A tripartite approach identifies the major sunflower seed albumins

Achala S. Jayasena[1]

Bastian Franke[2]

Johan Rosengren[2]

Joshua S. Mylne[1*]

[1]The University of Western Australia, School of Chemistry and Biochemistry & ARC Centre of Excellence in Plant Energy Biology, 35 Stirling Highway, Crawley, Perth 6009, Australia

[2]School of Biomedical Sciences, The University of Queensland, St Lucia, Queensland 4072, Australia

[*]Corresponding author

E mail: joshua.mylne@uwa.edu.au

Phone: +61 8 6488 4415
Fax: +61 8 6488 1005

Acknowledgements

A.S.J. is supported by an International Postgraduate Research Scholarship and an Australian Postgraduate Award. B.F. was supported by Australian Research Council grant DP120103369. J.R. and J.S.M. are supported by Australian Research Council Future Fellowships FT130100890 and FT120100013 respectively. Authors would like to acknowledge Prof. Loren Rieseberg (University of British Columbia) for providing access to a draft sunflower genome, the BAC/EST Resource Center of the Arizona Genomics Institute (University of Arizona) and David Kudrna for *H. annuus* cDNA clones and the Compositae Genome Project website (http://cgpdb.ucdavis.edu/) supported by the USDA IFAFS Programme and NSF Plant Genome Programme for EST data. The authors also thank Michelle Colgrave, Nicolas Taylor, Richard Jacoby and Mark Condina for valuable advice on proteomics. This work was supported by Australian Research Council grant DP130101191.

Abstract

Seed protein content is determined by the expression of what are typically large gene families. A major class of seed storage proteins is the napin-type, water soluble albumins. In this work we provide a comprehensive analysis of the napin-type albumin content of the common sunflower (*Helianthus annuus*) by analyzing a draft genome, a transcriptome and performing a proteomic analysis of the seed albumin fraction. We show that although sunflower contains at least 26 genes for napin-type albumins, only 15 of these are present at the mRNA level. We found protein evidence for eleven of these but the albumin content of mature seeds is dominated by the encoded products of just three genes. So despite high genetic redundancy for albumins, only a small sub-set of this gene family contributes to total seed albumin content. The three genes identified as producing the majority of sunflower seed albumin are potential future candidates for manipulation through genetics and breeding.

Key words: napin, genomics, transcriptomics, proteomics

Author contribution statement

J.S.M. conceived the study; A.S.J. performed all bioinformatics and cloned the genes; A.S.J., B.F., and J.S.M. did protein extraction and FPLC; B.F. performed proteomics; all authors analyzed the data; J.S.M. and A.S.J. wrote the manuscript with input and final approval from all authors.

Key message: We have used a combination of genomic, transcriptomic, and proteomic approaches to identify the napin-type albumin genes in sunflower and define their contributions to the seed albumin pool.

Introduction

Seeds are a major source of protein for humans and livestock. Laid down as a source of nitrogen and sulfur for the germinating seedling, seed storage proteins can constitute over 50% of total seed protein (Shewry and Halford 2002). Plant seed storage proteins are classified into four categories based on the solvent in which they dissolve most readily. Albumins, globulins, prolamins, and glutelins dissolve most readily in water, dilute saline, water/alcohol mixture, and dilute acid or alkali respectively (Shewry et al. 1995). The focus in this work is upon albumins. Found in most dicot seeds, these abundant proteins have also been referred to as napins or 2S albumins.

Despite their obvious importance as a source of nutrients for the emerging seedling, various albumins have been reported to possess bactericidal (Maria-Neto et al. 2011) and fungicidal activities (Freire et al. 2015; Ribeiro et al. 2012; Terras et al. 1992). Albumins have been considered as agronomic traits, but only when they have a deleterious effect such as the allergenic albumins Ber e 1 in *Bertholletia excelsa* (Brazil nut) (Nordlee et al. 1996), Sin a I in *Sinapis alba* L. (yellow mustard) (Menéndez-Arias et al. 1988), Ric c I in *Ricinus communis* (castor bean) (Thorpe et al. 1988), Ara h 2 and Ara h 6 in *Arachis hypogaea* (peanut) (Zhou et al. 2013). Food allergy has become a great concern and several methods have been employed to minimize the deleterious effects of allergens. An attempt was made to develop a hypoallergenic peanut by gene silencing of Ara h 2 and Ara h 6 with RNA interference and this significantly reduced human Immunoglobulin E binding to these two allergens (Chu et al. 2008). However as peanut allergens contribute to a significant proportion of total peanut proteins, allergen removal or depletion might affect the taste of the modified peanuts (Zhou et al. 2013).

The number of albumin seed storage protein genes can vary from as few as five in *Arabidopsis* (Krebbers et al. 1988; van der Klei et al. 1993) to at least ten in Brazil nut (Moreno et al. 2004) and over 16 in *Brassica napus* (rapeseed) (Scofield and Crouch 1987). Despite such genetic redundancy, only a subset might contribute to the seed storage protein pool. For example, it is reported that only a subclass of six napin-encoding genes are highly expressed during the embryogenesis in rapeseed (Blundy et al. 1991) although at least 16 genes are present (Scofield and Crouch 1987). Eight consensus isoforms of the mustard 2S albumin allergen Sin a 1 have been identified and they were shown to vary in their relative abundance (Hummel et al. 2015).

The seed proteomes of several species have been analysed. The lotus (*Nelumbo nucifera* Gaertn.) seed proteome was catalogued by combining 1 dimensional gel electrophoresis and 2 dimensional gel electrophoresis separation with tandem mass spectrometry. Among the 96 non-redundant proteins identified, proteins involved in carbohydrate metabolism and nutrient storage were identified as the most abundant (Moro et al. 2015). Various developmental stages of *Arabidopsis* seeds have been analysed to identify components of its proteome that were up-regulated and down-regulated at different stages of seed maturation (Chibani et al. 2006; Gallardo et al. 2001; Rajjou et al. 2008). Two dimensional gel electrophoresis coupled with mass spectrometry was used to identify differentially expressed seed proteins in mature-dry and germinating rubber seeds (Wong and Abubakar 2005).

Proteomic approaches employing tandem mass spectrometry (MS/MS) have been used to study fractions of the seed proteome. In this way the relative abundance of glutenin, albumin and globulin was determined by MS/MS for *Brachypodium distachyon* (Wang et al. 2010). Fractionation of rice protein into albumin-globulin, prolamine and glutelin before MS/MS revealed novel types of seed storage proteins in wild rice (Jiang et al. 2014).

Albumins have a highly conserved cysteine (Cys) residue pattern and are often rich in glutamine. Each mature albumin is processed from a precursor. A typical preproalbumin contains an endoplasmic reticulum (ER) signal and a pro-domain which are removed during the maturation process. During processing of the pro-albumin, the albumin is often cleaved into a heterodimer of small and large subunits with molecular weights of around 4 kDa and 9 kDa that remain connected by disulfide bonds (Ericson et al. 1986). Mature albumins generally have eight conserved Cys residues, two in the small subunit and six in the large subunit (Shewry et al. 1995), but there are albumin precursors with ten Cys residues such as peanut Ara h 6 (Lehmann et al. 2006).

Sunflower has several albumins that deviate from this typical structure and processing. For example, the albumin SFA8 is not cleaved into small and large subunits and so remains monomeric (Kortt et al. 1991). The sunflower albumin precursor HaG5 also deviates from the norm and is predicted to encode two mature albumins instead of one (Allen et al. 1987). PawS1 and PawS2 are especially unusual as these pre-proalbumins are each matured into two proteins; a typical hetero-dimeric albumin and a small, disulfide-bonded, macrocyclic peptide (Mylne et al. 2011). Sunflower *PawL1* is thought to be an ancestral form of *PawS1* (Elliott et al. 2014). Studies with seeds of *Arnica montana PawL1* showed that it is not restricted to sunflower and does indeed encode an albumin, but no additional peptide despite having some sequence similarity to the PawS-Derived Peptides of *PawS1* (Elliott et al. 2014).

Albumins, like most seed storage proteins, are typically encoded by multi-gene families. For sunflower at least seven genes are known already. These include the aforementioned *SFA8* (Kortt et al. 1991), *HaG5* (Allen et al. 1987), *PawS1* and *PawS2* (Mylne et al. 2011), *PawL1* (Elliott et al. 2014) as well as *pHAO* (Thoyts et al. 1996). A sequence called *BA* was published only in GenBank (Accession: AJ275962).

To fully understand the reason sunflower contains so many unusual albumins we combine the draft sunflower genome with transcriptomics data and a proteomic analysis of an albumin rich extract to determine the genes that encode sunflower albumins and the abundance of each mRNA transcript and their presence in protein extracts. This tripartite approach has elucidated the full albumin repertoire of sunflower seeds and reveals the proportion of genes that contribute to seed albumin content.

Materials and Methods

Searching the sunflower genome for albumins

To be able to sequence proteins in the sunflower albumin fraction by MS/MS it was first critical to fully annotate the genome for albumin-encoding genes. The complete genome of the common sunflower (*Helianthus annuus*) is still pending (http://www.sunflowergenome.org). We were provided access to a draft version of the genome (Nov22K22sspace2ext.final.scaffolds.fasta) consisting of 349,227 scaffolds. Using tBLASTn algorithm we queried the draft genome for putative napin-type albumin genes. The predicted protein sequences encoded by *PawS1* (GenBank ID: FJ749265)*, PawS2* (FJ469149)*, PawL1* (KF574811), *HaG5* (X06410)*, BA* (AJ275962)*, SFA8* (X56686), and *pHAO* (X76101) were used as the initial queries. The genomic DNA region for the predicted albumin-encoding gene along with 1.5 kb flanking regions either side (*i.e.* unless the gene is located in either end of a sequence scaffold) were annotated to predict the protein precursors. All new SEED STORAGE ALBUMIN (SESA) protein sequences were used to re-interrogate the sunflower genome by tBLASTn to identify as many napin-type albumins as possible. tBLASTn was performed with default settings (Matrix: BLOSUM62, Gap Penalties: Existence: 11, Extension: 1, Neighboring words threshold: 13). The full output was analysed for putative albumins rather than consider specific cut offs such as percentage ID, length etc. Conserved features of known albumins were sought, namely the ER signal, potential small and large subunit with the conserved

pattern of Cys residues. If only a partial gene was identified, the genome was searched again for any overlapping scaffold using the partial open reading frame (ORF) as a query.

Here onwards for simplicity all napin-type albumin precursors except *PawS1*, *PawS2*, and *PawL1* will be referred to as '*SESA*' genes for *SEED STORAGE ALBUMIN*. Known genes *HaG5*, *BA*, *SFA8*, and *pHAO* will be referred to as *SESA1*, *SESA2*, *SESA3*, and *SESA4* respectively for simplicity, but also in an effort to align with community efforts to standardise plant gene nomenclature (Meinke and Koornneef 1997), notably to favour three-letter gene names, avoid species prefixes and to give closely related proteins the same name with a different numerical identifier. Novel *PawS-Like* (*PawL)* genes identified were named as *PawL2* and *PawL3*.

Cloning novel *SESA* genes

All the novel full length sunflower albumin-encoding genes identified in the draft genome were confirmed by cloning. Genomic DNA was extracted from de-hulled mature sunflower seeds using the DNeasy plant mini kit (Qiagen) according to the manufacturer protocol. Unless stated otherwise, polymerase chain reaction (PCR) amplified fragments for each gene was cloned to the pGEM-Teasy vector (Promega) and three independent clones for each were sequenced and compared to identify potential errors introduced during PCR by the DNA polymerase.

In the initial search of the genome only a partial ORF was identified for *PawL2.* Searching the genome by tBLASTn with this partial ORF identified an overlapping scaffold. By combining these two scaffolds, the full length gene could be reconstructed. This information was used to design the primers AJ67 (5'-CTC CTT CAC TAG CAA CCA TCA-3') and AJ69 (5'-CGT ATA CAC ATA CAT AGG CAC ACG-3'). These amplified a 620 bp band by PCR using *Pfu Ultra* high fidelity DNA polymerase (Agilent) and sunflower seed genomic DNA as the template. The PawL2 amino acid sequence encoded by the genome matched the cloned product perfectly except for one amino acid difference in the predicted ER signal (Lys18Phe) caused by a one nucleotide difference.

PawL3 identified in the genome was only 91 amino acid residues long and had a premature stop codon. By analyzing a sunflower seed *de novo* transcriptome (Elliott et al. 2014) a possible full length transcript could be identified. Primers AJ221 (5'-TTA AGA ACA ATG GCC AAA GTT GC-3') and AJ222 (5'-CTA GGA AGT CGA TCG CAA CAC-3') were designed based on this putative full length transcript. AJ221 and AJ222 primers were used to amplify a 555 bp fragment by PCR from sunflower genomic DNA with *Taq* DNA polymerase. However, the cloned product had a premature stop codon exactly as observed in the genomic scaffold and the nucleotide sequence from the genome matched the cloned product perfectly.

The draft genome as well as the *de novo* transcriptome identified a partial ORF matching to the *HaG5* (*SESA1*) sequence published in GenBank (Accession: X06410). Primers AJ239 (5'-CCC ACA ATG GCA AAG CAA ATA G-3') and AJ240 (5'-CCA ACG ACT AGA GAT GCC ACT C-3') were designed based on the GenBank sequence. These primers were used to amplify a product of around 1100 bp by PCR from sunflower genomic DNA with *Taq* polymerase. Cloning followed by sequencing identified the intronic full length *SESA1* (*HaG5*) gene. *SESA1* showed 99% similarity to the publicly available *HaG5* sequence at the nucleotide level.

The SESA2 (BA) predicted amino acid sequence from the draft genome was 99% identical to the sequence deposited in GenBank (Accession: AJ275962). Primers AJ235 (5'-CCA ACA CCA TCT CCC ACA ATG GC-3') and AJ236 (5'-GCT TCC ATC ACA AAG CCA CAA TC-3') were used to amplify this gene by PCR from the sunflower seed genomic DNA with *Taq* DNA polymerase. Cloning and sequencing revealed

that the cloned product was intronic and matched the genome sequence perfectly at the nucleotide level.

Using SESA4 (pHAO) sequence encoded in the sunflower draft genome, the NCBI databases were queried by tBLASTn for matching Expressed Sequence Tags (ESTs). An EST which potentially contains the full-length *SESA4* gene sequence was identified (GenBank Accession: BQ971867.1). Its corresponding complementary DNA (cDNA) clone (QHB8O06) was ordered from the Arizona Genomics Institute, University of Arizona, USA. This cDNA clone was completely sequenced using M13R primer (5'-CAG GAA ACA GCT ATG AC-3') and a *SESA4* specific internal primer AJ49 (5'-AGC CCA TAT GAA CAG AGG CA-3'). By aligning the nucleotide sequences obtained from the genome and the full length cDNA clone (QHB8O06) a potential intron sequence could be predicted. Nucleotide sequences from the genome and the cDNA clone matched 100%, but six amino acids differed between this and the published SESA4/pHAO sequence (Thoyts et al. 1996).

SESA6 was identified from the genome and appeared to be partial, especially because it deviated from the conserved Cys pattern. Using Illumina raw reads from sunflower seed RNA sequencing (NCBI SRA: SAMN02569067) the transcriptome was assembled *de novo* using the CLC Genomics Workbench software as described (Jayasena et al. 2014). Searching this new seed *de novo* transcriptome using tBLASTn with the partial SESA6 sequence identified a matching full length transcript coding for a putative double albumin. Primers AJ31 (5'-CGT GTC CAC CCT CCA AAC CAC-3') and AJ39 (5'-GCA CTA CAC ATG CAT GTG CTC-3') were designed based on this transcript. Amplification of sunflower seed genomic DNA with these primers by PCR using Platinum Taq Hi Fidelity DNA polymerase (Invitrogen) resulted in a band of approximately 1.5 kb. The PCR product was cloned into pGEM-Teasy (Promega) and sequenced in either direction using SP6 (5'-ATT TAG GTG ACA CTA TAG-3') and T7 (5'-GTA ATA CGA CTC ACT ATA GGG C-3') primers. Aligning this nucleotide sequence with the full length transcript from the *de novo* transcriptome identified an intron. Nucleotide sequence of the cloned product matched the transcript from the *de novo* assembly perfectly.

SESA7 , *SESA8*, and *SESA9* full length genes were PCR amplified from the sunflower seed genomic DNA with *Taq* DNA polymerase using the primers AJ229 (5'-CCA ATT ATG GCT AAA CTT ACA AG-3'), AJ230 (5'-GGT TAT ACC GCA CAA CGT TGC-3') and AJ231 (5'-CAT CCT ATA ATG GCA AAA CTT GC-3'), AJ 232 (5'-CTT TGG GCC TGC GAT CGA TAC-3') and AJ225 (5'-CAA ACC ACG ATG GCA AAC CTA AC-3'), AJ226 (5'-GTT ACA CAT GAC CGA CCT ATA C-3') respectively. Primers were designed based on the full length genes identified from the draft genome sequence. Nucleotide sequences of the cloned products for all three genes matched the draft genome perfectly.

Primers AJ33 (5'-GAA CCA CCG TCC TCC ATG CAC-3') and AJ34 (5'-GAA TTA TCT TAT GTG CTC CTT-3') were designed based on the putative full length sequence of *SESA10* identified from the genome and amplified a product of about 600 bp from sunflower genomic DNA with these primers using Platinum Taq Hi Fidelity DNA polymerase (Invitrogen). The nucleotide sequence of *SESA10* matched 98% to the draft genome sequence.

Similarly *SESA12* and *SESA13* genes were amplified by PCR from the sunflower seed genomic DNA with *Taq* DNA polymerase using the primers AJ233 (5'-AAG ATG GAT AAA CTT GCA CTT-3'), AJ234 (5'-CAT TCA CCT AAA CTA TCA TCT TAC-3') and AJ223 (5'-TCT CCA ATG GCG ACA ACA CAA GC-3'), AJ224 (5'-GGA GGT GAA ATG GAC CCT AGA G-3') respectively. Primers were designed based on the sequences from the draft genome. Cloning and sequencing showed that the sequence similarity at the nucleotide level between the cloned product and the genome was 100%.

Primers AJ227 (5'-TCT CCC ACA ATG GCA AAG CTA ATA G-3') and AJ228 (5'-GGA AAC TAC TCG CTT TCA TCT-3') were designed for the *SESA20* sequence identified in the genome. These primers amplified full length *SESA20* by PCR with sunflower genomic DNA and *Taq* DNA polymerase. The PCR product was cloned and sequenced with SP6 and T7 universal primers. *SESA20* was intronic and its sequence matched the genome perfectly.

Pairwise comparison of sunflower albumin sequences

The sunflower SESA sequences (**Supplementary Data Set 1**) were aligned using CLC Genomics Workbench 8.5 software with a range of different gap open and gap extension cost values. No significant improvement in the alignment was observed with different parameters. The alignment created using a gap open cost: 5.0 and gap extension cost: 5.0 was further manually altered based on known albumin maturation sites with greater weight given to conserved Cys residues and the known enzyme target sites (Asn) (**Supplementary Data Set 2**). This manually altered sequence alignment was used to create a pairwise comparison using the same software. The two parameters used for the comparison are the percentage identities (the percentage of identical residues in alignment positions to overlapping alignment positions between the two sequences) and the differences (the number of alignment positions where one sequence is different from the other including the gap differences). A circular cladogram was created to summarise these data from the aforementioned manually altered alignment using CLC default settings (tree construction method: neighbor joining, protein distance measure: Jukes-Cantor).

mRNA level expression

To determine the abundance of each albumin at the mRNA level, raw reads from mature sunflower seed mRNA (SRA Biosample: SAMN02569067) were mapped on to genomic regions for each *SESA* gene using CLC Genomics Workbench 6.5.1 (CLC bio). Approximately 3 kb of genomic DNA sequence was available for the genes *PawS1, PawS2, PawL2, SESA2, SESA3, SESA9, SESA10, SESA13, SESA16,* and *SESA20*, but other genes had less than a 3 kb region. Mapping parameters were maintained as follows. Mismatch cost: 3, insertion cost: 3, deletion cost: 3, length fraction: 0.95, and similarity fraction: 0.95.

Preparation of an albumin-rich fraction from sunflower seeds

To prepare an albumin-rich extract for proteomic analysis, 40 g of mature sunflower seeds were ground to a fine meal. The seed meal was stirred in two tissue volume of hexane and the mixture was poured through a Whatman filter paper to separate oil from the meal. The defatted meal was left to dry and albumins were subsequently extracted essentially as described by (Kortt and Caldwell 1990). The crude albumin extract was lyophilized and 300 mg was dissolved in 5 mL of 50 mM sodium phosphate buffer pH 7.0, 150 mM sodium chloride, 1 mM dithiothreitol and separated by size exclusion through a Sephadex 75 column (exclusion limit >80 kDa). Thirty-five fractions of 3 mL were collected and 10 µL of each fraction was visualized on a NuPAGE 4-12% Bis-Tris gel (Life Technologies). Albumin-rich fractions were identified based on size. The albumin rich fractions were pooled and dialyzed against water overnight at 4°C to remove any salts.

Purification of mature albumins using reversed phase high performance liquid chromatography (RP-HPLC)

To isolate and purify mature albumins the desalted, albumin-rich fraction from the aforementioned size exclusion chromatography was separated further by RP-HPLC on a Shimadzu Prominence system with a semi-preparative Grace Vydac C18 column (250 mm x 10 mm, 10 µm, 300 Å) and an analytical Grace Vydac C18 column (250 mm x 4.6 mm, 5 µm, 300 Å) at a 1% gradient in which buffer A was 0.05% trifluoroacetic acid and buffer B was 90% acetonitrile 0.05% trifluoroacetic acid. Forty fractions were

collected but not all contained protein so only fractions 9 to 40 (F9-F40) are shown in **Fig. 3**. To determine the characteristic mass spectrometric pattern, each albumin fraction was analyzed by liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) on an SCIEX API 2000 LC-MS/MS electrospray mass spectrometer. A volume of 10 µL from each fraction was injected at a flow rate of 0.1 mL/min with buffer A to buffer B ratio of 30:70. Buffer A consisted of 0.1% formic acid and buffer B contained 0.1% formic acid in 90% acetonitrile. Specifically LC-MS/MS instrument settings were as Declustering Potential 88; Focusing Potential 220; Entrance Potential 8; Q1 MS.

In-solution digestion

To determine the identity of any unknown peaks in the analytical trace, the protein was dissolved in 100 µL of water. A 25 µL aliquot of each albumin fraction was mixed with 25 µL of a 50 mM ammonium bicarbonate and 1 mM calcium chloride buffer (pH 8.0). A volume of 5 µL of 100 mM dithiothreitol was added and incubated for 30 min at 60°C. To this, 5 µL of 250 mM iodoacetamide was added and incubated at room temperature in the dark for 60 min. Then 0.5 µg of trypsin (Sigma-Aldrich) was added to each aliquot and incubated at 37°C overnight. Prior to mass spectrometric analysis, 20 µL of 1% formic acid was added. Mass spectrometric analysis was performed using a 5600 TripleTOF™ (TOF = time of flight) mass spectrometer (SCIEX, Toronto, Canada) and in this way identified SESA2-1, SESA2-2 and SESA20-2.

1-D gel electrophoresis

Acrylamide gels (15%) were cast according to the Bio-Rad protocols and 1-D gel electrophoresis was performed at 4°C for 15 min at 300 V. Gels were stained with Coomassie Brilliant Blue G250 Bio-Safe (Bio-Rad) for 1 h and destained in water.

In-gel digestion

Manually excised gel pieces corresponding to visible protein bands in each fraction were washed in 500 µL of water for 5 min. A volume of 200 µL of acetonitrile was added to these gel pieces and they were dehydrated for 5 min at room temperature. Gel pieces were subsequently reduced in 30 µL of 10 mM dithiothreitol prepared in 100 mM ammonium bicarbonate for 30 min at room temperature and alkylated in 30 µL of 100 mM iodoacetamide prepared in 100 mM ammonium bicarbonate for another 30 min at room temperature. Then 60 µL of trypsin (Sigma-Aldrich) at a concentration of 50 ng/µL was added to each gel piece and incubated at 37°C overnight. Peptides were extracted by removing and collecting the supernatant. Gel pieces were washed in 30 µL extraction buffer (50% acetonitrile, 1% formic acid) and supernatant was collected and combined with the first extraction step. Then the gel pieces were washed in 30 µL extraction buffer (80% acetonitrile, 1% formic acid) and the supernatant was collected and combined with the supernatants collected from first and second extraction steps. Extracts were dried by vacuum centrifugation at ambient temperature and 20 µL of 1% formic acid was added. Mass spectrometric analysis was performed using a 5600 TripleTOF™ mass spectrometer (SCIEX, Toronto, Canada).

Nano HPLC, mass spectrometry and protein identification

To determine the protein mass of the intact albumins SESA3/SFA8, PawS1, and PawS2, they were analyzed by LC-MS on a Shimadzu Prominence Nano HPLC (Rydalmere, Australia) coupled to a TripleTOF™ 5600 mass spectrometer (SCIEX, Toronto, Canada) equipped with a nano electrospray ion source. A volume of 5 µL from each extract was injected onto a 50 mm x 300 µm C18 trap column (Agilent Technologies, Australia) at a rate of 60 µL/min. The samples were de-salted on the trap column for 5 min using 0.1% formic acid (aq). The trap column was then placed in-line with the Zorbax 300SB-C18 HPLC column (150 mm x 100 µm, 3.5 µm; Agilent Technologies, Australia). Linear gradients of 2% - 60% solvent B over 20 min at 500 nL/min flow rate, followed by a hold over 2 min and a steeper

gradient from 60% - 80% solvent B over 1 min were used for protein elution. Solvent B was held at 80% for 3 min to wash the column before returning to 2% solvent B for re-equilibration (15 min) prior to next sample injection. Solvent A consisted of 0.1% formic acid (aq) and solvent B contained 90:10 acetonitrile : 0.1% formic acid (aq). The ionspray voltage was set to 2000 V, declustering potential (DP) 100 V, curtain gas flow 25, nebuliser gas 1 12, gas 2 to 0. The mass spectrometer acquired 1 s full scan TOF-MS data over the mass range 500-2000. The data was acquired and processed using Analyst TF 1.6 software (SCIEX).

Peptide sequencing

The in-gel digested albumins were analyzed by LC-MS/MS on a Shimadzu Prominence Nano HPLC (Rydalmere, Australia) coupled to the 5600 TripleTOF™ mass spectrometer (SCIEX, Toronto, Canada) equipped with a nano electrospray ion source. A volume of 15 µL from each extract was injected onto a 50 mm x 300 µm C18 trap column (Agilent Technologies, Australia) at a rate of 60 µL/min. The samples were de-salted on the trap column for 5 min using 0.1% formic acid (aq) at a rate of 60 µL/min. The trap column was then placed in-line with the Zorbax 300SB-C18 HPLC column (150 mm x 100 µm, 3.5 µm; Agilent Technologies, Australia) for peptide separation. A linear gradient of 2% - 40% solvent B over 7 min at a flow rate of 500 nL/min followed by a steeper gradient from 40% - 80% over 2 min and a hold at 80% for 2 min allowing peptide separation. The column was subsequently washed by ramping the solvent to 98% B over 2 min which was held at 98% for 2 min prior to a return to 2% B for re-equilibration (15 min) prior to next sample injection. Solvent A consisted of 0.1% formic acid (aq) and solvent B contained 90:10 acetonitrile : 0.1% formic acid (aq). The ionspray voltage was set to 2200 V, declustering potential 100 V, curtain gas flow 25, nebuliser gas 1 to 12 and interface heater at 160 °C. The mass spectrometer was set to acquire TOF-MS data over the mass range 300-1800 for 250 ms followed by 20 full scan product ion spectra over the mass range 80 - 1400 with a maximum accumulation time of 100 ms in Information Dependant Acquisition (IDA) mode. The 20 most intense ions observed in the TOF-MS scan exceeding a threshold of 200 counts and a charge state of +2 to +5 were set to trigger the acquisition of product ion MS/MS spectra. The data was acquired and processed using Analyst TF 1.6 software (SCIEX). ProteinPilot™ software 4.0 (SCIEX) with the paragon algorithm 4.0.0.0 was used to identify albumins. MS/MS data were searched against a custom-built database or a Uniprot database (version 2014/05; Helianthus). Specifically the search parameter settings were iodoacetamide modification with cysteine alkylation; trypsin as digestive enzyme.

Results

Genome-based discovery of albumin precursor genes

To enable matching of peptide sequences to gene sequences we first had to fully annotate the draft genome for albumin and confirm their sequences. By searching the draft sunflower genome using the tBLASTn algorithm we rediscovered the seven previously described sunflower albumin precursors (Allen et al. 1987; Elliott et al. 2014; Kortt et al. 1991; Mylne et al. 2011; Thoyts et al. 1996). Furthermore, we identified eighteen new sequences **(Table 1, Supplementary Dataset 1, and Supplementary Fig. 1)**. Each sequence that encoded a conserved ER signal and a highly conserved Cys residue pattern for albumins were designated as a SEED STORAGE ALBUMIN (SESA) with a number.

Further analysis of the nucleotide sequences after cloning the genes helped predict that SESA6 and SESA20 encode two mature albumins like the previously described SESA1 (HaG5), SESA2 (BA), and SESA4 (pHAO). These five pre-proalbumin genes all appear to encode two mature albumins. The *HaG5* sequence in GenBank (GenBank accession: X06410) lacks a conserved Cys in the small subunit of the

first albumin, but the sequence we cloned had the expected number of Cys residues (GenBank accession: KR401266).

PawL2 and PawL3 possessed the characteristic amino acid arrangement of a PawL-type albumin (Elliott et al. 2014), where the predicted small subunit is preceded by an amino acid sequence which lacks Cys residues required to make a stable peptide. *PawL3* had a premature stop codon and appears to be a pseudo gene. Apart from *PawL2* and *PawL3*, no additional PawS-type albumins were identified than those previously described (Elliott et al. 2014). SESA7 possesses ten Cys residues similar to those of the Ara h 6 albumin precursor in peanut (Lehmann et al. 2006). Among the new *SESA* genes identified, most encode a single pre-proalbumin domain similar in structure to albumin precursors that are matured into heterodimeric albumin (e.g. SESA5, SESA7, SESA8, SESA9, SESA10, SESA11, SESA12, and SESA13).

A pairwise comparison was performed to understand how the identified 26 preproalbumin sequences are related to each other (**Fig. 1**). As the sequences vary greatly in length (double vs single albumin) and some sequences were partial, the chosen alignment and algorithm strongly affected the percentage identity that is calculated. Percentage identity between different preproalbumin sequences is generally low (<30% identity) with the main exception being the double preproalbumins where similarity for several ranged from 63-89%. PawS1 and PawS2 are 69% similar and SESA10 and SESA11 are 87% similar. These relationships are summarized in a cladogram (**Fig. 1B**). There are two main clusters; the double preproalbumins and the dimeric preproalbumins including the PawS-type. SESA9 is of the dimeric preproalbumin type, but grouped with double albumin SESA21 in the cladogram probably due to SESA21 being a partial sequence. PawS1 and are PawS2 grouped separately from the PawLs due to the presence of two extra Cys in the PDP region. The grouping of SESA13 close to PawS1 and PawS2 is likely due to the unusual positioning of Cys residues in SESA13 (**Supplementary Data Set 1**) relative to other dimeric albumins. The strong conservation of Cys and rapid evolution of the intervening sequences, insertion-deletion events and, in sunflower, internal duplication to create the double preproalbumins makes it challenging to elucidate their evolutionary history.

Transcript based discovery of seed storage albumin precursor genes

To confirm and quantitate expression of these *SESA* genes we performed high throughput RNA sequencing (RNA-seq) on mRNA from mature sunflower seeds. Mapping raw reads from sunflower seed RNA-seq onto DNA sequences from the draft genome revealed mRNA support for 15 of the predicted albumin-encoding genes (**Fig. 2**). No reads mapped to *SESA5*, *SESA15*, *SESA16* or *SESA19* regions of genome sequence. The reads that mapped to *PawL2, SESA7, SESA8, SESA12, SESA14, SESA16,* and *SESA18* genomic scaffolds were to regions adjacent to predicted ORF (**Fig. 2**) indicating that transcripts for these *SESA* genes are not present in the mRNA pool of dry seeds either.

When comparing the mature dry seed mRNA level expression, generally the double preproalbumins (SESA1, SESA2, SESA4, SESA6, and SESA20) showed the highest expression. A feature that separates the double preproalbumins is they all have introns. By comparison to the double preproalbumins, single preproalbumins (*e.g.* PawS1, PawS2, SESA3, SESA9, SESA10, and SESA13) showed moderate expression at the mRNA level. The expression of *PawL* genes is lower than the double and other single preproalbumins for mature dry seeds (**Fig. 2**).

Proteomic analysis of sunflower seed albumin fraction

Although there are 26 albumin precursor genes, the transcript data indicates mRNA for 15 of them are present in mature seeds. To determine which albumins are able to be found in mature seeds, we

extracted soluble protein from sunflower seeds and enriched for albumin by dialysis against distilled water, followed by size exclusion chromatography to remove high molecular weight proteins. High performance liquid chromatography (HPLC) was further used to separate the albumin-rich fraction and each fraction was visualized by 1D gel electrophoresis. Gel bands from the 1D electrophoresis were manually excised and their protein reduced, alkylated and digested with trypsin for sequencing by tandem mass spectrometry. In this way we could identify mature albumin protein matching eleven of the expressed genes (**Fig. 3** and **Table 2**).

Monomeric SESA3 (SFA8) was found to elute in a wide range of fractions (28 to 40) (**Fig. 3**), which might reflect a variety of oxidation states for its 16 Met residues (Kortt et al. 1991). Further proteomic analysis is needed to confirm this. Although the ProteinPilot database analysis did not predict fraction 33 as SESA3 with 99% confidence, visual inspection of the banding pattern on protein gel (**Fig. 3b** and **c**) suggests that all the fractions from 28 to 40 contain SESA3 or variants thereof.

The double albumins SESA1 (HaG5), SESA20, and SESA2 (BA) elute over a several fractions (**Fig. 3c**) and in these fractions were the top hits (N=1) in ProteinPilot analysis (**Table 2**). Together this suggests these are abundant albumins, but the overlapping elution pattern for most albumins except SESA6 and SESA3 (**Fig. 3 a**) makes it challenging to quantify their relative abundance accurately. Unique peptides corresponding to SESA4, SESA10, SESA11, PawS1, and PawS2 were found in the in-gel digested samples (**Table 2**) indicating that these albumin precursors are also present in mature dry sunflower seeds probably in low levels. PawL1 could not be detected at the protein level in sunflower, but other work detected PawL1-derived mature albumin in seeds of the Asteraceae plant *Arnica montana* (Elliott et al. 2014). The inability to detect sunflower PawL1 protein might be due to low abundance which is consistent with its low mRNA expression (**Fig. 2**).

To identify the major albumins at the protein level more precisely a combination of analytical HPLC and proteomic analysis was applied based on mass spectrometric peak pattern and elution time profile. Known albumin masses (e.g. PawS1, PawS2 and SFA8) correlate to distinct mass spectrometric peak patterns whereas unknown albumins were further reduced, alkylated and digested with trypsin for mass spectrometric analysis. The two albumins encoded by SESA2 (BA), monomeric SESA3, and the second albumin of the newly found SESA20 appeared to dominate the protein profile, whereas the unusual dual-purpose albumins PawS1 and PawS2 are minor components of total albumin (**Fig. 3a** inset: analytical HPLC trace, **Table 3**, **Fig. 4**).

For further relative quantitation of the major sunflower seed albumins (**Fig. 3a** inset: analytical HPLC trace) the ratio between absorbance at A_{280} (A) and extinction coefficient (ε) was calculated (**Table 4**). It is apparent that SESA2-2, SESA3, and SESA20-2 are more abundant when compared to PawS1 and PawS2. However, a higher A/ε ratio was observed for SESA2-1 when compared to other major albumins (**Table 4**). Tryptophan is the predominant contributor at A_{280} whereas tyrosine contributes as little as 5%. SESA2-1 lacks tryptophan and tyrosine residues, which might be the reason for high A/ε ratio observed.

Previous work on SESA1 (HaG5) by Allen *et al.* (1987) and SESA4 (pHAO) by Thoyts *et al.* (1996) suggest that their pro-proteins each appeared to be processed into two mature albumins. Similarly a double albumin found in castor bean was suggested to be matured into two hetero-dimeric albumins (Irwin et al. 1990). However, further analysis is needed to confirm whether sunflower SESA2 (BA) and SESA20 are each processed into two hetero-dimeric albumins or whether the mature albumins are monomeric like SESA3 (SFA8).

A sequencing analysis of the albumin-rich fraction by tandem mass spectrometry indicates that instead of equal expression at the protein level, one albumin encoded from certain transcripts appear to dominate in the protein pool. For example as shown in **Fig. 3a** and **Fig. 3c** SESA1-2, SESA2-2, SESA6-1, and SESA20-2 are readily detected in protein gels, whereas the other albumin present only in minor levels or not detectable at all (**Fig. 3a**). Further proteomic analysis is essential for accurate protein level measurements. Previous work on SESA4 (pHAO), Thoyts *et al.* (1996) suggest that the two potential albumins encoded by the single precursor differ in the extraction behavior. They suggest that only one albumin out of the pair is present in the albumin fraction extracted using the classical methods and the second albumin gets bound to oil bodies once released from the protein bodies. It is possible our albumin extraction protocol has similarly favoured purification of some albumins over others.

PawL3, SESA9, and *SESA13* are expressed at the mRNA level, but were not detected at the protein level. Proteomic analysis of the albumin fraction however, identified an extra mature albumin (SESA21) for which no sequence was detected in the genomic and transcriptomic data. The tryptic fragments for SESA21 match a partial protein sequence (126 amino acids) found in the UniProt database (UniProt ID: Q8RW54). This 126 amino acid sequence was fragment was 96% identical to the cloned SESA1 (HaG5), but had five amino acid differences. These data suggest that SESA21 is a novel albumin (**Table 2** and **Fig. 3c**), but searching the mature dry seed *de novo* transcriptome with this partial fragment could not identify the full length sequence. This might be due to high similarity to SESA1 (HaG5) which causes the *de novo* assembly process to fail to differentiate between the two transcripts.

Discussion

Relative to their importance as a major protein source for humans, seed proteins receive scant attention. The 2S albumins or napins constitute a major portion of seed storage proteins in dicots (Shewry et al. 1995), are highly stable and despite low sequence conservation share a five-helical fold and disulfide connectivity (Mylne et al. 2014). Despite retention of albumin genes and their structural conservation, the study of seed protein evolution is hampered by their rapid evolution and multi-genic families. In an attempt to resolve some of this complexity, we have catalogued the albumin content of sunflower seeds using a draft genome coupled with transcriptomic data and a proteomics analysis of the albumin fraction.

We identified at least 26 genes in sunflower with the potential to encode albumin(s). Of them we could map mRNA reads to 15 of these genes, although some were very weakly represented in the mRNA. A mature seed protein extract enriched for albumin by solubility in water and size exclusion chromatography, was separated further by HPLC and each fraction analyzed by sodium dodecyl sulfate polyacrylamide gel electrophoresis and MS/MS. This provided protein evidence for eleven albumin precursor genes. Further separation using analytical HPLC suggested that three albumins (*i.e.* SESA3/SFA8 and its modified analogues, SESA2/BA, and SESA20) dominate the protein profile. However it's worth noting that these albumins share high amino acid level sequence similarity, resulting in only one or two unique tryptic fragments to use for unambiguous identification. This makes further analysis by bottom-up proteomic approaches challenging.

We observed a correlation between protein level and the presence of introns. Although most of the 26 albumin-encoding genes in sunflower were intronless, the genes encoding the most abundant proteins possessed an intron. Intronic genes are common in eukaryotes. Introns are thought to help suppress gene silencing. This hypothesis has been tested in *Arabidopsis* by introducing an intron into an intronless transgenic reporter system. The introduction of the intron reduced transgene silencing more than four-fold (Christie et al. 2011). Similarly, placement of the sixth intron of the human triose

phosphate isomerase gene in two different places of the *Renilla* luciferase ORF, enhanced mRNA accumulation and enabled higher protein expression than the intronless version (Nott et al. 2003).

Despite the abundance of albumin and their ubiquity, an understudied area is what properties make a 'good' albumin? Specifically, what sequences and structures are favoured by evolution? All albumins studied structurally to date exhibit a similar five-helical structure and highly conserved disulfide bond connectivity (Mylne et al. 2014), but little else is known.

According to the literature, few groups have attempted to study the structure and evolution of seed storage proteins. Kreis and Shewry (1989) studied amino acid sequences in detail, suggesting that barley, maize, and rye prolamins evolved from a single ancestral gene. By studying the DNA and amino acid sequences of Brassicaceae members Boutilier *et al.* (1999) suggested that gene duplication and non-reciprocal DNA exchanges have influenced the evolution of 2S albumins. Similarly, it was suggested that gene duplication and divergence played a key role in the evolution of napin-encoding genes in radish and related crucifers (Raynal et al. 1991).

The rapid evolution and multi-gene nature of seed storage proteins makes it difficult to apply quantitative methods to their molecular evolution. In particular the high rates of insertion and deletion prevent the usual methods of assigning amino acids positive, neutral or negative selection values.

Here we show that although 26 genes for albumins exist in sunflower, only a few contribute to the mature albumin pool and in a storage context these few can be said to be 'important'. SESA3 (SFA8) is one of the most abundant albumins, despite flaunting features conserved in most albumins, such as being unusually methionine-rich and monomeric instead of hetero-dimeric. SESA2 and SESA20 are also abundant, but deviate from typical albumins, as they encode two albumins instead of one. Comparing the properties of important albumins like SESA2, SESA3, and SESA20 to those that are minor contributors to the storage pool might provide insights into the properties that evolution favours for protein storage. It is interesting that the double albumins (SESA1, SESA2, SESA4, SESA6, and SESA20) which are rare in other plant species are common in sunflower, where they represent 5 out of its 26 albumin precursor genes.

An interesting finding was that the two bifunctional albumin genes from sunflower that also make macrocyclic peptides (*PawS1* and *PawS2*) make a very minor contribution to the pool of mature seed albumin. A molecular evolutionary analysis of the *PawS1* gene within the Asteraceae (Elliott et al. 2014) could not use the traditional positive/negative selection analysis tools as these are confounded by the frequent insertions and deletions in genes encoding albumin precursor proteins (preproalbumins). Instead the analysis relied on the frequency of insertion and deletion to calculate the rates of evolution for the small peptide region and the adjacent albumin. Despite many biochemical constraints upon the peptide's processing the peptide region evolved over twice as fast as the adjacent albumin (Elliott et al. 2014). Here we show that mature PawS1 and PawS2 albumin are minor components of seed albumin, suggesting that this functional redundancy for storage provides the freedom for their buried peptides to evolve as rapidly as they do.

Gene redundancy usually implies functional redundancy, but as this work shows the number of genes for a particular protein and the number of gene that truly contribute to the encoded protein function may not be equivalent. Of the 26 genes for albumin in sunflower we found 15 are expressed at the mRNA level and only eleven are detectable in the protein profile which dominated by the products of three genes. By combining genomic, transcriptomic and proteomic approaches we have catalogued the

albumin content of sunflower in terms of the number of genes and their relative contributions to the final albumin content of seeds. This work provides a first step towards a better understanding of albumin evolution and identifies the major genetic loci responsible for sunflower seed albumin.

Accession codes

Helianthus annuus napin-type albumin-encoding genes *PawL2, PawL3, SESA1, SESA2, SESA4, SESA6, SESA7, SESA8, SESA9, SESA10, SESA12, SESA13,* and *SESA20* have been deposited in GenBank under the accession codes KR401276, KR401277, KR401266, KR401267, KR401278, KR401268, KR401270, KR401271, KR401272, KR401273, KR401274, KR401275, and KR401269 respectively.

Conflict of interest

The authors declare that they have no conflicting interests.

References

Allen RD, Cohen EA, Vonder Haar RA, Adams CA, Ma DP, Nessler CL, Thomas TL (1987) Sequence and expression of a gene encoding an albumin storage protein in sunflower. Molec Gen Genet 210:211-218

Blundy KS, Blundy MAC, Crouch ML (1991) Differential expression of members of the napin storage protein gene family during embryogenesis in *Brassica napus*. Plant Mol Biol 17:1099-1104

Boutilier K, Hattori J, Baum BR, Miki BL (1999) Evolution of 2S albumin seed storage protein genes in the Brassicaceae. Biochem Syst Ecol 27:223-234

Chibani K, Ali-Rachedi S, Job C, Job C, Jullien M, Grappin P (2006) Proteomic analysis of seed dormancy in *Arabidopsis*. Plant Physiol 142:1493–1510

Christie M, Croft LJ, Carroll BJ (2011) Intron splicing suppresses RNA silencing in *Arabidopsis*. Plant J 68:159-167

Chu Y et al. (2008) Reduction of IgE binding and nonpromotion of *Aspergillus flavus* fungal growth by simultaneously silencing Ara h 2 and Ara h 6 in peanut. J Agric Food Chem 56:11225-11233

Elliott AG et al. (2014) Evolutionary origins of a bioactive peptide buried within preproalbumin. Plant Cell 26:981-995

Ericson ML, Rödin J, Lenman M, Glimelius K, Josefsson LG, Rask L (1986) Structure of the rapeseed 1.7 S storage protein, napin, and its precursor. J Biol Chem 261:14576-14581

Freire JEC et al. (2015) *Mo*-CBP$_3$, an Antifungal chitin-binding protein from *Moringa oleifera* seeds, is a member of the 2S albumin family. PLoS ONE 10:e0119871

Gallardo K, Job C, Groot SPC, Puype M, Demol H, Vandekerckhove J, Job D (2001) Proteomic analysis of *Arabidopsis* seed germination and priming. Plant Physiol 126:835-848

Hummel M, Wigger T, Brockmeyer J (2015) Characterization of mustard 2S albumin allergens by bottom-up, middle-down, and top-down proteomics: A consensus set of isoforms of Sin a 1. J Proteome Res 14:1547-1556

Irwin SD, Keen JN, Findlay JBC, Lord JM (1990) The *Ricinus communis* 2S albumin precursor: A single preproprotein may be processed into two different heterodimeric storage proteins. Mol Gen Genet 222:400-408

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

Jiang C, Cheng Z, Zhang C, Yu T, Zhong Q, Shen J, Huang X (2014) Proteomic analysis of seed storage proteins in wild rice species of the *Oryza* genus. Proteome Sci 12:51

Kortt AA, Caldwell JB (1990) Low molecular weight albumins from sunflower seed: identification of a methionine-rich albumin. Phytochemistry 29:2805-2810

Kortt AA, Caldwell JB, Lilley GG, Higgins TJV (1991) Amino acid and cDNA sequences of a methionine-rich 2S protein from sunflower seed (*Helianthus annuus* L.). Eur J Biochem 195:329-334
Krebbers E et al. (1988) Determination of the processing sites of an *Arabidopsis* 2S albumin and characterization of the complete gene family. Plant Physiol 87:859-866
Kreis M, Shewry PR (1989) Unusual features of cereal seed protein structure and evolution. BioEssays 10:201-207
Lehmann K et al. (2006) Structure and stability of 2S albumin-type peanut allergens: implications for the severity of peanut allergic reactions. Biochem J 395:463-472
Maria-Neto S et al. (2011) Bactericidal activity identified in 2S albumin from sesame seeds and in silico studies of structure–function relations. Protein J 30:340-350
Meinke D, Koornneef M (1997) Community standards for *Arabidopsis* genetics. Plant J 12:247-253
Menéndez-Arias L, Moneo I, DomÍNguez J, RodrÍGuez R (1988) Primary structure of the major allergen of yellow mustard (*Sinapis alba* L.) seed, *Sin a* I. Eur J Biochem 177:159-166
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*). Biochim Biophys Acta 1698:175-186
Moro CF et al. (2015) Unraveling the seed endosperm proteome of the lotus (*Nelumbo nucifera* Gaertn.) utilizing 1DE and 2DE separation in conjunction with tandem mass spectrometry. Proteomics 00:1-19
Mylne JS et al. (2011) Albumins and their processing machinery are hijacked for cyclic peptides in sunflower. Nat Chem Biol 7:257-259
Mylne JS, Hara-Nishimura I, Rosengren KJ (2014) Seed storage albumins: biosynthesis, trafficking and structures. Funct Plant Biol 41:671-677
Nordlee JA, Taylor SL, Townsend JA, Thomas LA, Bush RK (1996) Identification of a Brazil-Nut allergen in transgenic soybeans. N Engl J Med 334:688-692
Nott A, Meislin SH, Moore MJ (2003) A quantitative analysis of intron effects on mammalian gene expression. RNA 9:607-617
Rajjou L, Lovigny Y, Groot SPC, Belghazi M, Job C, Job D (2008) Proteome-wide characterization of seed aging in *Arabidopsis*: A comparison between artificial and natural aging protocols. Plant Physiol 148:620-641
Raynal M, Depigny D, Grellet F, Delseny M (1991) Characterization and evolution of napin-encoding genes in radish and related crucifers. Gene 99:77-86
Ribeiro SFF et al. (2012) Antifungal and other biological activities of two 2S albumin-homologous proteins against pathogenic fungi. Protein J 31:59-67
Scofield SR, Crouch ML (1987) Nucleotide sequence of a member of the napin storage protein family from *Brassica napus*. J Biol Chem 262:12202-12208
Shewry PR, Halford NG (2002) Cereal seed storage proteins: structures, properties and role in grain utilization. J Exp Bot 53:947-958
Shewry PR, Napier JA, Tatham AS (1995) Seed storage proteins: structures and biosynthesis. Plant Cell 7:945-956
Terras FR et al. (1992) Analysis of two novel classes of plant antifungal proteins from radish (*Raphanus sativus* L.) seeds. J Biol Chem 267:15301-15309
Thorpe SC, Kemeny DM, Panzani RC, McGurl B, Lord M (1988) Allergy to castor bean: II. Identification of the major allergens in castor bean seeds. J Allergy Clin Immunol 82:67-72
Thoyts PJE, Napier JA, Millichip M, Stobart AK, Griffiths WT, Tatham AS, Shewry PR (1996) Characterization of a sunflower seed albumin which associates with oil bodies. Plant Sci 118:119-125

van der Klei H, Damme JV, Casteels P, Krebbers E (1993) A fifth 2S albumin isoform is present in *Arabidopsis thaliana*. Plant Physiol 101:1415-1416
Wang K et al. (2010) Characterization of seed proteome in *Brachypodium distachyon*. J Cereal Sci 52:177-186
Wong P-F, Abubakar S (2005) Post-germination changes in *Hevea brasiliensis* seeds proteome. Plant Sci 169:303-311
Zhou Y et al. (2013) Peanut allergy, allergen composition, and methods of reducing allergenicity: A review. Int J Food Sci Technol 2013:1-8

Tables

Table 1. Details of all sunflower albumins

Evidence for presence and absence of each albumin at the gene, mRNA, and protein level is marked with either + or - marks respectively. The structure was predicted by observing the Cys-residue pattern.

Gene name	GenBank ID	Gene	mRNA	Protein	Structure (or the predicted structure)
PawS1	FJ749265	+	+	+	Hetero-dimeric
PawS2	FJ469149	+	+	+	Hetero-dimeric
PawL1	KF574811	+	+	-	Hetero-dimeric
PawL2	KR401276	+	-	-	Hetero-dimeric
PawL3	KR401277	+	+	-	Premature stop codon
SESA1(*HaG5*)	X06410 & KR401266	+	+	+	Double albumin
SESA2 (*BA*)	AJ275962 & KR401267	+	+	+	Double albumin
SESA3 (*SFA8*)	X56686	+	+	+	Monomeric
SESA4 (*pHAO*)	KR401278	+	+	+	Double albumin
SESA5	N/A	+	-	-	Partial ORF
SESA6	KR401268	+	+	+	Double albumin
SESA7	KR401270	+	-	-	Hetero-dimeric
SESA8	KR401271	+	-	-	Hetero-dimeric
SESA9	KR401272	+	+	-	Hetero-dimeric
SESA10	KR401273	+	+	+	Hetero-dimeric
SESA11	N/A	+	+	+	Partial ORF
SESA12	KR401274	+	-	-	Hetero-dimeric
SESA13	KR401275	+	+	-	Hetero-dimeric
SESA14	N/A	+	-	-	Partial ORF
SESA15	N/A	+	-	-	Partial ORF
SESA16	N/A	+	-	-	Partial ORF
SESA17	N/A	+	-	-	Partial ORF
SESA18	N/A	+	-	-	Partial ORF
SESA19	N/A	+	-	-	Partial ORF
SESA20	KR401269	+	+	+	Double albumin
SESA21	UniProt: Q8RW54	-	+	+	Partial ORF

Table 2. Peptides detected in the in-gel digested samples

This table lists selected peptide fragments with a confidence level >95% as detected by the ProteinPilot software in each in-gel digested sample. N is the ProteinPilot database hit. Percentage coverage means the number of amino acids in all detected peptides (irrespective of confidence level) / total number of amino acids in protein sequence. Double underlined peptides are unique to the corresponding albumin. The top hit (N=1) for each fraction/band is shown in bold text. If the number of >95% peptides are less than two (*e.g.* fraction 29, 32), facts like the band size, elution time *etc.* were observed when assigning it as the corresponding SESA.

Fraction No.	N	Coverage (%)	No. of peptides (>95%)	Chosen detected peptides (>95% confidence)	Representing albumin
10	**1**	**25.9**	**5**	**NELQNVDKECQCEAVKK,GGGDYGSQEIQQLK**	**SESA6**
12	**1**	**23.4**	**3**	**ETEIQRPVGECQR,GLQQQCCNELQNVKR**	**SESA1 (HaG5)**
12	2	34.9	3	GLQQQCCNELQNVKR,ECQCEAIQEVAR	SESA21
13	**1**	**35.5**	**28**	**AQILPNVCNLQSR,IDIPFRDRPFGTGSQQCR,ETEIQRPVGECQR,GLQQQCCNELQNVKR,RVIQNLPNQCDLEVQQCNIPYG,FVEQQMQQSPR,RGLQQQCCNELQNVKR**	**SESA1 (HaG5)**
13	2	34.1	23	FVEQQMQQSPR,RGLQQQCCNELQNVKR,ECQCEAIQEVARR	SESA21
13	3	34	7	GQQHQQQQHQQQEQQLLQQCCQELQNIEGQCQCEAVK,MPFQGSQQSQQLK,AQILPNVCNLQSR	SESA4 (pHAO)
13	4	31	6	AQILPNVCNLQSR,IDIPFRDRPFGTGSQQCR,RPGQQQQEPELQQCCNELQNVKR	SESA20
13	5	26	3	AQILPNVCNLQSR,SQQCSETEIQRPVSQCQR,RVIQNLPNQCDLEVQQCNIPY	SESA2 (BA)
14 band 1	**1**	**55.2**	**37**	**QVFREAQQQVQQQQGR,AQILPNVCNLQSR,IDIPFRDRPFGTGSQQCRETEIQRPVGECQR,FVEQQMQQSPR,GLQQQCCNELQNVKRECHCEAIQEVAR,RVIQNLPNQCDLEVQQCNIPYGM**	**SESA1 (HaG5)**
14 band 1	2	74.6	24	FVEQQMQQSPR,GLQQQCCNELQNVKR,ECQCEAIQEVARR,	SESA21
14 band 1	3	44.4	8	GQQHQQQQHQQQEQQLLQQCCQELQNIEGQCQCEAVK,QQQQRPRMPFQGSQQSQQLK,MPFQGSQQSQQLKQR,AQILPNVCNLQSR	SESA4 (pHAO)
14 band 1	4	44.6	6	PLSEQRQCSQQVQGQRLNQCR,QVFREAQQQVQQQQGR,AQILPNVCNLQSR,SQQCSETEIQRPVSQCQR,RVIQNLPNQCDLEVQQCNIPY	SESA2 (BA)
14 band 2	**1**	**36.1**	**6**	**IDIPFRDRPFGTGSQQCR,GLQQQCCNELQNVKR,RVIQNLPNQCDLEVQQCNIPYGM**	**SESA1 (HaG5)**
14 band 2	2	22.2	3	GLQQQCCNELQNVKRECQCEAIQEVAR	SESA21
15	**1**	**40.8**	**19**	**AQILPNVCNLQSR,ETEIQRPVGECQR,FVEQQMQQSPR,GLQQQCCNELQNVKR,ECHCEAIQEVAR**	**SESA1 (HaG5)**
15	2	44.4	20	FVEQQMQQSPR,GLQQQCCNELQNVKR,ECQCEAIQEVARR	SESA21
15	3	29.5	4	AQILPNVCNLQSR,SQQCSETEIQRPVSQCQR,ECQCEAVQEVAR	SESA2 (BA)
15	4	17.9	3	MPFQGSQQSQQLK,AQILPNVCNLQSR	SESA4 (pHAO)
16 band 1	**1**	**50.8**	**49**	**MFLQQGQNIPR,AQILPNVCNLQSR,IDIPFRDRPFGTGSQQCR,ETEIQRPVGECQR,FVEQQMQQSPR,GLQQQCCNELQNVKR,ECHCEAIQEVARR,RGQFGGQEMETAR,RVIQNLPNQCDLEVQQCNIPYG**	**SESA1 (HaG5)**
16 band 1	2	47.9	27	MFLQQGQNIPR,AQILPNVCNLQSR,IDIPFRDRPFGTGSQQCR,ETEIQRPVSQCQR,YVEQQMQSPMPYIR,RPGQQQQEPELQQCCNELQNVKR,ECQCEAVQEVAR,RGQFGGQEMDIAR,RVIQNLPNQCSLQVQQCNIPY	SESA20
16 band 1	3	70.6	29	DRPFGTGSQQCR,FVEQQMQQSPR,GLQQQCCNELQNVKR,ECQCEAIQEVARR,RGQFGGQEMETAR	SESA21
16 band 1	4	42.5	16	AQILPNVCNLQSR,SQQCSETEIQRPVSQCQR,YVEQQMQSPMPYIR,ECQCEAVQEVAR,RGQFGGQEMDIAR,RVIQNLPNQCDLEVQQCNIPY	SESA2 (BA)
16 band 1	5	20.4	3	MPFQGSQQSQQLK,AQILPNVCNLQSR,QSEIQRPVSQCQR	SESA4 (pHAO)
16 band 2	**1**	**34.5**	**11**	**MFLQQGQNIPR,AQILPNVCNLQSR,IDIPFRDRPFGTGSQQCR,GLQQQCCNELQNVKR,ECHCEAIQEVAR,RVIQNLPNQCDLEVQQCNIPYGM**	**SESA1 (HaG5)**
16 band 2	2	35.5	7	MFLQQGQNIPR,AQILPNVCNLQSR,IDIPFRDRPFGTGSQQCR,RPGQQQQEPELQQCCN,ECQCEAVQEVAR,GQFGGQEMDIAR,RVIQNLPNQCDLEVQQCNIPY	SESA20

16 band 2	3	34.4	7	AQILPNVCNLQSR,SQQCSETEIQRPVSQCQR,RPGQQQQEPELQQCCN,ECQCEAVQEVAR,GQFGGQEMDIAR,RVIQNLPNQCDLEVQQCNIPY	SESA2 (BA)
17 band 1	**1**	**60.2**	**28**	**QVFREAQQQVQQQQGR,AQILPNVCNLQSR,IDIPFRDRPFGTGSQQCR,ETEIQRPVGECQR,FVEQQMQQSPR,GLQQQCCNELQNVKR,ECHCEAIQEVARR,RVIQNLPNQCDLEVQQCNIPYG**	**SESA1 (HaG5)**
17 band 1	2	48.3	25	AQILPNVCNLQSR,IDIPFRDRPFGTGSQQCR,ETEIQRPVSQCQR,YVEQQMQSPMPYIR,RPGQQQQEPELQQCCNELQNVKR,ECQCEAVQEVARR,RGQFGGQEMDIAR	SESA20
17 band 1	3	74.6	7	FVEQQMQQSPR,GLQQQCCNELQNVKR,ECQCEAIQEVAR	SESA21
17 band 1	4	56.8	20	QVFREAQQQVQQQQGR,AQILPNVCNLQSR,SQQCSETEIQRPVSQCQR,YVEQQMQSPMPYIR,ECQCEAVQEVARR,RGQFGGQEMDIAR,RVIQNLPNQCDLEVQQCNIPY	SESA2 (BA)
17 band 2	**1**	**25.8**	**3**	**IDIPFRDRPFGTGSQQCR,GLQQQCCNELQNVKR,RVIQNLPNQCDLEVQQCNIPYG**	**SESA1 (HaG5)**
17 band 2	2	14.1	2	IDIPFRDRPFGTGSQQCR,ECQCEAVQEVAR	SESA20
18 band 1	**1**	**43.8**	**46**	**IDIPFRDRPFGTGSQQCR,ETEIQRPVSQCQR,YVEQQMQSPMPYIR,RPGQQQQEPELQQCCNELQNVKR,ECQCEAVQEVARR,RGQFGGQEMDIAR,RVIQNLPNQCSLQVQQCNIPY**	**SESA20**
18 band 1	2	38.5	11	QVFREAQQQVQQQQGR,QSVPFRGSQQTQQLKQK,IDIPFRDRPFGTGSQQCR,GLQQQCCNELQNVKRECHCEAIQEVAR	SESA1 (HaG5)
18 band 1	3	52.3	23	QVFREAQQQVQQQQGR,SQQCSETEIQRPVSQCQR,YVEQQMQSPMPYIR,ECQCEAVQEVARR,RGQFGGQEMDIARR	SESA2 (BA)
18 band 1	4	38.1	4	GLQQQCCNELQNVKR,ECQCEAIQEVAR	SESA21
18 band 2	**1**	**54.8**	**9**	**IDIPFRDRPFGTGSQQCR,YVEQQMQSPMPY,RPGQQQQEPELQQCCNELQNVK,ECQCEAVQEVAR,GQFGGQEMDIAR,RVIQNLPNQC**	**SESA20**
18 band 3	**1**	**47.9**	**10**	**MFLQQGQNIPR,AQILPNVCNLQSR,YVEQQMQSPMPYIR,RPGQQQQEPELQQCCNELQNVKR,ECQCEAVQEVAR,RGQFGGQEMDIAR**	**SESA20**
18 band 3	2	40.1	5	MFLQQGQNIPR,QVFREAQQQVQQQQGR,AQILPNVCNLQSR,GLQQQCCNELQNVKR,ECHCEAIQEVAR	SESA1 (HaG5)
18 band 3	3	41.8	8	QVFREAQQQVQQQQGR,AQILPNVCNLQSR,SQQCSETEIQRPVSQCQR,YVEQQMQSPMPYIR,ECQCEAVQEVAR,RGQFGGQEMDIAR	SESA2 (BA)
19 band 1	**1**	**34.8**	**17**	**MFLQQGQNIPR,AQILPNVCNLQSR,ETEIQRPVSQCQR,YVEQQMQSPMPYIR,RPGQQQQEPELQQCCNELQNVKR,ECQCEAVQEVAR,RGQFGGQEMDIAR**	**SESA20**
19 band 1	2	40.4	18	AQILPNVCNLQSR,SQQCSETEIQRPVSQCQR,YVEQQMQSPMPYIR,RPGQQQQEPELQQCCNQLQNVNR,ECQCEAVQEVAR,RGQFGGQEMDIAR	SESA2 (BA)
19 band 1	3	28.8	5	MFLQQGQNIPR,AQILPNVCNLQSR,ETEIQRPVGECQR,GLQQQCCNELQNVKR,ECHCEAIQEVAR	SESA1 (HaG5)
19 band 2	**1**	**46**	**31**	**AQILPNVCNLQSR,IDIPFR,SQQCSETEIQRPVSQCQR,YVEQQMQSPMPY,RPGQQQQEPELQQCCNQLQNVNR,ECQCEAVQEVARR,RGQFGGQEMDIAR,RVIQNLPNQCDLEVQQC**	**SESA2 (BA)**
19 band 2	2	43.8	35	MFLQQGQNIPR,AQILPNVCNLQSR,IDIPFR,ETEIQRPVSQCQR,YVEQQMQSPMPY,RPGQQQQEPELQQCCNELQNVKR,ECQCEAVQEVARR,RGQFGGQEMDIAR,RVIQNLPNQCSLQVQQC	SESA20
19 band 2	3	26.8	8	MFLQQGQNIPR,AQILPNVCNLQSR,IDIPFR,GLQQQCCNELQNVK,RVIQNLPNQCDLEVQQC	SESA1 (HaG5)
20 band 1	**1**	**47.7**	**38**	**AQILPNVCNLQSR,SQQCSETEIQRPVSQCQR,YVEQQMQSPMPYIR,RPGQQQQEPELQQCCNQLQNVNR,ECQCEAVQEVARR,RGQFGGQEMDIAR,RVIQNLPNQCDLEVQQCNIPY**	**SESA2 (BA)**
20 band 1	2	43.8	31	MFLQQGQNIPR,AQILPNVCNLQSR,ETEIQRPVSQCQR,YVEQQMQSPMPYIR,RPGQQQQEPELQQCCNELQNVKR,ECQCEAVQEVARR,RGQFGGQEMDIAR	SESA20
20 band 1	3	22.1	6	MFLQQGQNIPR,AQILPNVCNLQSR,GLQQQCCNELQNVKR,RVIQNLPNQCDLEVQQCNIPY	SESA1 (HaG5)
20 band 1	4	25	2	QLYQEALQMVK,QQQSVPIFGSQR	SESA11
20 band 1	5	31.3	2	ELQNVDEKCQCEAVK,QQQSVPIFGSQR	SESA10
20 band 2	**1**	**13.7**	**0**	**Peptides with >95% confidence level were not detected.**	**SESA2 (BA)**
20 band 3	**1**	**36.5**	**8**	**SQQCSETEIQRPVSQCQR,RPGQQQQEPELQQCCN,ECQCEAVQEVAR,GQFGGQEMDIAR**	**SESA2 (BA)**
20 band 3	2	24.4	2	MFLQQGQNIPR,GLQQQCCNELQNVKR	SESA1 (HaG5)
20 band 3	3	23.2	1	QLQQGQGGQQQVQQMVK	PawS1
21 band 1	**1**	**57.5**	**26**	**QVFREAQQQVQQQQGR,AQILPNVCNLQSR,SQQCSETEIQRPVSQCQR,**	**SESA2 (BA)**

				YVEQQMQSPMPYIR,RPGQQQQEPELQQCCNQLQNVNR,ECQCEAVQEVARR ,RGQFGGQEMDIAR,RVIQNLPNQCDLEVQQCNIPY	
21 band 1	2	58.3	16	PISEQRQCSQQLQGQR,LNECDMYFMK,RPEEVIQQACCK,ELQNVDEKCQC EAVK,QLFQEALQMVK,QQQSVPIFGSQRQK,QRAQILPNVCNFQSK, AQILPNVCNFQSKR	SESA10
21 band 1	3	52.4	19	MFLQQGQNIPR,AQILPNVCNLQSR,IDIPFRDRPFGTGSQQCR,ETEIQRP VSQCQR,YVEQQMQSPMPYIR,RPGQQQQEPELQQCCNELQNVKR,ECQCEA VQEVARR,RGQFGGQEMDIAR	SESA20
21 band 1	4	50.8	9	MFLQQGQNIPR,QVFREAQQQVQQQQGR,AQILPNVCNLQSR,IDIPFRDRP FGTGSQQCR,GLQQQCCNELQNVKR,ECHCEAIQEVAR,RVIQNLPNQCDLE VQQCNIPY	SESA1 (HaG5)
21 band 1	5	30.2	3	GLQQQCCNELQNVKR,ECQCEAIQEVAR	SESA21
21 band 2	**1**	**28.9**	**1**	**IQDKEGIPPDQQR**	**Seed tetra-ubiquitin**
21 band 2	2	19	0	Peptides with >95% confidence level were not detected.	SESA2 (BA)
22 band 1	**1**	**43.5**	**23**	**MFLQQGQNIPR,EAQQQVQQQQGPQSVPFLHSQQAQR,AQILPNVCNLQSR, ETEIQRPVSQCQR,YVEQQMQSPMPYIR,RPGQQQQEPELQQCCNELQNVKR ,ECQCEAVQEVARR,GQFGGQEMDIAR**	**SESA20**
22 band 1	2	53.4	15	LNECDMYFMK,GMTIDEGYSIPMR,RPEEVIQQACCK,ELQNVDEKCQCEAV K,QLFQEALQMVK,QQQSVPIFGSQR,AQILPNVCNFQSKR	SESA10
22 band 1	3	33.3	20	AQILPNVCNLQSR,SQQCSETEIQRPVSQCQR,YVEQQMQSPMPYIR,RPGQ QQQEPELQQCCNQLQNVNR,ECQCEAVQEVARR,GQFGGQEMDIAR	SESA2 (BA)
22 band 1	4	21.4	3	GLQQQCCNELQNVKR,ECQCEAIQEVAR	SESA21
22 band 1	5	22.7	5	MFLQQGQNIPR,AQILPNVCNLQSR,GLQQQCCNELQNVKR,ECHCEAIQEV AR	SESA1 (HaG5)
22 band 2	**1**	**29.1**	**12**	**RSQQCSETEIQRPVSQCQR,YVEQQMQSPMPY,IRRPGQQQQEPELQQCCNQ LQNVNR,ECQCEAVQEVAR,GQFGGQEMDIAR**	**SESA2 (BA)**
22 band 2	2	52.8	17	LNECDMYFMK,GMTIDEGYSIPMR,RPEEVIQQACCK,ELQNVDEKCQCEAV K,QLFQEALQMVK,QQQSVPIFGSQR,AQILPNVCNFQSK	SESA10
22 band 2	3	30.3	8	MFLQQGQNIPR,YVEQQMQSPMPY,IRRPGQQQQEPELQQCCNELQNVKR,E CQCEAVQEVAR,GQFGGQEMDIAR	SESA20
22 band 2	4	11.9	1	QLQQGQGGQQQVQQMVK	PawS1
22 band 2	5	8.8	2	MFLQQGQNIPR,GLQQQCCNELQNVKR	SESA1 (HaG5)
23 band 1	**1**	**37.9**	**14**	**SQQCSETEIQRPVSQCQR,YVEQQMQSPMPY,RPGQQQQEPELQQCCNQ,EC QCEAVQEVAR,GQFGGQEMDIAR,RVIQNLPNQC**	**SESA2 (BA)**
23 band 1	2	31.4	12	YVEQQMQSPMPY,RPGQQQQEPELQQCCNELQNVK,ECQCEAVQEVAR,GQF GGQEMDIAR,RVIQNLPNQC	SESA20
23 band 1	3	24.5	1	QLQQGQGGQQQVQQMVK	PawS1
23 band 1	4	17.4	2	GMTIDEGYSIPMR	SESA10 or SESA11
23 band 2	**1**	**49.3**	**23**	**MFLQQGQNIPR,EAQQQVQQQQGPQSVPFLHSQQAQR,AQILPNVCNLQSR, IDIPFRDRPFGTGSQQCR,ETEIQRPVSQCQR,YVEQQMQSPMPYIR,RPGQ QQQEPELQQCCNELQNVKR,ECQCEAVQEVARR,GQFGGQEMDIAR**	**SESA20**
23 band 2	2	42.1	20	AQILPNVCNLQSR,SQQCSETEIQRPVSQCQR,YVEQQMQSPMPYIR,RPGQ QQQEPELQQCCNQLQNVNR,ECQCEAVQEVARR,GQFGGQEMDIAR	SESA2 (BA)
23 band 2	3	38	7	LNECDMYFMK,ELQNVDEKCQCEAVK,QLFQEALQMVK,QQQSVPIFGSQR, AQILPNVCNFQSKR	SESA10
23 band 2	4	21.4	2	GLQQQCCNELQNVKR,ECQCEAIQEVAR	SESA21
24 band 1	**1**	**40.4**	**23**	**AQILPNVCNLQSR,SQQCSETEIQRPVSQCQR,YVEQQMQSPMPYIR,RPGQ QQQEPELQQCCNQLQNVNR,ECQCEAVQEVARR,RGQFGGQEMDIAR**	**SESA2 (BA)**
24 band 1	2	41.7	14	MFLQQGQNIPR,AQILPNVCNLQSR,YVEQQMQSPMPYIR,RPGQQQQEPEL QQCCNELQNVKR,ECQCEAVQEVARR,RGQFGGQEMDIAR	SESA20
24 band 1	3	21.4	2	GLQQQCCNELQNVKR,ECQCEAIQEVAR	SESA21
24 band 1	4	9.2	1	ELQNVDEKCQCEAVK	SESA10
24 band 2	**1**	**41.1**	**16**	**SQQCSETEIQRPVSQCQR,YVEQQMQSPMPY,IRRPGQQQQEPELQQCCNQL QNVNR,ECQCEAVQEVAR**	**SESA2 (BA)**
24 band 2	2	24.5	10	MFLQQGQNIPR,YVEQQMQSPMPY,RPGQQQQEPELQQCCNELQNVKR,ECQ CEAVQEVAR,GQFGGQEMDIAR	SESA20
24 band 2	3	23.4	2	QLQQGQGGQQQVQQMVK	PawS2
24 band 2	4	23.3	1	ELQNVDEKCQCEAVK	SESA10
26	**1**	**43.5**	**8**	**SQQCSETEIQRPVSQCQR,YVEQQMQSPMPYIR,RPGQQQQEPELQQCCNQL QNVNR,ECQCEAVQEVAR,GQFGGQEMDIAR**	**SESA2 (BA)**

26	2	32.8	6	YVEQQMQSPMPYIR,RPGQQQQEPELQQCCNELQNVKR,ECQCEAVQEVAR, GQFGGQEMDIAR	SESA20
26	3	15.3	1	ELQNVDEKCQCEAVK	SESA10
26	4	19.7	1	GLQQQCCNELQNVKR	SESA1 (HaG5)
27	**1**	**19**	**0**	**Peptides with >95% confidence level were not detected.**	**SESA2 (BA)**
28	**1**	**47.5**	**2**	**MREEDHKQLCCMQLK**	**SESA3 (SFA8)**
29	**1**	**31.2**	**1**	**MREEDHKQLCCMQLK**	**SESA3 (SFA8)**
32	**1**	**36.9**	**1**	**MREEDHKQLCCMQLK**	**SESA3 (SFA8)**
33	**1**	**44**	**0**	**Peptides with >95% confidence level were not detected.**	**SESA3 (SFA8)**
34	**1**	**36.2**	**3**	**MREEDHKQLCCMQLKNLDEK**	**SESA3 (SFA8)**
35	**1**	**41.1**	**0**	**Peptides with >95% confidence level were not detected.**	**SESA3 (SFA8)**
36	**1**	**70.2**	**10**	**GRTESGCYQQMEEAEMLNHCGMYLMK,MREEDHKQLCCMQLKNLDEK, CMCPAIMMMLNEPMWIR,MRDQVMSMAHNLPIECNLMSQPCQM**	**SESA3 (SFA8)**
37 band 1	**1**	**27.7**	**0**	**Peptides with >95% confidence level were not detected.**	**SESA3 (SFA8)**
37 band 2	**1**	**53.9**	**7**	**GRTESGCYQQMEEAEML,AHNLPIECNLMSQPCQM**	**SESA3 (SFA8)**
38	**1**	**54.6**	**1**	**MREEDHKQLCCMQLK**	**SESA3 (SFA8)**
39	**1**	**43.3**	**0**	**Peptides with >95% confidence level were not detected.**	**SESA3 (SFA8)**
40	**1**	**51.8**	**1**	**MREEDHKQLCCMQLK**	**SESA3 (SFA8)**

Table 3. Peptides detected in in-solution digested samples

This table lists the peptide fragments with a confidence level >95% as detected by the ProteinPilot software in in-solution digested analytical HPLC fractions. Double underlined peptides are unique to the corresponding albumin.

N	Coverage (%)	No. of peptides (>95%)	Chosen detected peptides (>95% confidence)	Representing albumin
1	57.5	12	QCSQQVQGQR,MFLQQGQR,GQQHQQQQEQQLLQQCCQELQNIDQQCQCEAVKQVFR,SSQQTQQLK,AQILPNVCNLQSR	SESA2-1 (BA-1)
1	48.4	17	SQQCSETEIQRPVSQCQR,RPGQQQQEPELQQCCNQLQNVNR,ECQCEAVQEVAR,GQFGGQEMDIAR	SESA2-2 (BA-2)
1	57.6	39	DRPFGTGSQQCR,ETEIQRPVSQCQR,YVEQQMQSPMPYIR,RPGQQQQEPELQQCCNELQNVKR,ECQCEAVQEVAR,GQFGGQEMDIAR,RVIQNLPNQCSL,VIQNLPNQCSLQVQQCNIPY	SESA20-2

Table 4. Absorbance (A_{280})/extinction coefficient ratio of the major sunflower seed albumins

* Note that SESA2-1 does not contain any tyrosine and tryptophan residues.

**For SESA3 (SFA8), A/ε ratio was not calculated of any of its analogues.

Albumin	Absorbance A_{280} (μV)	Extinction coefficient (ε)	A/ε
SESA2-1	7963	500*	15.9
SESA2-2	29154	4970	5.9
SESA20-2	35073	4970	7.1
PawS1	9860	1990	5.0
PawS2	9183	1990	4.6
SESA3 (SFA8)**	61363	10470	5.9

Figure Captions

Fig. 1 Pairwise comparison of sunflower preproalbumin sequences
A. Graphical illustration of different preproalbumin types. Predicted ER signal (ER), small albumin subunit (SSU), large albumin subunit (LSU), PawS-derived peptide (PDP in PawS) or the PDP-like (in PawL) region, and spacer regions are marked in rose, green, orange, aqua, and black respectively. Region delimitation is based on the Cys residue pattern and the potential sites for albumin maturation by proteolytic cleavage.
B. A circular cladogram summarising the similarity between different sunflower preproalbumins.
C. Pairwise comparison of preproalbumin sequences. Upper right triangle shows the percentage identities and the lower left triangle shows the differences. Maximum to minimum similarity is indicated by a variation of dark red to dark blue colour.

Fig. 2 mRNA expression of sunflower preproalbumins
Raw reads from RNA-seq of the mature sunflower seed RNA were mapped on to the 3 kb genomic sequence containing preproalbumin genes identified from the sunflower genome. X-axis is the alignment position and Y-axis is the coverage. Coverage is the number of reads contributing to a given position in the mapping. ORF, open reading frame; EX-exon; IN-intron.

Fig. 3 Separation of sunflower albumins by HPLC
(A) First inset shows the albumin-rich fractions from size-separated sunflower seed extract. Lane 2 is the crude extract. Based on size, lanes 6-9 were selected as the albumin-rich fractions.
The albumin-rich fraction was further purified using RP-HPLC. HPLC fractions from 20-50 min are shown with a second inset showing a labelled analytical HPLC trace of the most abundant sunflower albumins.

(B) 1-dimensional gel electrophoresis of equal volumes of HPLC fractions number 9 to 40. Squares indicate where bands were excised from for proteomic analysis. If multiple bands were present, they were excised separately and numbered.

(C) The same trace as B, but with proteomic information from Table 1 overlaid. Only the top scored (N=1) protein from each fraction/band is shown. Additionally, inferences were made based on the banding pattern in fractions 25 to 40.

Fig. 4 MS evidence for SESA3 (SFA8), PawS1, and PawS2
High resolution mass spectra of the albumin fractions corresponding to native SESA3 (SFA8), PawS1, and PawS2 purified using RP-HPLC. M_c is the calculated mass and M_O is the observed mass calculated based on peaks with most intense charge states using protein reconstruction in Analyst TF 1.6.1 (SCIEX).

Supplementary Figure Caption

Supplementary Fig. 1 Alignment of the predicted sunflower albumin sequences
(A) Alignment of dimeric albumins

(B) Alignment of double albumins

(C) Alignment of PawS-type albumins

Sequences were aligned using CLC Genomics Workbench 7.5.1 setting the alignment parameters specifically as gap open cost: 9, gap extension cost: 2 and rendered using BOXSHADE. Based on similarity sequences were ordered by CLC automatically. Predicted ER signal (ER), small albumin subunit (SSU), large albumin subunit (LSU), PawS-derived peptide (PDP in PawS) or the PDP-like (in PawL) region, and spacer regions are marked in rose, green, orange, aqua, and black respectively. Region delimitation was inferred by observing the Cys residue pattern and the potential albumin maturation sites. Red boxes indicate the conserved Cys residues. Partial ORFs are indicated by two lines at the distal ends of the sequence.

Supplementary Data Set 1.
FASTA list of all 26 sunflower preproalbumin sequences (SESA1-SESA21, PawL1-PawL3, PawS1 and PawS2).

Supplementary Data Set 2.
ClustalW alignment of sunflower preproalbumin sequences used to create the pairwise comparison and cladogram in **Fig. 1.**