i>	Ast., Eupatorieae	60,312,841	59,941,689	Single	CLC ws30	479	121,250	99%	80%	60%	100%	61%	98%

								12,082	16	1,467	19	9	23
<i>Chromolaena odorata</i>	Ast., Eupatorieae	28,966,774	26,027,156	Paired	Trinity	262	138,505	100%	100%	100%	100%	100%	69%
								112	6,883	347	1,967	275	29
<i>Tithonia rotundifolia</i>	Ast., Heliantheae	234,389,862	206,194,086	Paired	CLC ws30	607	71,753	99%	99%	80%	100%	100%	100%
								40,522	99	683	16	6,783	28
<i>Helianthus annuus</i>	Ast., Heliantheae	42,809,404	40,742,686	Paired	Trinity	425	161,498	100%	100%	100%	100%	100%	100%
								187,373	16,382	1,422	5,253	93	138
<i>Helianthus annuus</i>	Ast., Heliantheae	45,416,165	42,183,557	Single	CLC ws60	454	25,925	90%	69%	72%	56%	63%	73%
								3,031	702	342	3	14	25
<i>Ambrosia trifida</i>	Ast., Heliantheae	27,209,442	24,308,178	Paired	CLC ws23	324	141,073	99%	100%	88%	100%	100%	85%
								3,441	38,149	535	465	361	57
<i>Parthenium argentatum</i>	Ast., Heliantheae	237,665,922	196,800,414	Paired	CLC ws60	487	100,874	68%	100%	99%	100%	100%	59%
								306,091	29,616	4,054	30	110	30
<i>Parthenium integrifolium</i>	Ast., Heliantheae	242,049,214	220,863,239	Paired	CLC ws60	365	379,536	97%	89%	98%	100%	83%	100%
								21,861	19,973	2,137	7,114	104	93
<i>Sanvitalia procumbens</i>	Ast., Heliantheae	230,909,080	205,833,064	Paired	CLC ws60	427	109,350	100%	91%	54%	100%	77%	100%
								23,977	56	149	57	54	94
<i>Echinacea purpurea</i>	Ast., Heliantheae	231,413,396	207,641,295	Paired	CLC ws40	434	210,919	100%	81%	100%	100%	100%	50%
								6,262	50	4,081	32	75	73
<i>Zinnia haageana</i>	Ast., Heliantheae	76,764,180	64,649,107	Paired	CLC ws60	435	80,460	79%	95%	89%	100%	98%	97%
								5,584	12,436	43	13	66	40
<i>Zinnia elegans</i>	Ast., Heliantheae	52,624,683	42,020,906	Single	CLC ws60	846	24,956	74%	80%	100%	100%	40%	80%
								1,415	611	47	45	10	20
<i>Heliopsis helianthoides</i>	Ast., Heliantheae	219,255,188	169,128,716	Paired	CLC ws60	497	151,272	73%	45%	97%	100%	94%	99%
								77,866	123,416	5,303	51	291	69
<i>Ratibida columnifera</i>	Ast., Heliantheae	137,764,288	125,996,190	Paired	CLC ws60	302	163,253	66%	84%	96%	100%	62%	66%
								50,326	17	1,257	133	59	73
<i>Rudbeckia hirta</i>	Ast., Heliantheae	94,730,298	87,062,280	Paired	CLC ws60	348	131,098	94%	77%	79%	100%	80%	100%
								101,042	25,724	6,160	684	22	22
<i>Engelmannia peristenia</i>	Ast., Heliantheae	87,568,879	83,655,125	Single	CLC ws40	455	62,642	100%	100%	100%	100%	100%	92%
								4	419	158	91	109	83
<i>Lindheimera texana</i>	Ast., Heliantheae	31,193,094	26,536,102	Paired	CLC ws60	439	36,378	100%	100%	100%	100%	99%	56%
								101,482	38,635	823	1,089	292	30
<i>Tagetes erecta</i>	Ast., Tageteae	119,271,184	111,034,049	Paired	CLC ws60	516	83,424	82%	100%	89%	100%	95%	98%
								3,073	2,073	177	49	97	43
<i>Tagetes patula</i>	Ast., Tageteae	46,800,051	43,705,495	Single	CLC ws60	421	47,748	100%	100%	78%	100%	100%	85%

								754	2,774	195	205	235	22
<i>Gaillardia pulchella</i>	Ast., Helenieae	62,546,810	56,658,373	Single	CLC ws23	245	273,507	98%	59%	96%	100%	100%	71%
								4,147	167	119	1,288	19	9
<i>Gaillardia aristata</i>	Ast., Helenieae	40,820,201	39,464,777	Single	CLC ws60	327	100,147	100%	80%	56%	94%	94%	72%
								24,636	7,764	56	66	51	36
<i>Marshallia caespitosa</i>	Ast., Helenieae	74,500,863	72,843,576	Single	CLC ws40	487	126,238	75%	66%	93%	100%	100%	100%
								9,625	4	397	356	380	35
<i>Coreopsis lanceolata</i>	Ast., Coreopsideae	23,651,946	20,352,577	Single	CLC ws40	429	70,580	100%	100%	100%	100%	98%	100%
								26	99	624	20,556	23	22
<i>Coreopsis tinctoria</i>	Ast., Coreopsideae	29,834,448	26,938,300	Paired	CLC ws40	378	60,095	81%	89%	99%	100%	52%	100%
								508	15	348	4,736	106	171
<i>Coreopsis grandiflora</i>	Ast., Coreopsideae	33,427,711	28,201,911	Single	CLC ws60	572	35,699	100%	70%	100%	100%	82%	88%
								2,530	5,101	605	2,421	340	97
<i>Cosmos bipinnatus</i>	Ast., Coreopsideae	129,914,594	119,978,233	Paired	CLC ws60	386	80,555	78%	55%	72%	100%	45%	77%
								29	22,497	252	3,259	35	111
<i>Cosmos sulphureus</i>	Ast., Coreopsideae	55,132,348	38,597,544	Single	CLC ws60	573	55,604	100%	100%	100%	100%	78%	57%
								3,788	2,392	1,283	191	33	27
<i>Bidens pilosa</i>	Ast., Coreopsideae	92,089,417	65,294,097	Single	CLC ws60	348	156,267	100%	82%	67%	100%	61%	50%
								1,763	18	114	196	14	27
<i>Dahlia variabilis</i>	Ast., Coreopsideae	24,436,540	20,967,556	Single	CLC ws60	372	56,066	55%	44%	34%	100%	29%	82%
								611	6,635	800	329	8	30
<i>Inula helenium</i>	Ast., Inuleae	22,215,787	21,201,303	Single	CLC ws60	435	51,881	92%	90%	100%	98%	47%	49%
								4,889	20	134	107	17	14
<i>Inula racemosa</i>	Ast., Inuleae	47,085,168	31,325,354	Single	CLC ws60	414	96,648	100%	99%	54%	100%	100%	100%
								6,532	2,878	274	594	313	356
<i>Inula ensifolia</i>	Ast., Inuleae	89,301,446	57,016,848	Single	CLC ws60	386	107,208	80%	99%	83%	100%	100%	100%
								43,682	4,996	72	219	100	88
<i>Bupthalmum salicifolium</i>	Ast., Inuleae	105,907,022	74,653,164	Single	CLC ws60	358	144,143	78%	91%	73%	100%	100%	47%
								2,508	3,358	301	3,311	50	20
<i>Streptoglossa macrocephala</i>	Ast., Inuleae	56,239,109	48,515,631	Single	CLC ws60	802	41,325	100%	91%	85%	100%	100%	100%
								5	3,619	18	16	27	42
<i>Senecio pinnatifolius</i> (var <i>maritimus</i>)	Ast., Senecioneae	96,655,442	92,907,450	Paired	CLC ws60	383	35,283	99%	99%	48%	96%	57%	59%
								14	8,750	376	16	265	57
<i>Senecio pinnatifolius</i> (ssp. <i>latilobus</i>)	Ast., Senecioneae	31,801,068	27,331,894	Single	CLC ws23	340	138,520	99%	98%	28%	100%	64%	100%
								4,083	9	908	7	10	32

<i>Senecio vulgaris</i>	Ast., Senecioneae	74,683,487	48,542,380	Single	CLC ws30	510	80,276	71% 11,155	100% 1,845	55% 5	85% 35	63% 26	99% 132
<i>Senecio magnificus</i>	Ast., Senecioneae	60,746,328	56,181,820	Single	CLC ws60	384	96,732	94% 417	78% 12,899	73% 2,028	100% 202	100% 223	69% 73
<i>Senecio abrotanifolius</i>	Ast., Senecioneae	84,774,709	54,910,251	Single	CLC ws60	376	98,099	100% 2,217	100% 4,852	100% 857	100% 272	100% 113	54% 20
<i>Steirodiscus tagetes</i>	Ast., Senecioneae	69,082,418	39,748,780	Single	CLC ws40	468	71,571	87% 6,578	100% 22,504	40% 1,062	100% 97	73% 29	100% 35
<i>Othonna arborescens</i>	Ast., Senecioneae	58,549,826	38,280,891	Single	CLC ws40	522	72,778	77% 6,474	63% 7	89% 412	100% 498	100% 168	91% 62
<i>Tussilago farfara</i>	Ast., Senecioneae	52,249,807	49,944,830	Single	CLC ws60	543	93,034	100% 4,259	89% 1,160	100% 681	100% 43	100% 165	100% 49
<i>Ligularia przewalskii</i>	Ast., Senecioneae	66,489,253	43,470,219	Single	CLC ws60	402	79,120	79% 9,493	100% 2,451	95% 473	100% 124	100% 105	100% 44
<i>Doronicum grandiflorum</i>	Ast., Doroniceae	44,327,022	42,747,609	Single	CLC ws60	498	63,829	73% 23,930	100% 1,718	100% 1,340	99% 152	87% 49	100% 134
<i>Doronicum orientale</i>	Ast., Doroniceae	76,108,548	48,535,014	Single	CLC ws60	387	74,105	63% 39,409	71% 33	100% 1,733	100% 381	75% 49	100% 83
<i>Doronicum caucasicum</i>	Ast., Doroniceae	44,901,545	38,184,298	Single	CLC ws60	415	79,121	73% 20,318	90% 1,011	100% 1,456	100% 134	100% 81	86% 68
<i>Pyrethrum coccineum</i>	Ast., Anthemideae	49,403,101	43,414,411	Single	CLC ws60	437	92,803	65% 3,308	87% 4,085	92% 486	100% 463	100% 91	81% 59
<i>Tanacetum coccineum</i>	Ast., Anthemideae	48,575,317	45,174,778	Single	CLC ws60	506	46,944	100% 756	83% 1,697	100% 125	100% 73	100% 31	69% 18
<i>Tanacetum balsamita</i>	Ast., Anthemideae	23,094,047	20,435,460	Single	CLC ws60	387	68,223	83% 2,153	100% 1,602	100% 223	100% 59	80% 182	74% 45
<i>Oncosiphon africanus</i>	Ast., Anthemideae	33,164,153	28,488,537	Single	CLC ws60	468	68,625	100% 2,278	100% 4,744	100% 527	100% 22	100% 30	100% 41
<i>Pyrethrum cinerariifolium</i>	Ast., Anthemideae	43,819,987	42,849,228	Single	CLC ws60	467	89,106	77% 3,352	96% 8,645	82% 1,028	100% 2,915	100% 140	94% 60
<i>Anacyclus depressus</i>	Ast., Anthemideae	90,984,447	85,903,036	Single	CLC ws60	369	163,047	100% 22,820	100% 1,498	100% 905	100% 1,069	100% 1,125	87% 26
<i>Leucanthemum vulgare</i>	Ast., Anthemideae	34,400,715	33,598,126	Single	CLC ws60	390	60,095	51% 2,481	40% 18,044	48% 611	95% 271	71% 24	52% 16
<i>Leucanthemum maximum</i>	Ast., Anthemideae	46,964,607	41,293,685	Single	CLC ws60	545	28,286	63%	85%	62%	84%	53%	85%

								18,296	23,587	673	298	343	37
<i>Mauranthemum paludosum</i>	Ast., Anthemideae	55,228,215	54,105,457	Single	CLC ws40	426	130,810	100% 8,666	98% 35,804	78% 47	89% 305	100% 17	100% 66
<i>Lonas annua</i>	Ast., Anthemideae	35,319,761	30,141,996	Single	CLC ws60	639	32,334	100% 88	99% 48	100% 457	100% 154	100% 507	92% 47
<i>Artemisia absinthium</i>	Ast., Anthemideae	38,819,357	33,613,253	Single	CLC ws60	607	38,236	71% 5,406	73% 4,082	100% 1,001	100% 284	69% 28	89% 32
<i>Achillea millefolium</i>	Ast., Anthemideae	41,339,362	38,263,122	Single	CLC ws60	346	59,797	72% 8,554	96% 767	92% 337	100% 94	58% 277	100% 56
<i>Achillea filipendulina</i>	Ast., Anthemideae	72,231,895	67,137,040	Single	CLC ws60	415	121,147	100% 2,692	98% 8,429	88% 1,270	100% 80	88% 83	73% 39
<i>Achillea ptarmica</i>	Ast., Anthemideae	52,558,260	50,098,884	Single	CLC ws60	401	112,728	77% 11,454	73% 14	100% 1,182	100% 81	52% 17	96% 33
<i>Ursinia anthemoides</i>	Ast., Anthemideae	66,143,868	62,019,963	Single	CLC ws30	581	101,149	82% 6	100% 119,937	100% 1,820	100% 266	73% 25	87% 153
<i>Callistephus chinensis</i>	Ast., Astereae	209,661,666	190,217,871	Paired	CLC ws60	686	70,246	100% 62,414	99% 41	100% 3,777	100% 195	100% 362	100% 116
<i>Brachyscome iberidifolia</i>	Ast., Astereae	28,990,205	25,002,349	Single	CLC ws60	520	38,852	76% 1,380	58% 554	32% 173	70% 11	51% 271	83% 23
<i>Olearia lyrata</i>	Ast., Astereae	75,072,783	72,731,237	Single	CLC ws40	345	181,211	70% 3,360	100% 2,418	88% 234	100% 8	96% 23	96% 46
<i>Osteospermum sinuatum</i>	Ast., Calenduleae	59,135,285	39,092,859	Single	CLC ws40	452	84,609	100% 6,117	100% 15	95% 270	100% 39	48% 138	100% 32
<i>Osteospermum jucundum</i>	Ast., Calenduleae	36,163,781	20,221,990	Single	CLC ws60	420	52,372	58% 1,603	91% 952	100% 164	100% 20	83% 29	93% 105
<i>Schoenia filifolia</i>	Ast., Gnaphalieae	49,711,790	46,293,567	Single	CLC ws60	445	51,789	100% 6,855	81% 2,487	53% 96	100% 466	33% 21	57% 11
<i>Ixodia achilleoides</i> (ssp. <i>alata</i>)	Ast., Gnaphalieae	34,413,431	33,434,513	Single	CLC ws30	326	105,143	99% 102	100% 81	73% 11	61% 5	43% 19	35% 4
<i>Waitzia nitida</i>	Ast., Gnaphalieae	76,822,084	73,963,638	Paired	CLC ws60	337	43,195	79% 1,078	52% 1,173	45% 45	78% 31	55% 38	68% 19
<i>Helichrysum bracteatum</i>	Ast., Gnaphalieae	42,955,662	40,272,625	Single	CLC ws60	409	80,554	51% 2,788	85% 1,934	50% 498	78% 78	64% 35	76% 33
<i>Helichrysum macranthum</i>	Ast., Gnaphalieae	39,894,741	34,445,216	Single	CLC ws60	575	25,782	96% 7,412	47% 4,894	61% 163	75% 60	100% 30	100% 45

<i>Xerochrysum viscosum</i>	Ast., Gnaphalieae	39,537,058	38,000,736	Single	CLC ws60	371	68,458	87% 808	39% 385	43% 482	100% 37	58% 29	100% 37
<i>Arctotis gumbletonii</i>	Cic., Arctotideae	31,595,689	31,033,452	Single	CLC ws40	447	64,684	88% 16,284	52% 12,536	100% 68	82% 6	100% 22	100% 16
<i>Arctotis scapigera</i>	Cic., Arctotideae	37,403,241	34,871,583	Single	CLC ws60	357	88,084	82% 3,875	61% 4,670	100% 60	100% 19	26% 11	78% 110
<i>Arctotis venusta</i>	Cic., Arctotideae	36,762,489	36,096,530	Single	CLC ws23	285	167,329	68% 20,011	47% 3	57% 2,006	92% 14	100% 201	100% 115
<i>Arctotis fastuosa</i>	Cic., Arctotideae	38,323,077	37,489,836	Single	CLC ws30	463	60,961	88% 9,860	99% 344	54% 2,079	100% 382	98% 19	89% 7
<i>Arctotheca populifolia</i>	Cic., Arctotideae	58,780,183	54,956,295	Single	CLC ws60	525	67,810	93% 29,956	100% 5,794	77% 1,267	98% 36	99% 84	100% 57
<i>Gazania krebsiana</i>	Cic., Arctotideae	33,026,167	32,465,268	Single	CLC ws40	487	73,922	79% 5	70% 5,902	100% 664	99% 125	80% 16	74% 14
<i>Gazania rigens</i>	Cic., Arctotideae	61,773,632	52,164,242	Single	CLC ws60	432	77,919	55% 3,463	99% 59	99% 65	84% 341	72% 54	100% 70
<i>Gazania x hybrida</i>	Cic., Arctotideae	100,882,542	96,067,367	Paired	CLC ws60	386	23,311	99% 667	80% 466	82% 99	87% 56	100% 30	74% 21
<i>Gazania splendens</i>	Cic., Arctotideae	26,133,716	22,432,350	Single	CLC ws40	511	53,969	100% 1,446	100% 16	90% 445	100% 279	71% 27	88% 22
<i>Hirpicium armerioides</i>	Cic., Arctotideae	37,321,609	36,731,543	Single	CLC ws60	438	65,181	72% 7,754	66% 1,970	100% 66	54% 56	42% 17	100% 46
<i>Taraxacum officinale</i>	Cic., Cichorieae	83,391,923	78,034,917	Single	CLC ws40	411	25,495	99% 467	99% 1,783	100% 177	58% 16	100% 47	51% 5
<i>Sonchus asper</i>	Cic., Cichorieae	44,257,348	43,206,332	Single	CLC ws60	431	86,871	82% 1,788	100% 26,314	60% 2,463	100% 134	55% 19	96% 38
<i>Tragopogon porrifolius</i>	Cic., Cichorieae	41,666,745	40,572,819	Single	CLC ws60	459	49,268	100% 4,203	100% 4,918	100% 1,429	100% 70	72% 71	74% 41
<i>Scorzonera hispanica</i>	Cic., Cichorieae	37,374,881	36,202,461	Single	CLC ws60	444	44,837	87% 2,060	100% 10,433	100% 427	67% 96	62% 279	100% 26
<i>Cichorium endivia</i>	Cic., Cichorieae	251,818,672	225,578,699	Paired	CLC ws60	872	72,831	92% 28	100% 228,163	100% 3,264	100% 3,121	100% 73	100% 45
<i>Centaurea pseudophrygia</i>	Cic., Cardueae	30,932,695	30,201,420	Single	CLC ws30	560	64,626	100% 21,471	72% 4,231	100% 402	100% 26	59% 18	92% 16
<i>Centaurea macrocephala</i>	Car., Cardueae	36,811,790	36,020,954	Single	CLC ws60	454	88,940	80%	99%	100%	91%	79%	100%

								7	2,287	599	25	36	44
<i>Centaurea cyanus</i>	Car., Cardueae	89,683,424	83,762,853	Paired	CLC ws60	437	82,620	99%	87%	95%	89%	45%	100%
								3,712	5,265	581	289	338	65
<i>Centaurea montana</i>	Car., Cardueae	30,414,531	29,692,754	Single	CLC ws30	497	64,520	81%	85%	70%	89%	67%	98%
								14,760	9	1,181	13	21	11
<i>Centaurea cyanoides</i>	Car., Cardueae	32,652,078	31,889,259	Single	CLC ws30	522	84,551	74%	51%	100%	100%	100%	99%
								32,083	7,975	334	14	20	42
<i>Centaurea dealbata</i>	Car., Cardueae	29,212,848	28,566,884	Single	CLC ws60	429	76,079	100%	68%	77%	100%	56%	100%
								4,367	1,478	323	89	35	27
<i>Centaurea pulcherrima</i>	Car., Cardueae	30,008,981	29,285,799	Single	CLC ws23	306	240,724	63%	98%	31%	100%	88%	71%
								8,005	2,066	58	597	8	16
<i>Plectocephalus americanus</i>	Car., Cardueae	44,527,130	42,954,468	Single	CLC ws60	477	76,747	61%	40%	100%	100%	95%	100%
								6,763	11,800	347	190	14	16
<i>Cynara scolymus</i>	Car., Cardueae	71,866,405	50,288,010	Single	CLC ws60	406	128,199	72%	100%	65%	78%	86%	57%
								29,401	29	126	1,816	32	19
<i>Xeranthemum annuum</i>	Car., Cardueae	89,968,206	86,390,939	Paired	CLC ws60	600	72,718	99%	99%	100%	100%	100%	100%
								13,978	24,330	23	88	2,231	25
<i>Echinops ritro</i>	Car., Cardueae	23,887,753	20,173,225	Single	CLC ws60	425	62,534	100%	100%	100%	100%	60%	85%
								131	6,758	73	175	31	35
<i>Mutisia ilicifolia</i>	Mut., Mutisieae	34,766,512	33,906,511	Single	CLC ws60	446	69,219	100%	99%	100%	95%	98%	100%
								4,530	2,358	3,299	3,869	173	86
<i>Gerbera jamesonii</i>	Mut., Mutisieae	121,137,318	110,410,018	Paired	CLC ws30	510	24,623	94%	93%	100%	100%	99%	99%
								788	659	26	6	15	15
<i>Amblysperma spathulata</i>	Mut., Mutisieae	52,998,961	50,593,058	Single	CLC ws60	414	36,802	65%	81%	100%	56%	45%	100%
								319	558	145	846	34	65

Supplementary Table 2. Primers used to clone genes

Primer	Sequence (5' to 3')	Purpose	PCR template
AJ73	ACA ATG ACT AAC GTT GCA GCG	Clone <i>Senecio pinnatifolius</i> var <i>maritimus</i> PawL1a	Genomic DNA
AJ74	AAC TCG ATC GCT TCG CCG TTG	Clone <i>Senecio pinnatifolius</i> var <i>maritimus</i> PawL1a	Genomic DNA
AJ75	CCA ATA ATG TCG AAG CTT GCT CTT	Clone <i>Senecio pinnatifolius</i> var <i>maritimus</i> PawL1b	Genomic DNA
AJ76	AAC CAC ACG ATT GTG AAT ACG	Clone <i>Senecio pinnatifolius</i> var <i>maritimus</i> PawL1b	Genomic DNA
AJ206	TAT ATA ATG TCA AAG CTT GCT C	Clone <i>Senecio abrotanifolius</i> PawL1a	Genomic DNA
AJ156	CTC TAT AGA GTA ACC ACA CGA CTG	Clone <i>Senecio abrotanifolius</i> PawL1a	Genomic DNA
AJ241	CGG GCC GCC GGT AGA TTT TGA CAG TC	3' RACE <i>Othonna arborescens</i> PawL1b, PawL1f, and PawL1g	cDNA
AJ165	CCA CTT GCC TAA GCT GGT TAC AAC	5' RACE <i>Othonna arborescens</i> PawL1b, PawL1f, and PawL1g	cDNA
AJ259	CAT CCA ACA ACA ATG GCA AAA C	Clone <i>Othonna arborescens</i> PawL1b, PawL1f, and PawL1g	cDNA
AJ260	CAT AAT TTA CGG TTC GCA TAC G	Clone <i>Othonna arborescens</i> PawL1b, PawL1f, and PawL1g	cDNA
AJ175	GAT GAC AAC TAT TTG GAG GAG TAT AC	3' RACE <i>Steirodiscus tagetes</i> PawL1c	cDNA
AJ237	CCA ATA ATG GCA AAA CTT GCT C	Clone <i>Steirodiscus tagetes</i> PawL1c	cDNA
AJ238	TAA ACA TCT GAA GCC ATA CGT AC	Clone <i>Steirodiscus tagetes</i> PawL1c	cDNA
AJ268	CGA TCT ACA AAA CAT CCG ACA ATG	Clone <i>Ligularia przewalskii</i> PawL1a	Genomic DNA
AJ269	CCT CTT TAC CGG TTT GAC GCA TAC	Clone <i>Ligularia przewalskii</i> PawL1a	Genomic DNA
AJ116	GCA AGT TGC AGT CAT TTG GTA GT	5' RACE <i>Tanacetum coccineum</i> PawL1a	cDNA
AJ124	ACA ACA ATG GCA AAA CTT GCA C	Clone <i>Tanacetum coccineum</i> PawL1a	cDNA
AJ125	CCG TTT GAT TAG TGT ACA CTG	Clone <i>Tanacetum coccineum</i> PawL1a	cDNA
AJ79	CAA CAT GGG ATT GAA AGA GGT CC	Clone <i>Artemisia absinthium</i> PawL-CC.a	Genomic DNA
AJ80	GAG TTC ATT AGC CTT CGT CAA CAT CTC	Clone <i>Artemisia absinthium</i> PawL-CC.a	Genomic DNA
AJ147	ATC ATA ATG GCA AAT CAG CTC GC	Clone <i>Inula ensifolia</i> PawL1a	Genomic DNA
AJ148	GAT ACG TAC GAA CTG CCT TCA C	Clone <i>Inula ensifolia</i> PawL1a	Genomic DNA
AJ263	CAT ACT ACA ATG GCA AAA CTT GC	Clone <i>Inula racemosa</i> PawL1a and PawL1b	Genomic DNA
AJ264	TAA CAA AGC GTA ATT AAA AGC	Clone <i>Inula racemosa</i> PawL1a	Genomic DNA
AJ265	GTA CGT GCG ACC TAG CTA GC	Clone <i>Inula racemosa</i> PawL1b	Genomic DNA
AJ266	CAC CGG CAA TTA AAC ATC TCG ATC	Clone <i>Inula racemosa</i> PawL1c	Genomic DNA
AJ267	GCG AGT AAA TAT GCA CAA GAG TAC	Clone <i>Inula racemosa</i> PawL1c	Genomic DNA
AJ261	CGA TCT CTC ATA ATG GCA AAT C	Clone <i>Buphthalmum salicifolium</i> PawL1a	Genomic DNA
AJ262	GAA ACT GCC GGT AGA TAT GCA C	Clone <i>Buphthalmum salicifolium</i> PawL1a	Genomic DNA
AJ83	ACA ATG GTA AAA CTT GTA ATC GTT GC	Clone <i>Streptoglossa macrocephala</i> PawL1a	Genomic DNA
AJ84	CTG GCT TTC ACA TCT CGA TCG AGC	Clone <i>Streptoglossa macrocephala</i> PawL1a	Genomic DNA

AJ85	TAA TCC ACA ATG GCT AAA GCT GTA	Clone <i>Streptoglossa macrocephala</i> PawL1b	Genomic DNA
AJ86	TAG CTT GGA ACT CGC TTT CAG	Clone <i>Streptoglossa macrocephala</i> PawL1b	Genomic DNA
AJ113	ACC ATG GCA AAA TTA CTT GCA G	Clone <i>Gaillardia pulchella</i> PawL1a	cDNA
AJ114	CAA TTC TAC CAA ACT TCA ACA GC	Clone <i>Gaillardia pulchella</i> PawL1a	cDNA
AJ111	GCA CTA ACA GCA CTC GCC TTT GCG CTC	3' RACE <i>Gaillardia pulchella</i> PawS(-)alb.b and PawS(-)alb.e	cDNA
AJ142	ATG ACA AAA GTT GCA CTA ACA GCA CTC	Clone <i>Gaillardia pulchella</i> PawS(-)alb.b and PawS(-)alb.e	cDNA
AJ144	CCG TAT AAT ATA ACT CGT TTA AGC	Clone <i>Gaillardia pulchella</i> PawS(-)alb.b	cDNA
AJ143	GCA GCT AGT ACA TAC AAG CCC	Clone <i>Gaillardia pulchella</i> PawS(-)alb.e	cDNA
AJ108	GTT TGT AGT CCC TTC CCC TTC TG	3' RACE <i>Gaillardia pulchella</i> PawS1a	cDNA
AJ109	GAT GGC CTC ACA CTG ACA CTC CTC	5' RACE <i>Gaillardia pulchella</i> PawS1a	cDNA
AJ117	TCC ACA ATG ACA AAA GTT GCC GC	Clone <i>Gaillardia pulchella</i> PawS1a	cDNA
AJ118	CAT AGA AAG CAC ACT ACA TGA G	Clone <i>Gaillardia pulchella</i> PawS1a	cDNA
AJ192	GCT TGG TCA AAC ACT TGC TGG ATC	5' RACE <i>Dahlia variabilis</i> PawL1b, PawL1e, and PawL1f	cDNA
AJ193	GCC CGG ACA GCC TGA AGG ACA AC	3' RACE <i>Dahlia variabilis</i> PawL1b, PawL1e, and PawL1f	cDNA
AJ251	GAC GAT ATG ATG ATA AAA GTT TC	Clone <i>Dahlia variabilis</i> PawL1b and PawL1e	cDNA
AJ252	GAT GAA AAC CTT AAA CCA TGA TC	Clone <i>Dahlia variabilis</i> PawL1b and PawL1e	cDNA
AJ253	CCG ACG ATG ACA AAA GTT TCA G	Clone <i>Dahlia variabilis</i> PawL1f	cDNA
AJ254	GGA GCT ACC GAC CAT TAC TGT	Clone <i>Dahlia variabilis</i> PawL1f	cDNA
AJ149	ACC AAC ATG GCA AAA CTT GCG	Clone <i>Coreopsis tinctoria</i> PawL1a	Genomic DNA
AJ150	GTT CGC TTC ACC AGC TTA TAA G	Clone <i>Coreopsis tinctoria</i> PawL1a	Genomic DNA
AJ25	CAA CCA TCA ACC ACC GTC AAG	Clone <i>Coreopsis tinctoria</i> PawL1b	Genomic DNA
AJ26	CTC ATG ATT TAT AGC GAT CGG TGA	Clone <i>Coreopsis tinctoria</i> PawL1b	Genomic DNA
AJ153	TCC ATC CGA CTT ACT ATA ATG AAA C	Clone <i>Coreopsis tinctoria</i> PawL(-)alb.a and PawL(-)alb.b	Genomic DNA
AJ154	CAT TTC ATA GTG GAA GCA AGT G	Clone <i>Coreopsis tinctoria</i> PawL(-)alb.a and PawL(-)alb.b	Genomic DNA
AJ199	ACA ATG GCG AAA CTT GCA GTT	Clone <i>Cosmos bipinnatus</i> PawL1b	cDNA
AJ200	GGA ACT GAT CGC ATC ACA ATA AGC	Clone <i>Cosmos bipinnatus</i> PawL1b	cDNA
AJ207	GAT CGT GTT ACA ACA ATG GCG	Clone <i>Bidens pilosa</i> PawL1a	Genomic DNA
AJ208	CAT GAT GGG CAC CGG GTT GTT TG	Clone <i>Bidens pilosa</i> PawL1a	Genomic DNA
AJ205	GCA CTC ACC TTG ACG GCC ATC GTA G	3' RACE <i>Hymenopappus tenuifolius</i> PawL1b	cDNA
AJ257	CAT CCG GCA TTA ATT CCT GCA ATG	Clone <i>Hymenopappus tenuifolius</i> PawL1b	cDNA
AJ258	GAT TAG CAC GAC CAC ACA CAT GG	Clone <i>Hymenopappus tenuifolius</i> PawL1b	cDNA
AJ128	CAA CCA CCT TCA ACC ATG GCA	Clone <i>Zinnia elegans</i> PawL1a	Genomic DNA
AJ129	TCT TCT TAT TCA AAC CCG CGG GGC	Clone <i>Zinnia elegans</i> PawL1a	Genomic DNA
AJ130	TCC ACA ATG ACT AAA CTC ACA CTC	Clone <i>Zinnia elegans</i> PawL-C.a	Genomic DNA

AJ131	AAG CAC AAG GAC CAA CCA AGC	Clone <i>Zinnia elegans</i> PawL-C.a	Genomic DNA
AJ21	GAC CTC CTT CTC CGT CAC TAC	Clone <i>Echinacea purpurea</i> PawL1a and PawL1b	Genomic DNA
AJ22	TCG ATC GCT TTA CAA GCT AAC	Clone <i>Echinacea purpurea</i> PawL1a and PawL1b	Genomic DNA
AJ23	CTA GCT ATG ACA AAA TTC GAT CTT	Clone <i>Ambrosia trifida</i> PawL1a	Genomic DNA
AJ24	TAC AAC TAA CGG CGA TCG ATC ATC	Clone <i>Ambrosia trifida</i> PawL1a	Genomic DNA
AJ138	CAT CCC ACA ATG GCA AAA CTG GCA	Clone <i>Tetragonotheca texana</i> PawL1a	Genomic DNA
AJ139	GAG GCA ATA TAA CGC ACA CGA T	Clone <i>Tetragonotheca texana</i> PawL1a	Genomic DNA
AJ157	GTA CAC ATC CCA TCC ATA ATG GCA	Clone <i>Melampodium paludosum</i> PawL1a	Genomic DNA
AJ158	GCT GGA AAC ACA CTA CTT AGC	Clone <i>Melampodium paludosum</i> PawL1a	Genomic DNA
AJ159	CGT ACA CAT CCC ACA ATG GCA	Clone <i>Melampodium paludosum</i> PawL1b	Genomic DNA
AJ160	GAC ACA ACC ATA CGT ACG TAG	Clone <i>Melampodium paludosum</i> PawL1b	Genomic DNA

Supplementary Figures

Supplementary Figure 1: Box plots of core gene expression vs *PawS1* and *PawL1* transcripts

Average coverage was normalized for sequencing depth by dividing by the total number of clean reads used for assembly. Each whisker extends to 1.5 times the inter-quartile range, with outliers shown as dots.

Supplementary Figure 2: LC-MS detection of all masses corresponding to novel PDPs in seed extracts

From the transcriptomic data, evidence of novel PDPs was sought by searching LC-MS data for masses matching predictions from the corresponding *PawS* gene based on the conserved asparaginyl endopeptidase processing sites as elucidated for sunflower PawS1. In each panel, on the left is the predicted peptide structure. M_e is the expected mass-to-charge and M_o is the observed value. Some species contained a transcript encoding more than one peptide and some species had multiple transcripts encodings more than one peptide, explaining multiple LC-MS data panels for some species. Each panel contains the extracted ion chromatogram (XIC) and inset is the mass spectrum of the predicted native peptides. The average retention time for each peptide is labelled on the XIC.

Supplementary Datasets

Supplementary Dataset 1: Core transcripts used to assess the quality of assemblies

All 110 transcriptomes were queried with the six sunflower core transcripts described in Jayasena *et al.* (2014) using the tBLASTn procedure. The transcript with the top matching score was selected from the tBLASTn output. Data is shown for one among the four assemblies (the same assembly as in **Table S1**).

Supplementary Dataset 2: All *PawS1/PawS1* intermediate transcripts identified from *de novo* transcriptomes

Partial transcripts are indicated by a dash. Cys residues are underlined and the Pro-Pro hinge is double underlined. A color scheme was used to denote predictions that were made in line with the known processing of *Arnica montana* PawL1, *H. annuus* PawS1 and PawS2 and several other albumins; colors indicate ER signal (pink), spacer (black), PDP region (aqua), small albumin subunit (green) and large albumin subunit (orange).

Supplementary Dataset 3: Cloned genes

Details of the primers used to clone the genes and the PCR template are shown in **Table S2**. Color scheme is as **Dataset 2**. Region delimitation was done by comparison with sunflower PawS1.

Supplementary Dataset 4: Alignment of 105 SESA sequences used to create the SESA WebLogo

These seed storage albumin precursor (SESA) sequences are from 105 different Asteraceae seed *de novo* transcriptomes that we assembled. The albumin small sub-unit (SSU) sequence was used for the alignment, including four amino acids preceding the SSU and 1-4 amino acids from the albumin large sub-unit (LSU) to maintain a consistent length. Sequences were aligned using CLC Genomics Workbench 8.2 maintaining alignment parameters as gap open cost: 5.0, gap extension cost: 5.0, and end gap cost: free and manually altered based on known processing sites and Cys residue location.

Supplementary Dataset 5: Alignment of 99 PawL1 sequences used to create the PawL1 WebLogo.

These 99 PawL1 sequences represent all that were identified in the 110 *de novo* transcriptomes. The SSU sequence containing the PDP-like region was used for the alignment including four amino acids preceding the SSU and 1-4 amino acids from the LSU to maintain a consistent length. Sequences were aligned as done for **Dataset 4**.

Supplementary Dataset 6: Alignment of 49 PawS1 sequence used to create the PawS1 WebLogo.

These 49 PawS1 sequences represent all that were identified in the 110 *de novo* transcriptomes. The SSU sequence containing the PDP region was used for the alignment including four amino acids preceding the SSU and 1-4 amino acids from the LSU to maintain a consistent length. Sequences were aligned as done for **Dataset 4**.

Supplementary Dataset 7: Alignment used to generate phylogeny in FASTA format.

Sequences were edited and assembled into primer-based matrices prior to concatenation. Using Mesquite, nbrf files were created which were aligned using the program MAFFT. The aligned data was edited to a simplified Nexus interleaved file prior to performing concatenation. The concatenated file was used for time analyses using Mesquite.

Supplementary Dataset 8: ITS and chloroplast gene GenBank numbers used to generate phylogeny.

This table lists the GenBank codes for the ITS and chloroplast genes used to generate the phylogeny in **Figure 2**. The table is listed in the same order as Figure 2 and includes the following information: sub-family, tribe, accepted name followed by the GenBank codes for ITS, matK, ndhF or rbcL. Some species synonyms are included in the final table column.