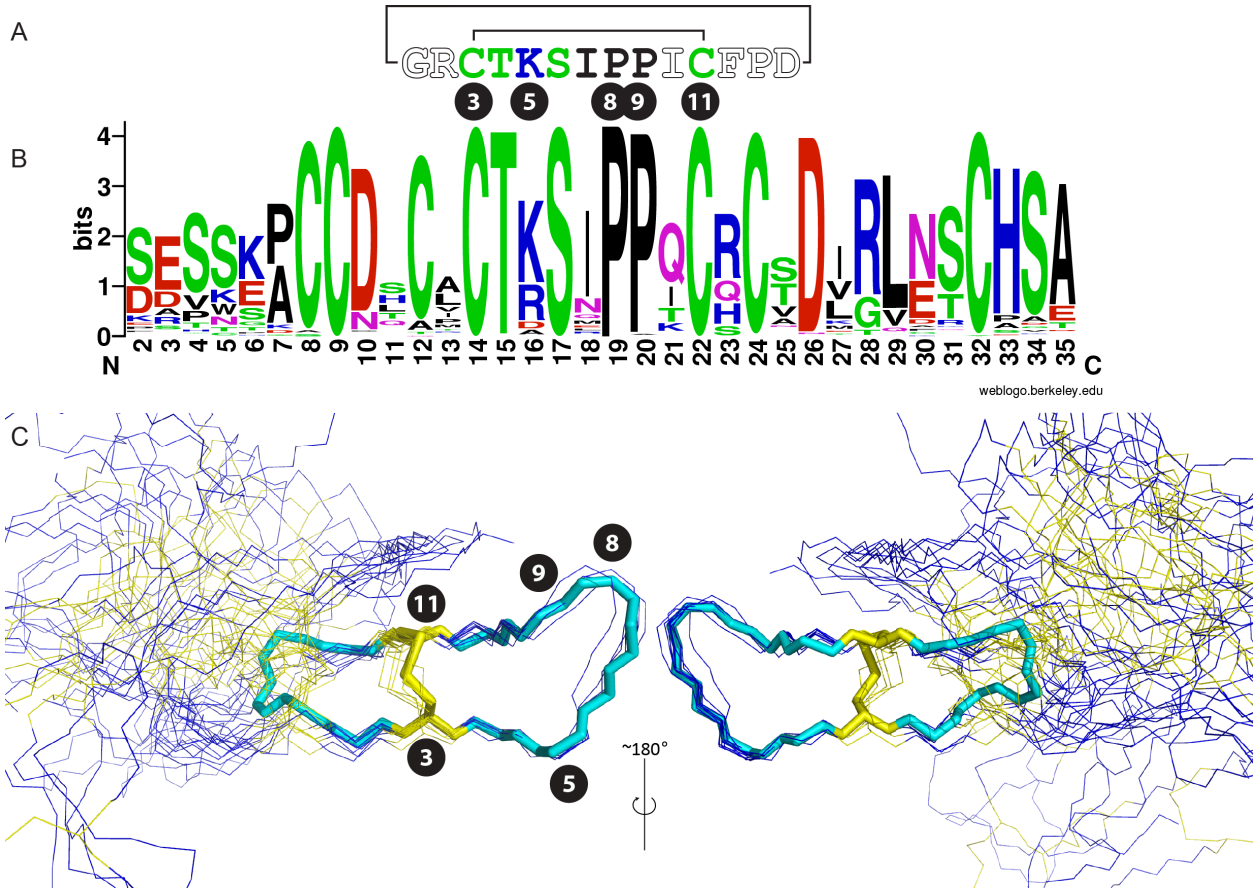
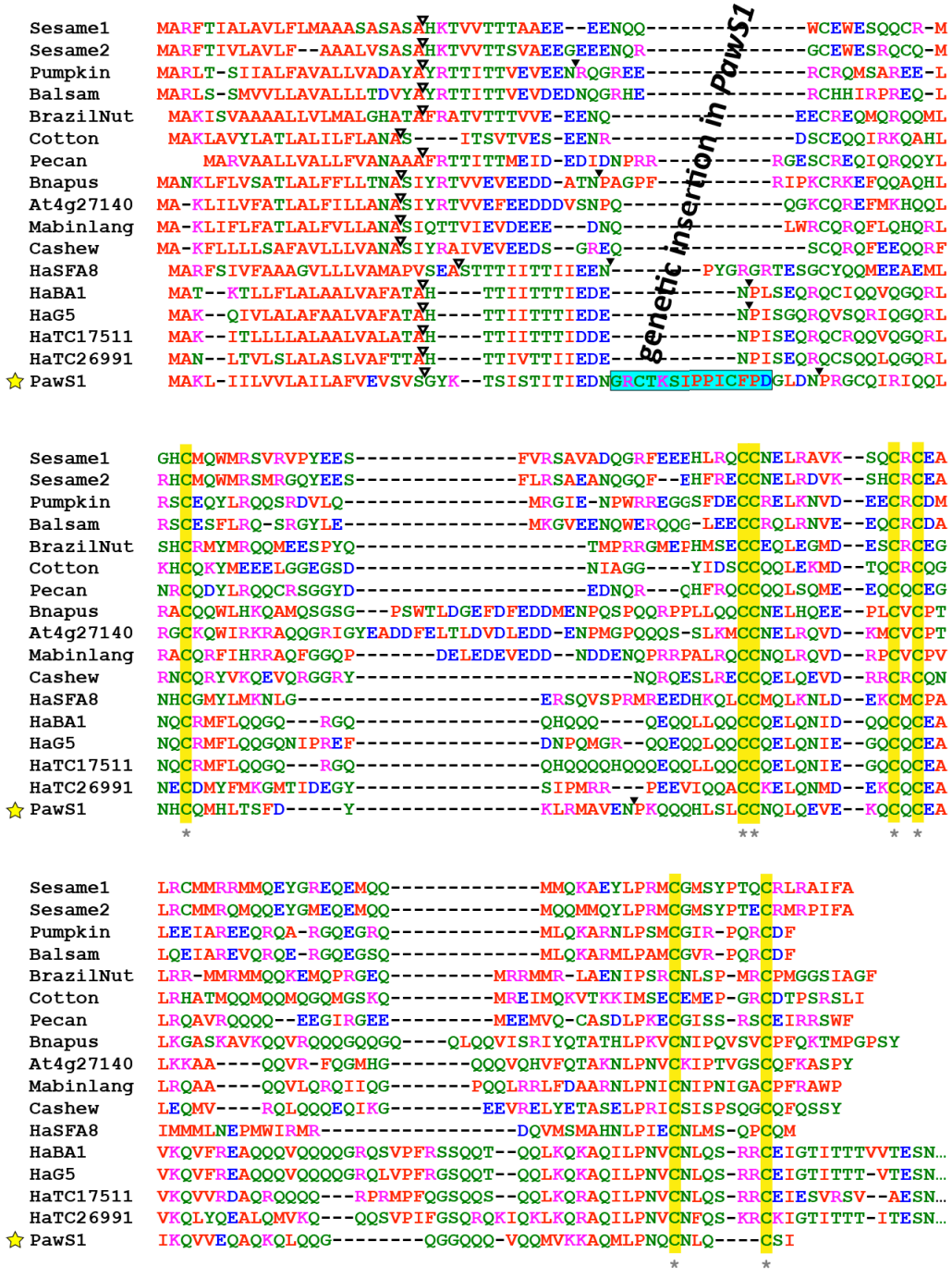
Supplemental Information

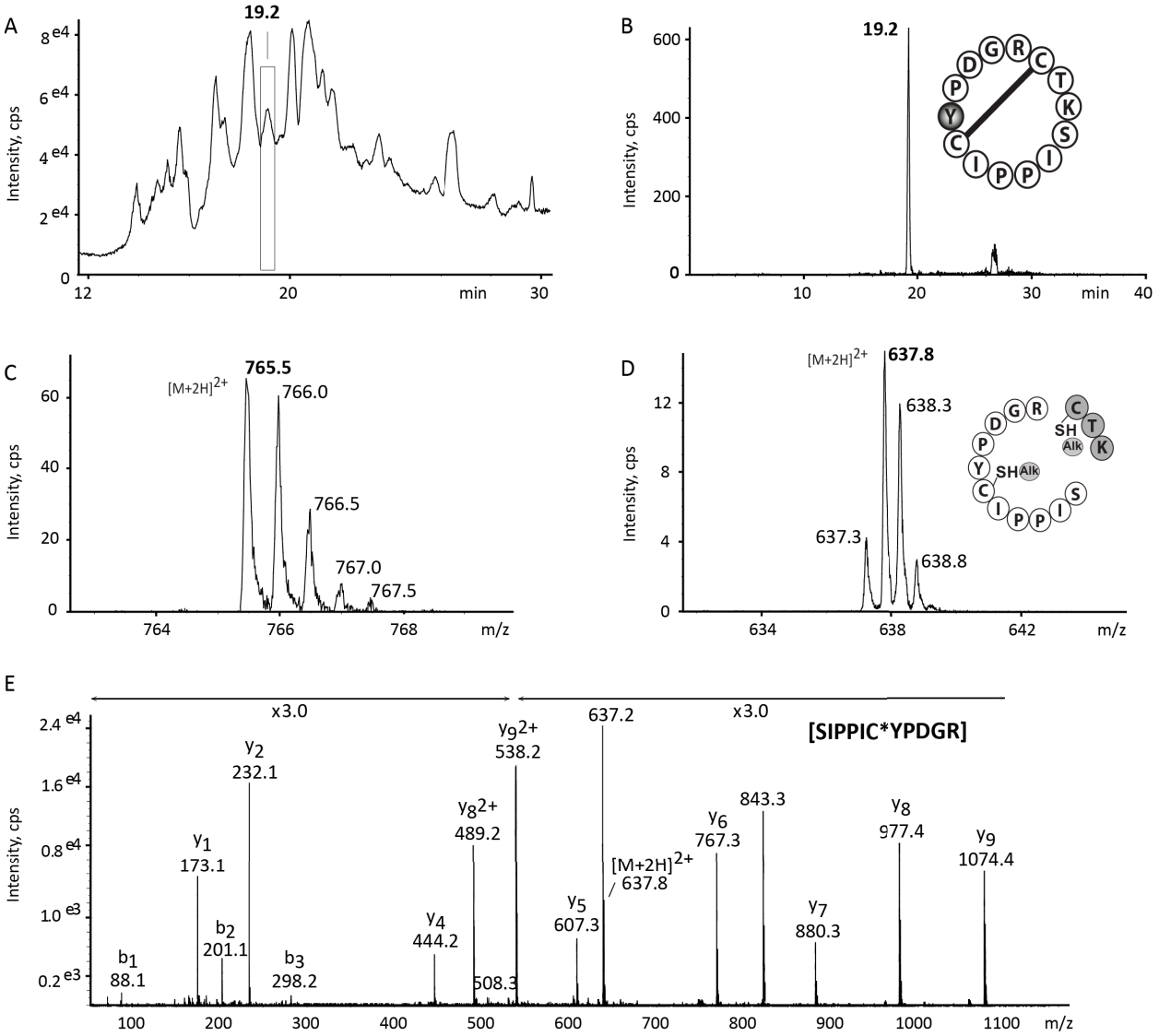
Supplemental Figures



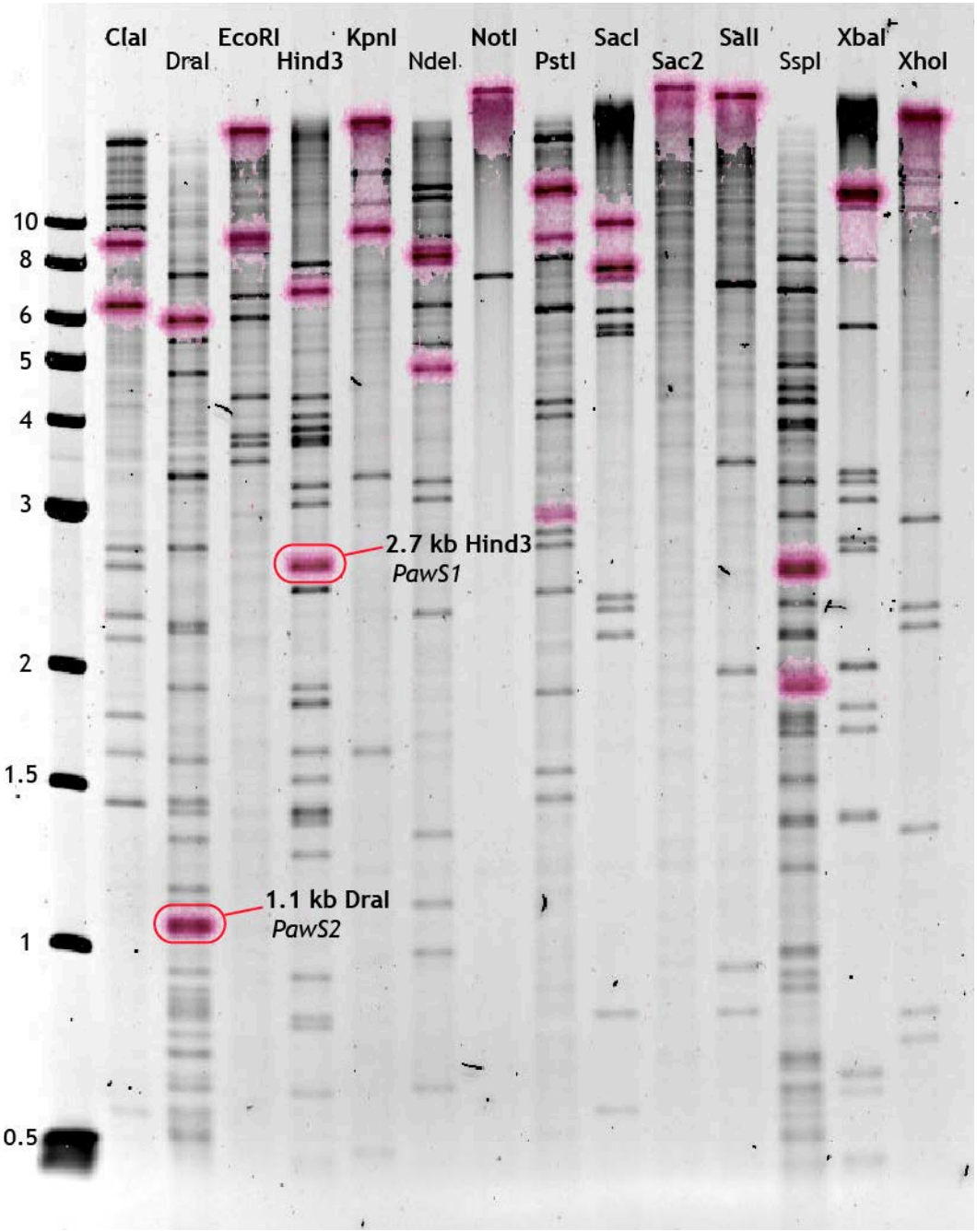
Supplemental Fig. 1: Structural similarity between SFTI-1 and the inhibitory arm of Bowman-Birk Inhibitors (**A**) Sequence of the backbone cyclic and disulfide bonded SFTI-1; (**B**) Weblogo display of an alignment of 150 partial Bowman-Birk sequences. For a full list of sequences used in the alignment see Supplemental Dataset 1; (**C**) Structural overlay of SFTI-1 (PDB code: 1SFI, stick format) with 10 Bowman-Birk inhibitors (PDB codes: 1BBI, 1C2A, 1D6R, 1G9I, 1H34, 1MVZ, 1PBI, 1SMF, 1TAB, 2BBI, line format). The figure was prepared with Pymol. Only the main chain and disulfide bonds are shown; irrelevant parts of PDB files are hidden, all Cys residues are coloured yellow.



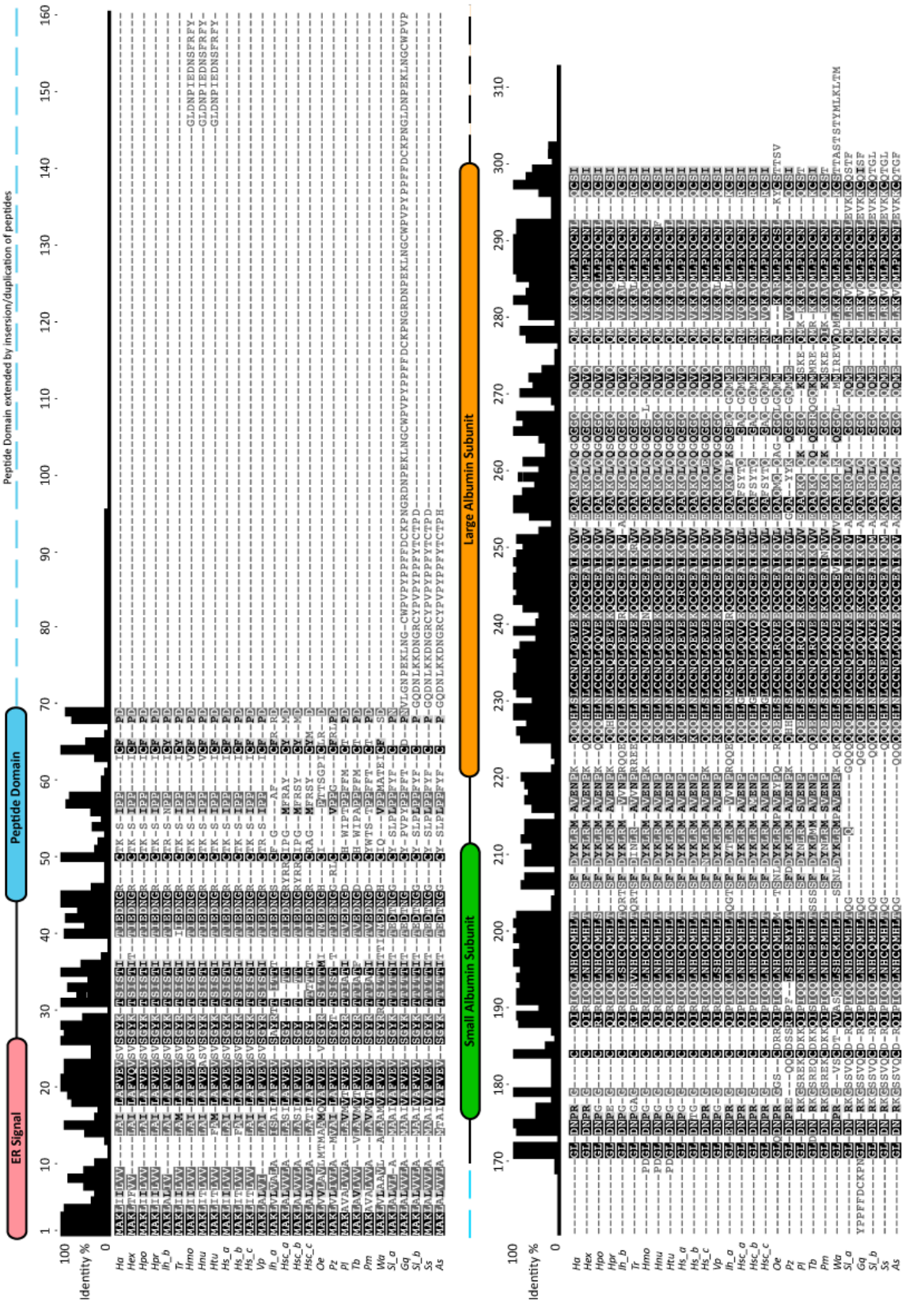
Supplemental Fig. 2: A multiple sequence alignment of PawS1 with preproalbumins from sunflower compared to other plant species highlights a large alignment break covering the SFTI-1 region, indicative of it being the result of genetic insertion. Figure adapted from Mylne et al. (2011). Color schemes for residues are hydrophobic (red), positive (pink), polar (green) and charged (blue). SFTI-1 is boxed with aqua highlight. Predicted or known ER signal sequence cleavage sites are shown with open triangles. Known protease cleavage sites are shown with solid triangles. Conserved Cys residues are asterisked (*). The star marks PawS1. GenBank accessions: Sesame1 (ABB60053), Sesame2 (Q9XHP1), Pumpkin (Q39649), Balsam (CAD32938), BrazilNut (1905414A), Cotton (AAA33066), Pecan (AAO32314), Bnapus (AAA81908), At4g27140 (AGI At4g27140), Mabinlang (P30233), Cashew (AAL91665) and sunflower preproalbumins, HaSFA8 (X56686), HaBA1 (AJ275962), HaG5 (X06410), HaTC17511 (TIGR TC17511), HaTC26991 (TIGR TC26991), PawS1 (FJ469149).



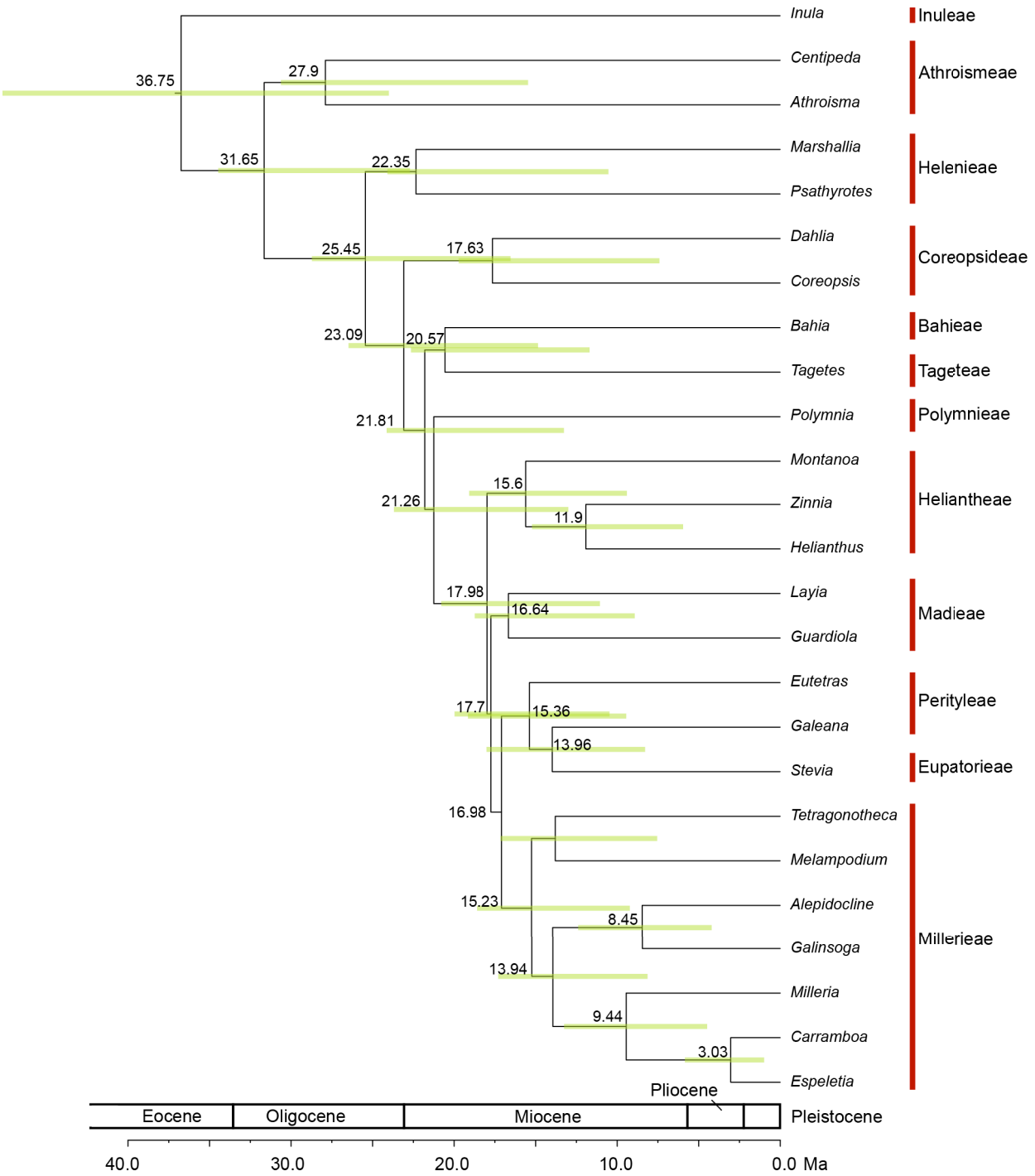
Supplemental Fig. 3: *In planta* confirmation of PDP-3 in seed peptide extracts of *Tithonia rotundifolia.* Y-axis for all panels is Intensity, counts per second (cps). (**A**) LC-MS profile of a seed extraction for *T. rotundifolia*; (**B**) Extracted ion chromatogram (XIC) for the peptide of *m/z* 765.5 reveals a distinct peak at 19.2 min and a predicted schematic of the cyclic peptide sequence, cyclo-GRCTKSIPPICYPD, the Tyr (Y) residue different from the SFTI-1 sequence is shaded in grey; (**C**) Average mass spectrum from 19.1 min - 19.3 min showing the doubly charged ion at *m/z* 765.5, equating to the mass for PDP-3 of 1529.0 Da; (**D**) ESI-TOF-MS of the doubly charged ion at *m/z* 637.8 corresponding to the tryptic digestion product Ser6-Arg2 of PDP-3 with cysteine alkylation The peak at *m/z* 637.3 is an interference and does not equate to the peptide of interest; (**E**) Full scan product ion mass spectrum of the ion at *m/z* 637.8 confirmed the peptide sequence to be PDP-3 [Ser6-Arg2]. The y- and b- sequence ions are labelled.



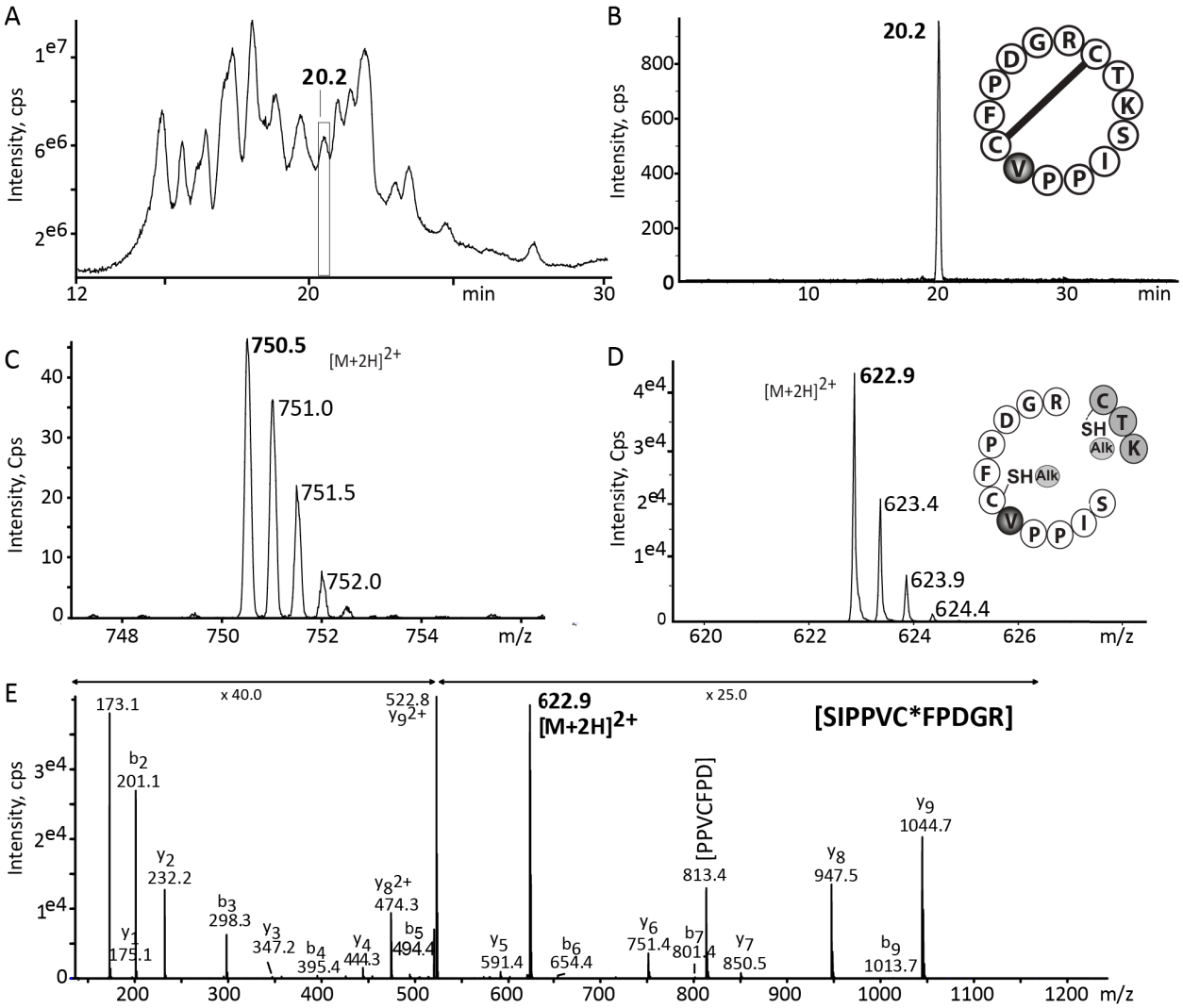
Supplemental Fig. 4: Fingerprinted *Helianthus annuus* bacterial artificial chromosome (BAC) 122C14 from Bouzidi et al. (2006), supplied by The French Plant Genomic Resource Centre (INRA-CNRGV-Toulouse). We digested BAC DNA with a range of restriction enzymes, separated the digests on a 0.8% agarose gel, stained the gel with ethidium bromide and scanned it before it was treated and capillary transferred to a nitrocellulose membrane. The ethidium bromide stained image of DNA fragments has been overlaid (magenta) with its corresponding Southern blot in which a ^{32}P-labelled probe that would bind both *PawS1* and *PawS2* was used to identify the *PawS*-positive DNA fragments. The 1.1 kb band was cloned and sequenced (GenBank JX910423) and found to contain *PawS2*. The 2.7 kb HindIII band was similarly sequenced (GenBank JX910422) and found to contain *PawS1*. This allowed us to design primers to the sequences flanking *PawS1* and *PawS2* ORFs.



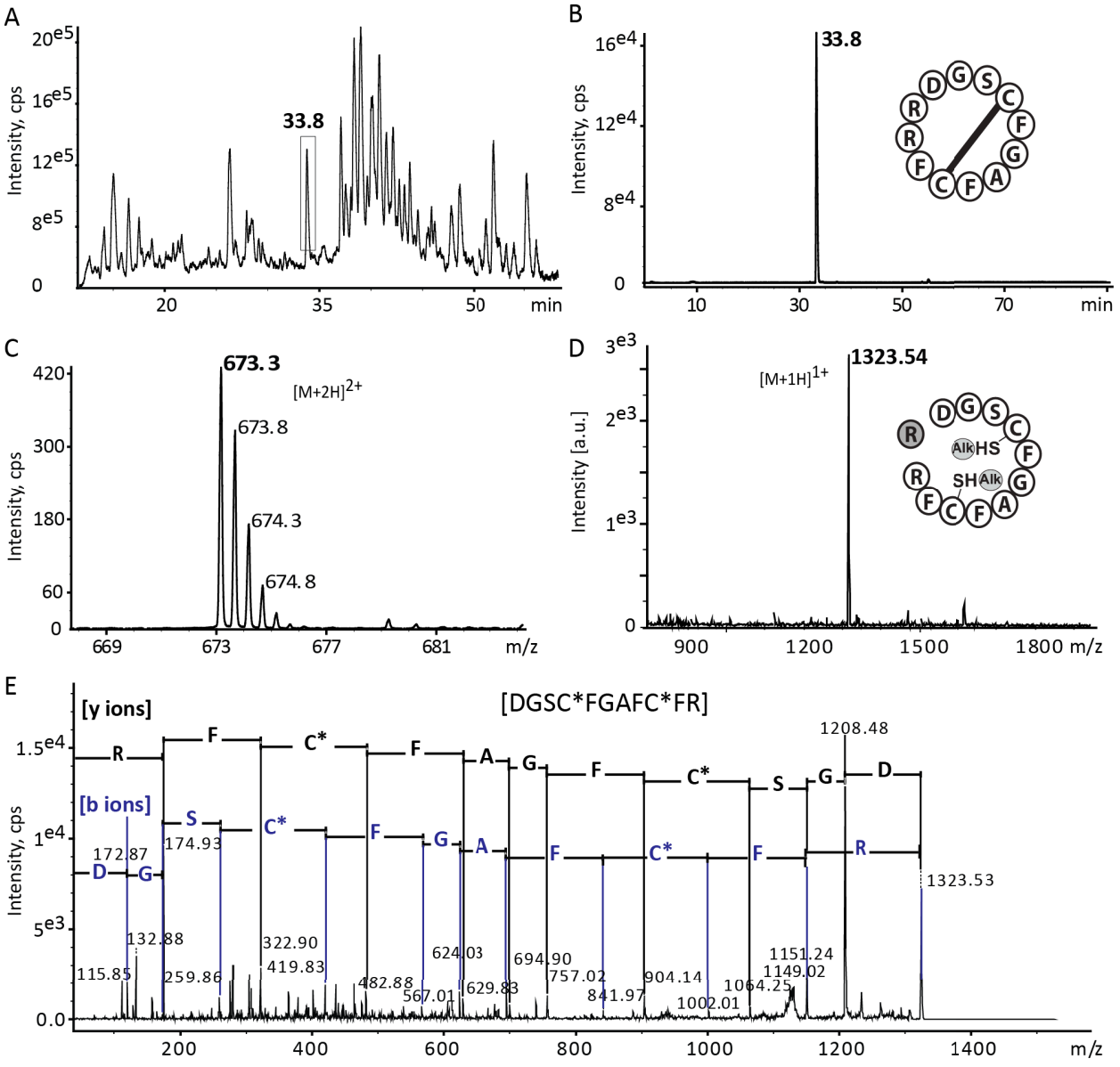
Supplemental Fig. 5: Alignment of predicted protein sequences for 27 *PawS1* genes aligned against *PawS1* from *Helianthus annuus*. At the top is the percent identity graph followed by the sequence alignment. See Supplemental Table 1 for details of each gene. The alignment was generated in Geneious version 6.5.4.



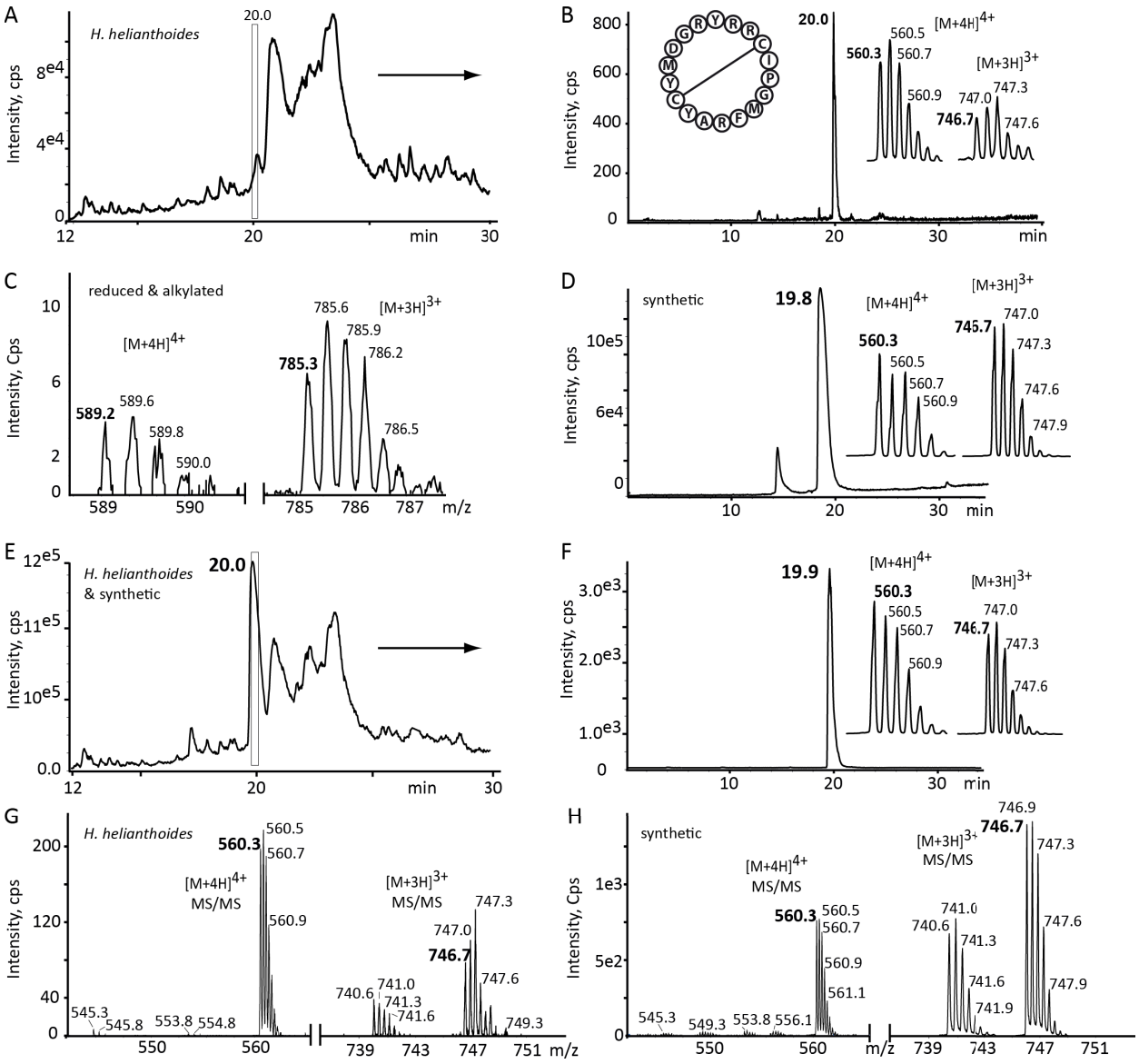
Supplemental Fig. 6: Chronogram for 25 genera of Asteraceae subfamily Asteroideae produced by BEAST under a Yule prior uncorrelated lognormal relaxed clock model. Green bars on node ages represent 95% highest posterior densities of divergence times. Ma, million years ago. Mean age estimates are shown above branches. For sequences and alignment, see Supplemental Dataset 5-6.



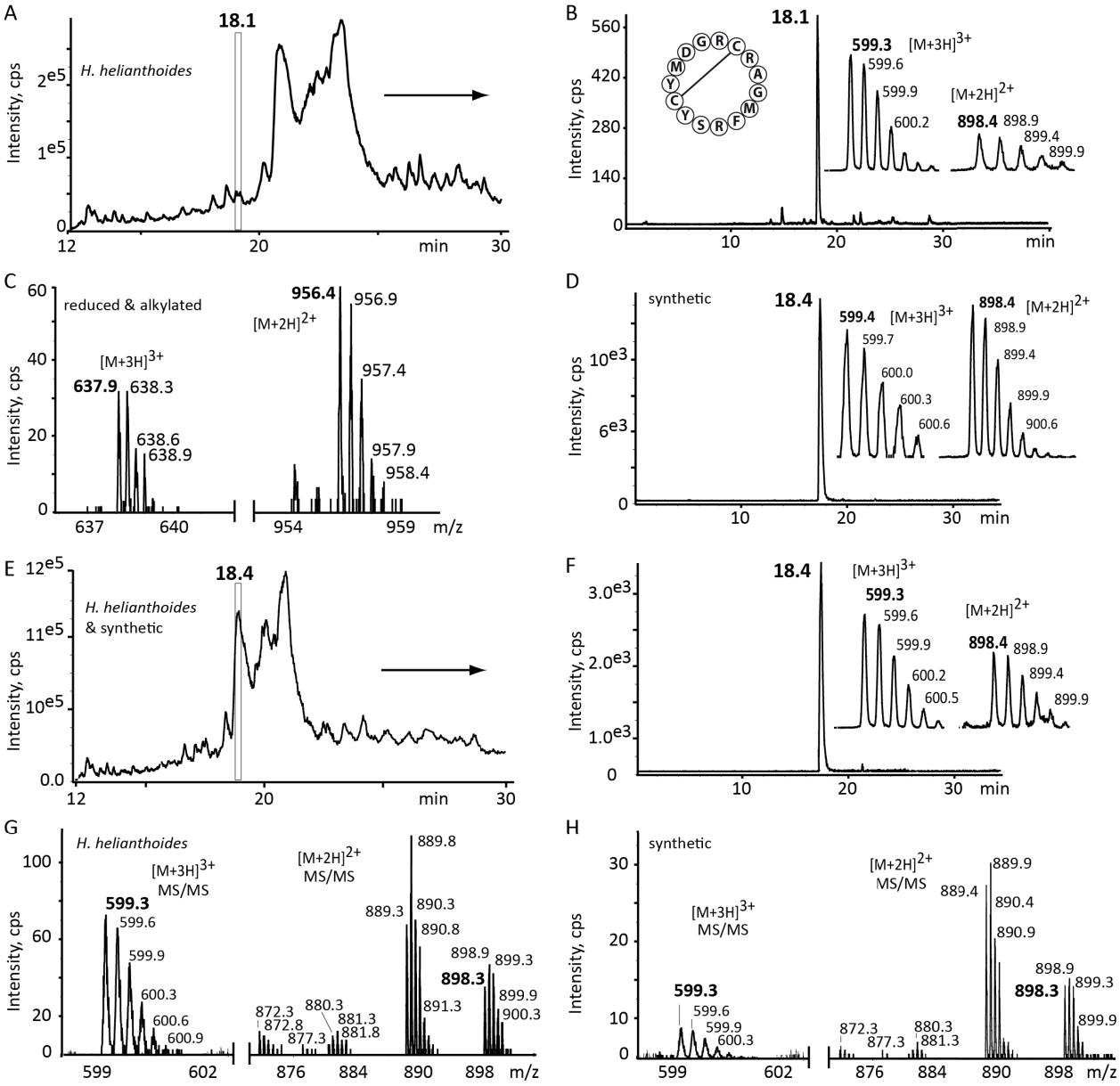
Supplemental Fig. 7: *In planta* confirmation of PDP-12 within peptide extracts from *Helianthus schweinitzii* seeds (**A**) LC-MS profile of a seed extraction for *H. schweinitzii*; (**B**) Extracted ion chromatogram (XIC) for the peptide of *m/z* 750.5 and a predicted schematic of the cyclic peptide sequence, cyclo-GRCTKSIPPVCFPD, the Val (V) residue different from SFTI-1 sequence is shaded in grey; (**C**) Average mass spectrum from 20.1 min – 20.3 min showing the doubly charged ion at *m/z* 750.5, equating to the mass of PDP-I2 1499.0 Da; (**D**) ESI-TOF-MS of the doubly charged ion at *m/z* 622.9 corresponding to the tryptic digestion product Ser6-Arg2 of PDP-12 with cysteine alkylation (Alk). The CTK shaded (grey) were cleaved from the peptide by tryptic digestion so are not present in the sequencing in the next panel; (**E**) Full scan product ion mass spectrum of the ion at *m/z* 622.9 confirmed the peptide sequence to be PDP-12 [Ser6-Arg2]. The y- and b- sequence ions are labelled.



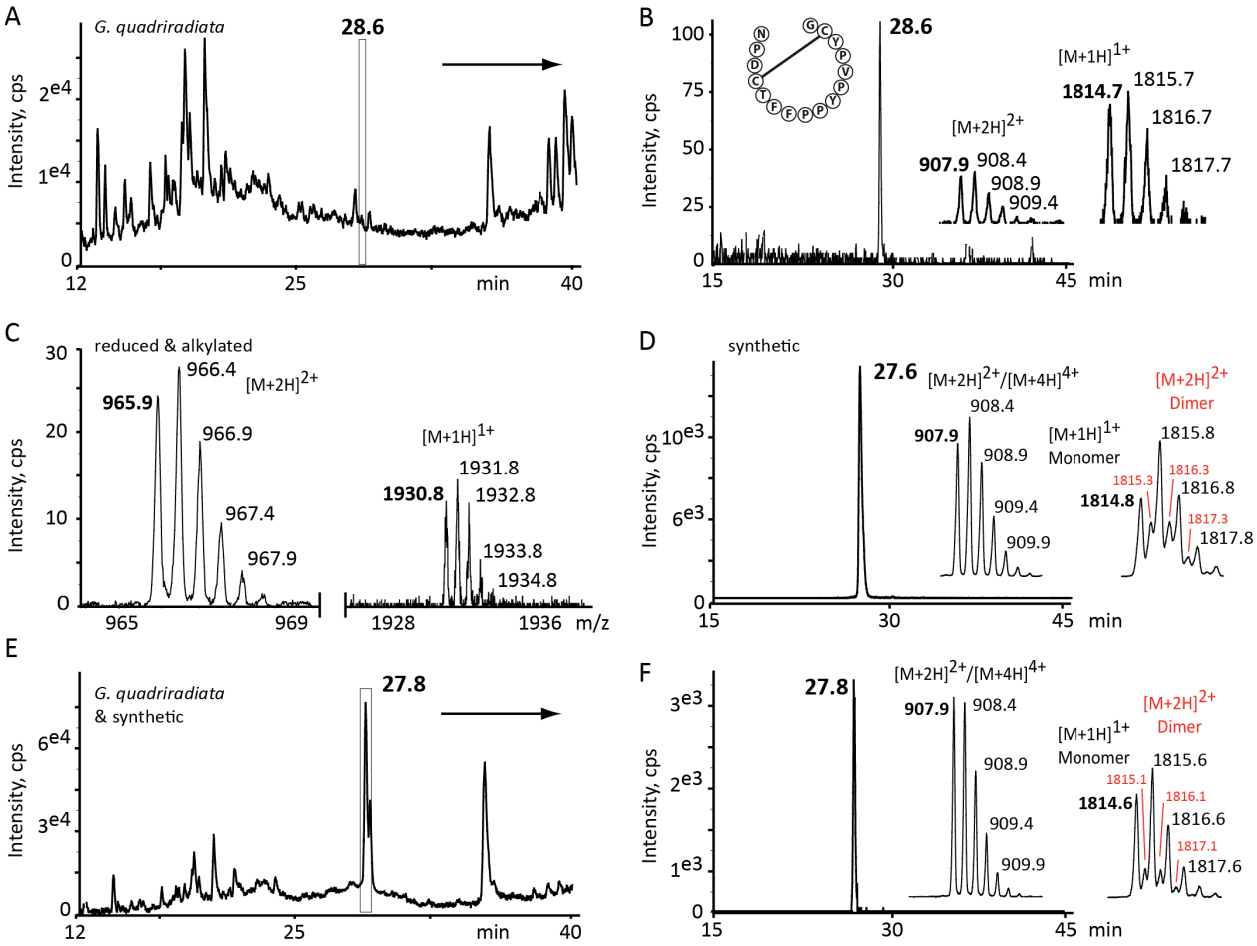
Supplemental Fig. 8: *In planta* confirmation of PDP-4 in seed peptide extracts of *Iostephane heterophylla* (**A**) LC-MS profile of a seed extraction for *I. heterophylla*; (**B**) Extracted ion chromatogram (XIC) for the peptide of *m/z* 673.3 and a predicted schematic of the cyclic peptide sequence, cyclo-GSCFGAFCFRRD; (**C**) Average mass spectrum from 33.7 min – 33.9 min showing the doubly charged ion at *m/z* 673.3, equating to the mass of PDP-4 1344.6 Da; (**D**) MALDI-TOF-MS of the singly charged ion at 1323.5 corresponding to the tryptic digestion product Asp12-Arg10 of PDP-4 with cysteine alkylation (Alk). A schematic representation of the peptide fragment is depicted, the Arg (R) shown in the grey circle is the released residue by cleavage at both Arg (R) residues; (**E**) Full scan product ion mass spectrum of the ion at *m/z* 1323.5 confirmed the peptide sequence to be PDP-4 [Asp12-Arg10]. The y- and b- sequence ions are labelled.



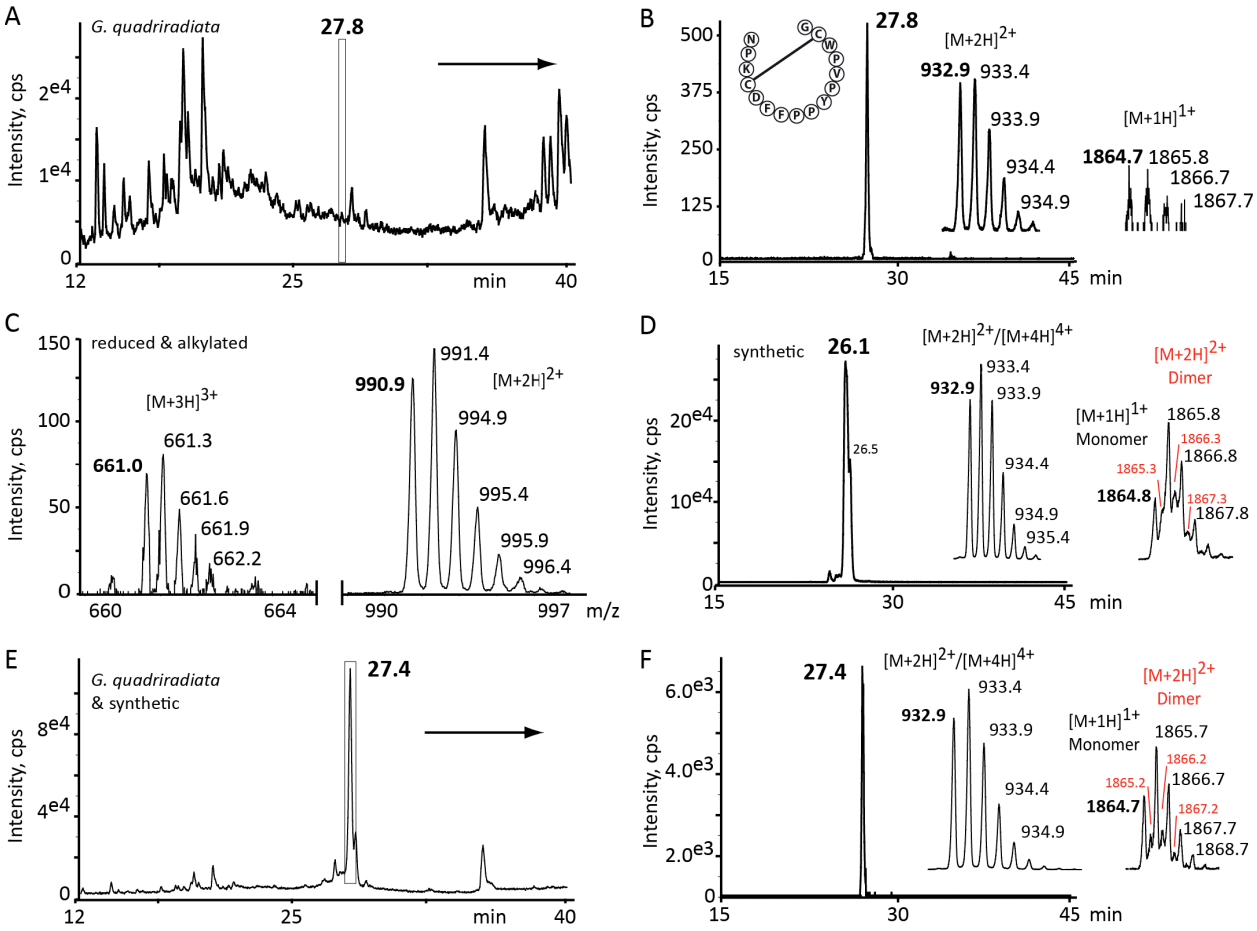
Supplemental Fig. 9: *In planta* confirmation of PDP-5 within peptide extracts made from *Heliopsis helianthoides scabra* seeds (**A**) LC-MS profile of a seed extraction for *H. helianthoides*; (**B**) Extracted ion chromatogram (XIC) and average MS spectrum from 19.9 min – 20.1 min showing the $[M+3H]^{3+}$ and $[M+4H]^{4+}$ ions at *m/z* 746.7 and 560.3, equating to the mass of PDP-5 2237.1 Da and a predicted schematic of the cyclic peptide sequence, cyclo-GRYRRCIPGMFRAYCYMD; (**C**) $[M+3H]^{3+}$ and $[M+4H]^{4+}$ ions from reduced and alkylated crude plant extract identifies a two cysteine (single disulfide bond) containing peptide. The masses are consistent with an increase of two alkyl groups (final mass 2353.0 Da), one per cysteine; (**D**) XIC for synthetic PDP-5 displaying a profile consistent to the native; (**E**) LC-MS profile of *H. helianthoides* spiked with synthetic PDP-5 to show an increase in the subject peptide peak and no other differences; (**F**) XIC and average MS from **E**; (**G-H**) Zoomed in MS/MS spectra showing the $[M+3H]^{3+}$ and $[M+4H]^{4+}$ ions for PDP-5 native (**F**) and for synthetic (**G**) respectively, showing the same precursor ion pattern. At the lower *m/z* range there was no evidence of ion fragmentation for either the synthetic or the native peptide, as expected for a cyclic peptide.



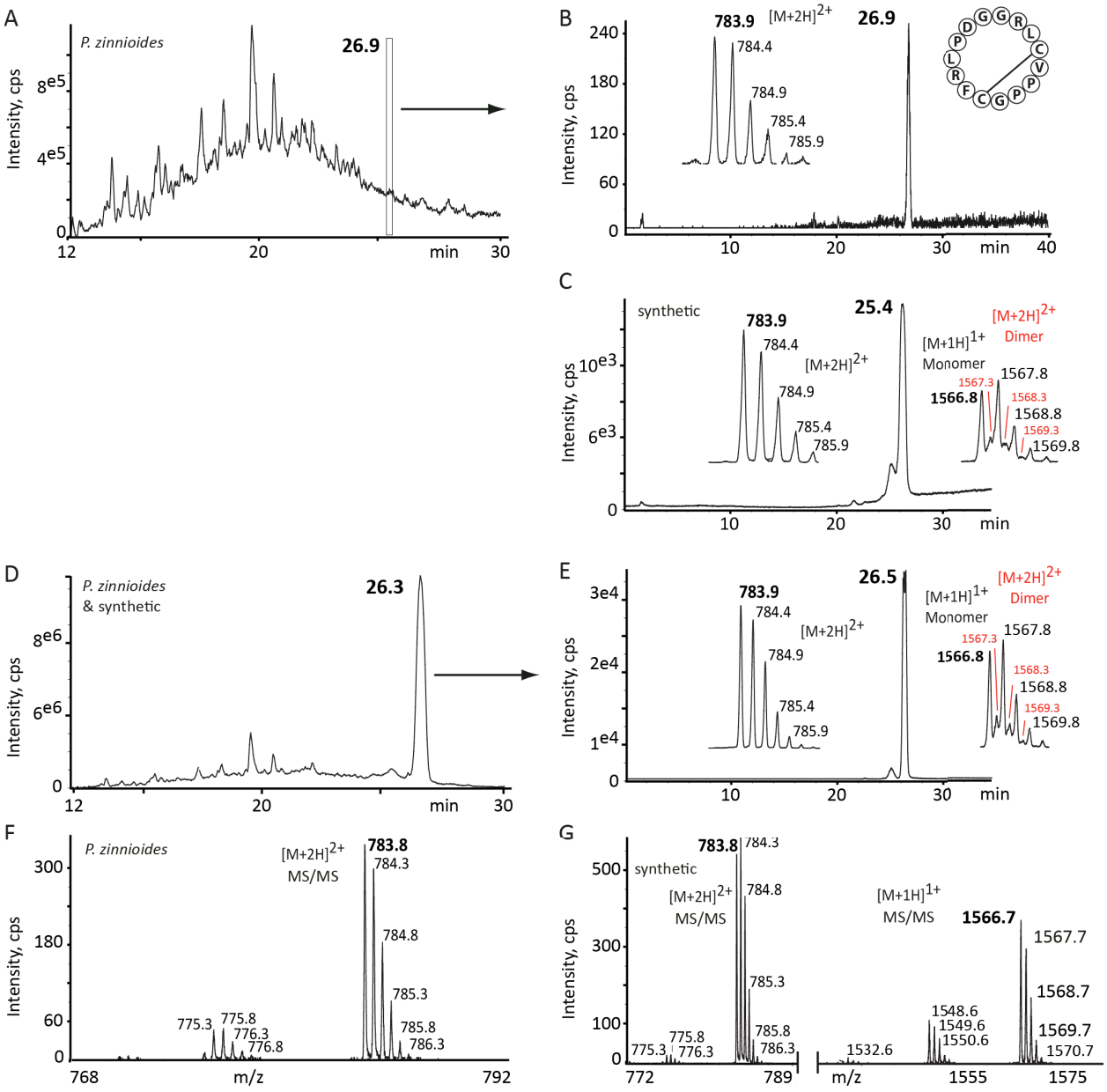
Supplemental Fig. 10: *In planta* confirmation of PDP-14 within peptide extracts made from *Heliopsis helianthoides scabra* seeds (**A**) LC-MS profile of a seed extraction for *H. helianthoides*; (**B**) Extracted ion chromatogram (XIC) and average MS spectrum from 18.0 min – 18.2 min showing the $[M+2H]^{2+}$ and $[M+3H]^{3+}$ ions at *m/z* 898.4 and 599.3, equating to the mass of PDP-14 1794.9 Da and a predicted schematic of the cyclic peptide sequence, cyclo-GRCRAGMFRSYCYMD; (**C**) Resulting $[M+2H]^{2+}$ and $[M+3H]^{3+}$ ions from reducing and alkylating the crude plant extract to identify a two cysteine (single disulfide bond) containing peptide. The masses are consistent with an increase of two alkyl groups (final mass 1910.7 Da), one per cysteine; (**D**) XIC for synthetic PDP-14 displaying a profile consistent to the native; (**E**) LC-MS profile of *H. helianthoides* spiked with synthetic PDP-14 to show an increase in the subject peptide peak and no other differences; (**F**) XIC and average MS from **E**; (**G-H**) Zoomed in MS/MS spectra showing the $[M+2H]^{2+}$ and $[M+3H]^{3+}$ ions for PDP-14 native (**G**) and for the synthetic (**H**) respectively, showing the same precursor ion pattern. At the lower *m/z* range there was no evidence of ion fragmentation for either the synthetic or the native peptide, as expected for a cyclic peptide.



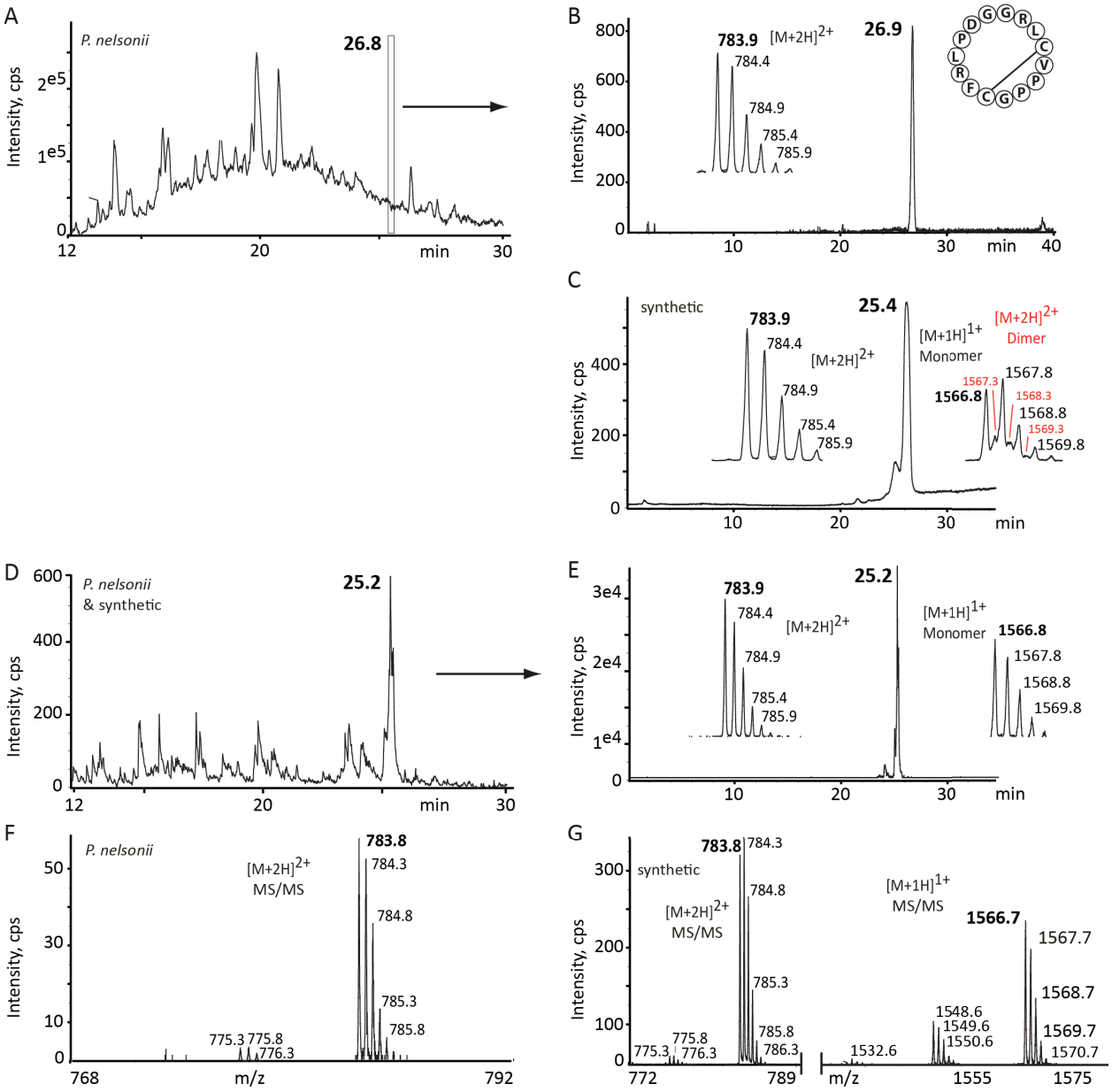
Supplemental Fig. 11: *In planta* confirmation of PDP-10 within peptide extracts made from *Galinsoga quadriradiata* seeds (**A**) LC-MS profile of a seed extraction for *G. quadriradiata*; (**B**) Extracted ion chromatogram (XIC) and average MS spectrum from 28.5 min – 28.7 min showing the $[M+1H]^{1+}$ and $[M+2H]^{2+}$ ions at *m/z* 1814.7 and 907.9, equating to the mass of 1813.8 Da for PDP-10 and a predicted schematic of the cyclic peptide sequence, acyclic-GCYPVPYPPFFTCDPN; (**C**) $[M+1H]^{1+}$ and $[M+2H]^{2+}$ ions from reduced and alkylated crude plant extract identified a two cysteine (single disulfide bond) containing peptide. The masses are consistent with an increase of two alkyl groups (final mass 1929.8 Da), one per cysteine; (**D**) XIC for synthetic PDP-10 displaying a profile closely consistent to the native. A portion of the synthetic peptide forms a dimer and co-elutes with the monomer form, decreasing the retention time by 1 min compared to the native; this is evident by the doubly charged 1814.8 MS ion that displays an overlapping pattern with the singly charged ions of the monomer, the additional masses not attributed to the monomer are in red. The $[M+2H]^{2+}$ ion also displayed here may represent both the 2+ charged state of the monomer and the 4+ charged state of the dimer; (**E**) LC-MS profile of *G. quadriradiata* spiked with synthetic PDP-10 to show an increase in the subject peptide peak and no other differences; (**F**) XIC and average MS from panel **E**.



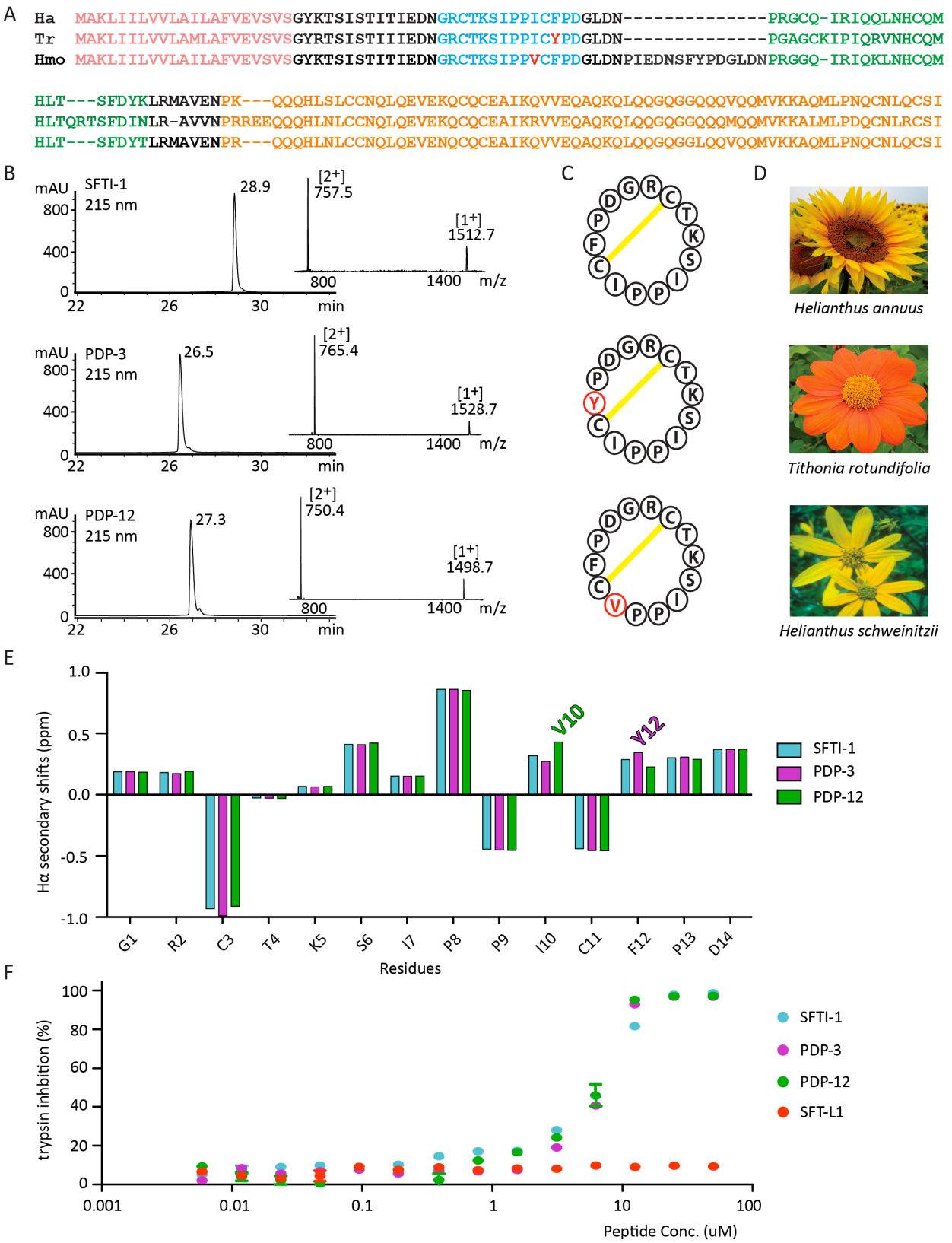
Supplemental Fig. 12: *In planta* confirmation of PDP-11 within peptide extracts made from *Galinsoga quadriradiata* seeds (**A**) LC-MS profile of a seed extraction for *G. quadriradiata*; (**B**) Extracted ion chromatogram (XIC) and average MS spectrum from 27.7 min – 27.9 min showing the $[M+1H]^{1+}$ and $[M+2H]^{2+}$ ions at *m/z* 1864.7 and 932.9, equating to the mass of 1863.8 Da for PDP-11 and a predicted schematic of the cyclic peptide sequence, acyclic-GCWPVPYPPFFDCKPN; (**C**) Resulting $[M+2H]^{2+}$ and $[M+3H]^{3+}$ charged ions from reducing and alkylating the crude plant extract to identify a two cysteine (single disulfide bond) containing peptide. The masses are consistent with an increase of two alkyl groups (final mass 1979.8 Da), one per cysteine; (**D**) XIC for synthetic PDP-11 displaying a profile closely consistent to the native. A portion of the synthetic peptide forms a dimer and co-elutes with the monomer form, decreasing the retention time by 1 min compared to the native, this is evident by the doubly charged 1864.8 MS ion that displays an overlapping pattern with the singly charged ions of the monomer, and the additional masses not attributed to the monomer are in red. The $[M+2H]^{2+}$ ion also displayed here may represent both the 2+ charged state of the monomer and the 4+ charged state of the dimer; (**E**) LC-MS profile of *G. quadriradiata* spiked with synthetic PDP-11 to show an increase in the subject peptide peak and no other differences; (**F**) XIC and average MS from panel **E**.



Supplemental Fig. 13: *In planta* confirmation of PDP-8 within peptide extracts made from *Philactis zinnioides* seeds (**A**) LC-MS profile of a seed extraction for *P. zinnioides*; (**B**) Extracted ion chromatogram (XIC) and average MS spectrum from 26.8 min – 30 min showing the $[M+2H]^{2+}$ ions at *m/z* 783.9, equating to the mass of PDP-8 (1565.8 Da) and a predicted schematic of the cyclic peptide sequence, cyclic-GGRLCVPPGCFRLPD; (**C**) XIC and average MS of synthetic PDP-8 displaying a profile consistent to the native; (**D**) LC-MS profile of *P. zinnioides* spiked with synthetic PDP-8 to show an increase in the subject peptide peak and no other differences; (**E**) XIC and average MS from **D**; (**F-G**) Zoomed in MS/MS spectra showing the $[M+2H]^{2+}$ ion for PDP-8 native (**F**) and the $[M+1H]^{1+}$ and $[M+2H]^{2+}$ ions for synthetic peptide (**G**) respectively, showing the same precursor ion pattern. At the lower *m/z* range there was no evidence of ion fragmentation for either the synthetic or the native peptide, as expected for a cyclic peptide.

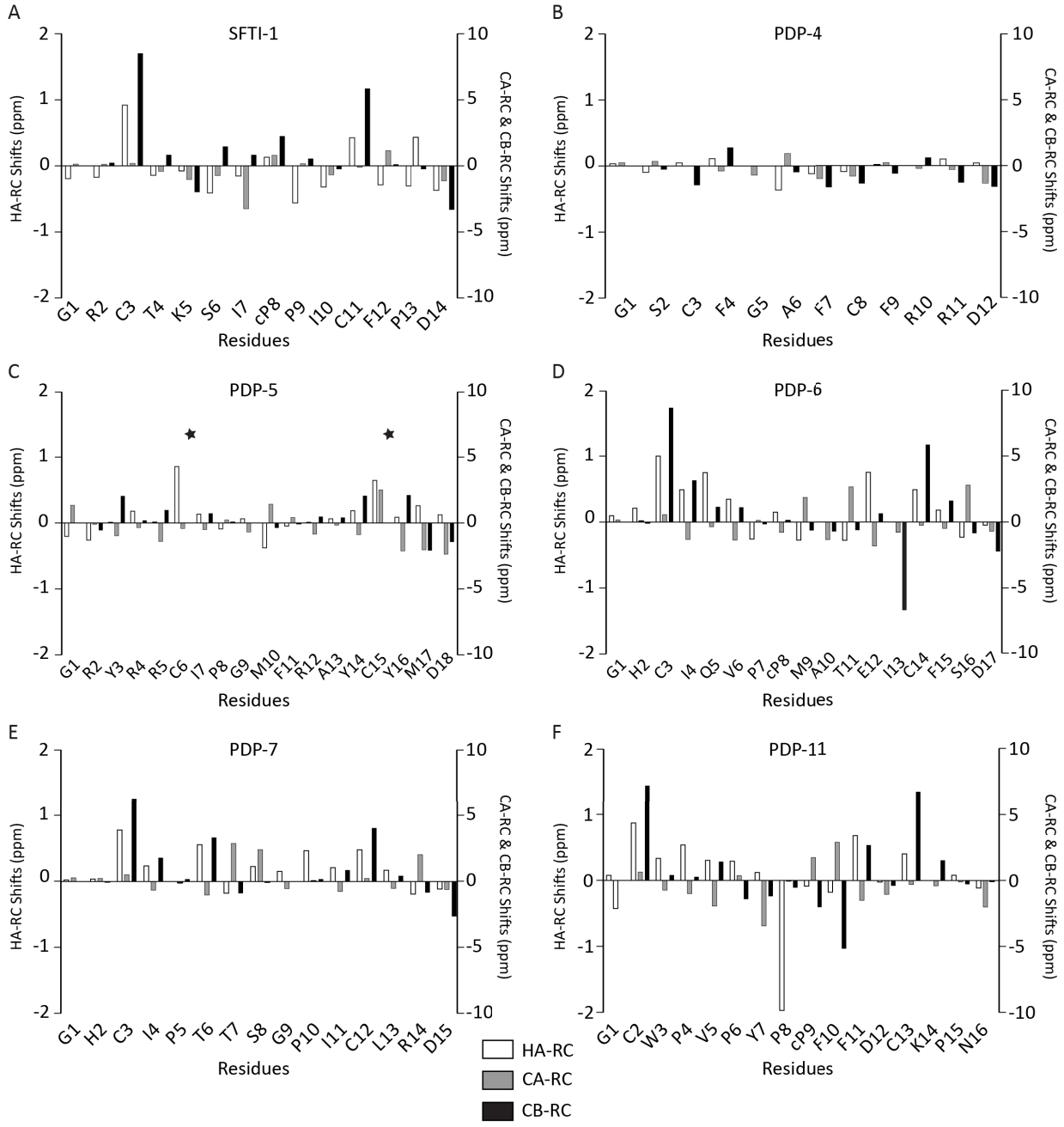


Supplemental Fig. 14: *In planta* confirmation of PDP-8 within peptide extracts made from *Philactis nelsonii* seeds (**A**) LC-MS profile of a seed extraction for *P. nelsonii*; (**B**) Extracted ion chromatogram (XIC) and average MS spectrum from 26.8 min – 30 min showing the $[M+2H]^{2+}$ charged ions at *m/z* 783.9, equating to the mass of PDP-8 (1565.8 Da) and a predicted schematic of the cyclic peptide sequence, cyclic-GGRLCVPPGCFRLPD; (**C**) XIC and average MS of synthetic PDP-8 displaying a profile consistent to the native; (**D**) LC-MS profile of *P. nelsonii* spiked with synthetic PDP-8 to show an increase in the subject peptide peak and no other differences; (**E**) XIC and average MS from **D**; (**F-G**) Zoomed in MS/MS spectra showing the $[M+2H]^{2+}$ ion for PDP-8 native (**F**) and the $[M+1H]^{1+}$ and $[M+2H]^{2+}$ ions for synthetic peptide (**G**) respectively, showing the same precursor ion pattern. At the lower *m/z* range there was no evidence of ion fragmentation for either the synthetic or the native peptide, as expected for a cyclic peptide.

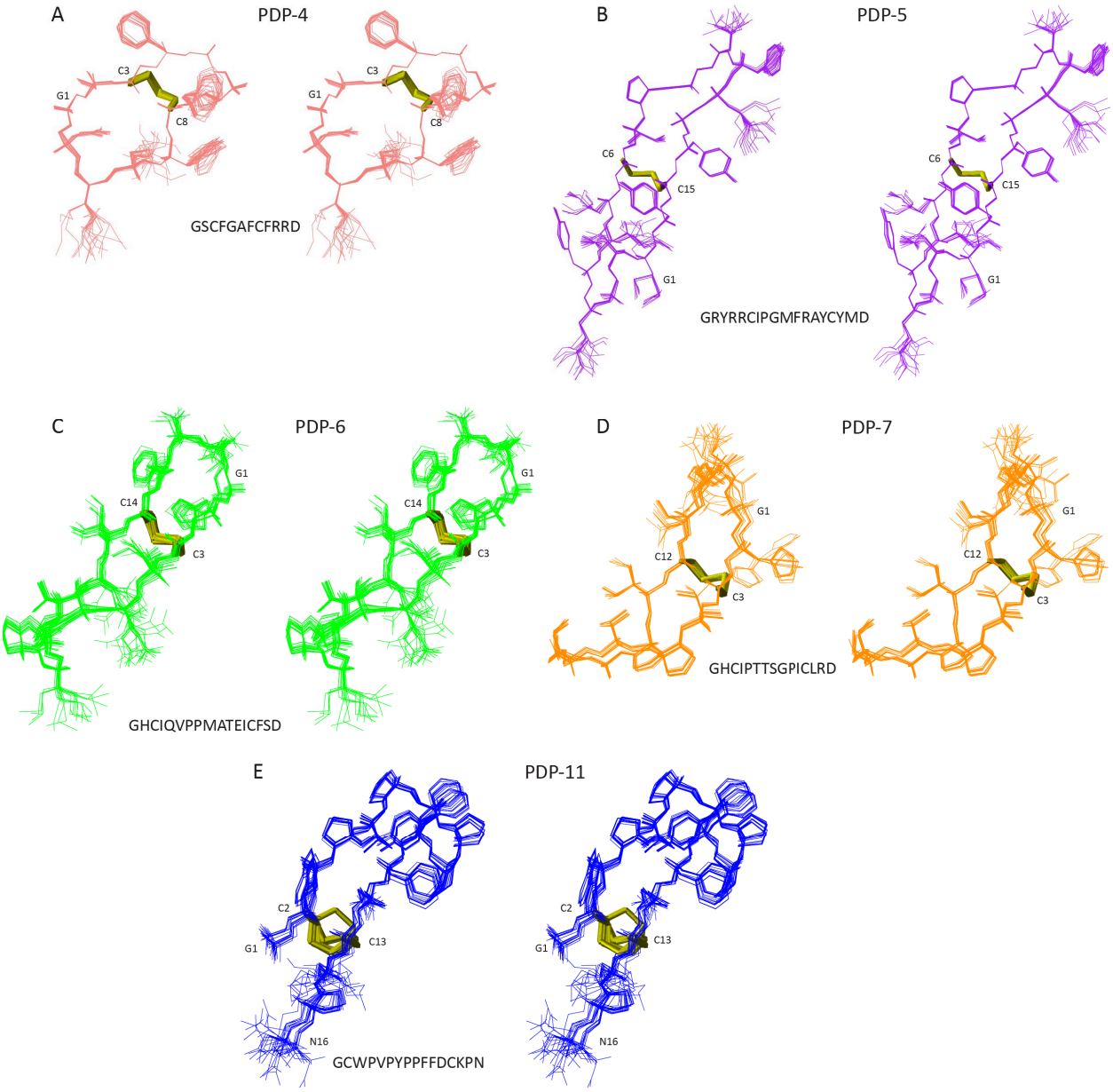


Supplemental Fig. 15: Structure and activity comparison for SFTI-1 and variant peptides (**A**) Alignment of *PawS1* gene sequences for *H. annuus* (Ha), *T. rotundifolia* (Tr) and *H. mollis* (Hmo) encoding peptides SFTI-1, PDP-3 and PDP-12, respectively; (**B**) RP-HPLC profile and MS of chemically synthesized versions of

these peptides; (**C**) Schematic of each cyclic peptide with the residue that varies from the sequence of SFTI-1 highlighted in red, i.e. F12Y in PDP-3 and I10V in PDP-12; (**D**) The flowers of *H. annuus, T. rotundifolia* and *H. schweinitzii* represent the plants in which the seed peptides were extracted and sequenced, for PDP-3 see Supplemental Fig. 3 and for PDP-12 see Supplemental Fig. 7 for full sequencing details; (**E**) Hα secondary shifts calculated by comparison of the random coil shifts assigned by structural studies with theoretically determined shifts (Wishart et al., 1995). These shifts show the backbone of the peptides to be significantly similar and only differing slightly at the altered residues; (**F**) trypsin inhibitory assay for the three peptides shows no variation in activity and SFT-L1 is included as a negative control, as previously described (Mylne et al., 2011).

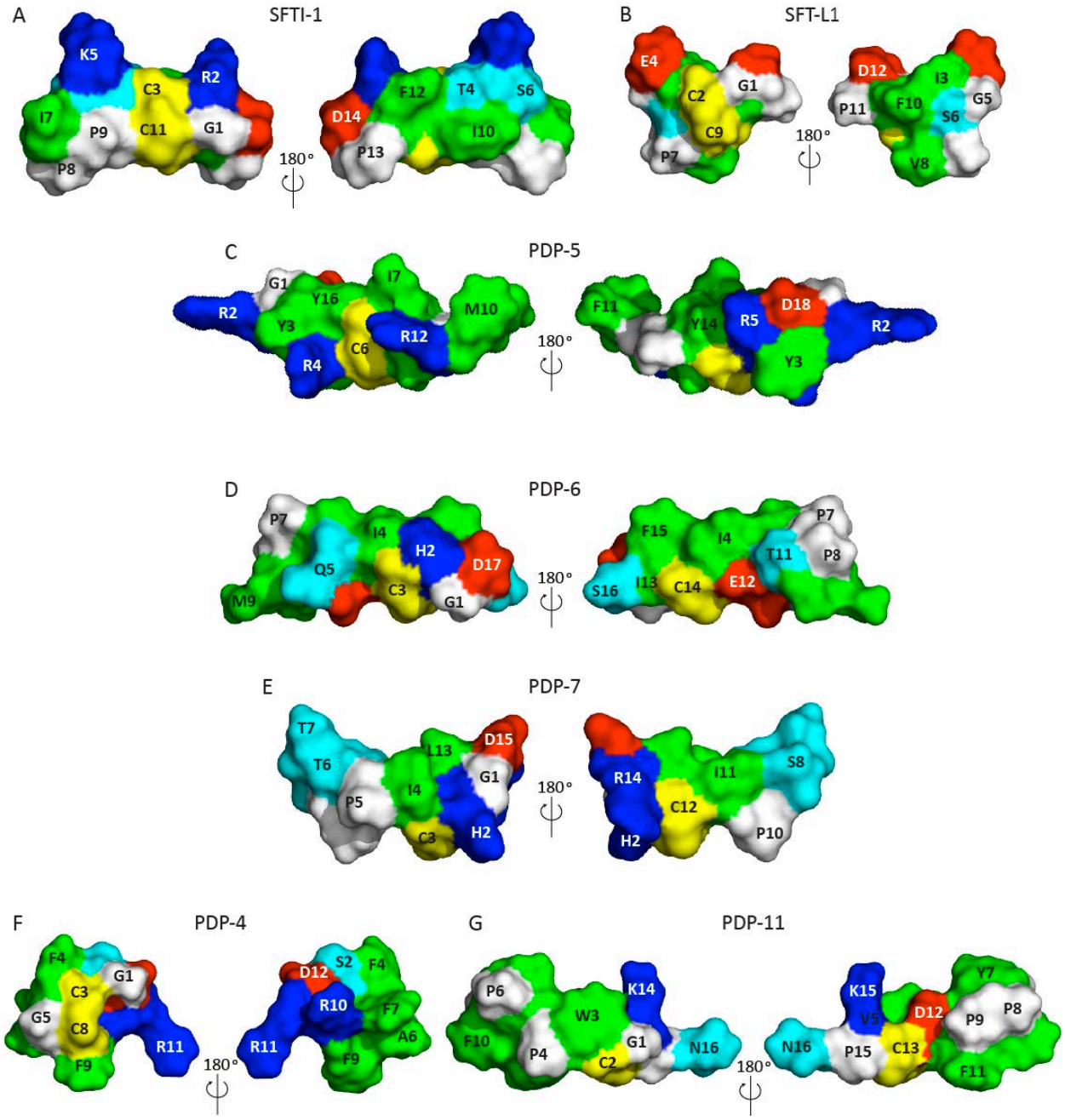


Supplemental Fig. 16: PDP secondary chemical shifts. Secondary chemical shifts were calculated by comparing the experimentally observed Hα chemical shifts (HA) for each residue with the chemical shifts observed for the corresponding residues in random coil (RC) peptides (Wishart et al., 1995). (**A**) SFTI-1; (**B**) PDP-4; (**C**) PDP-5, the stars signify that the Cβ shifts for Cys6 and Cys15 are not zero yet were not seen due to resonance broadening; (**D**) PDP-6; (**E**) PDP-7; (**F**) PDP-11. Natural abundance ^{13}C HSQC experiments to obtain Cα and Cβ shifts were not reported in the original study for SFTI-1 (Korsinczky et al., 2001). Therefore, all data shown here were collected in this study and compared to the original structure to cross check accuracy of assignments. Hα-RC shifts are in white, Cα-RC shifts are in grey and Cβ-RC shifts are in black. All PDPs have disulfide bonds with short right-handed hook conformations that produce consistent patterns with the Cys Hα and Cβ chemical shifts being downfield shifted by 0.5-1 ppm and 5-10 ppm, respectively.

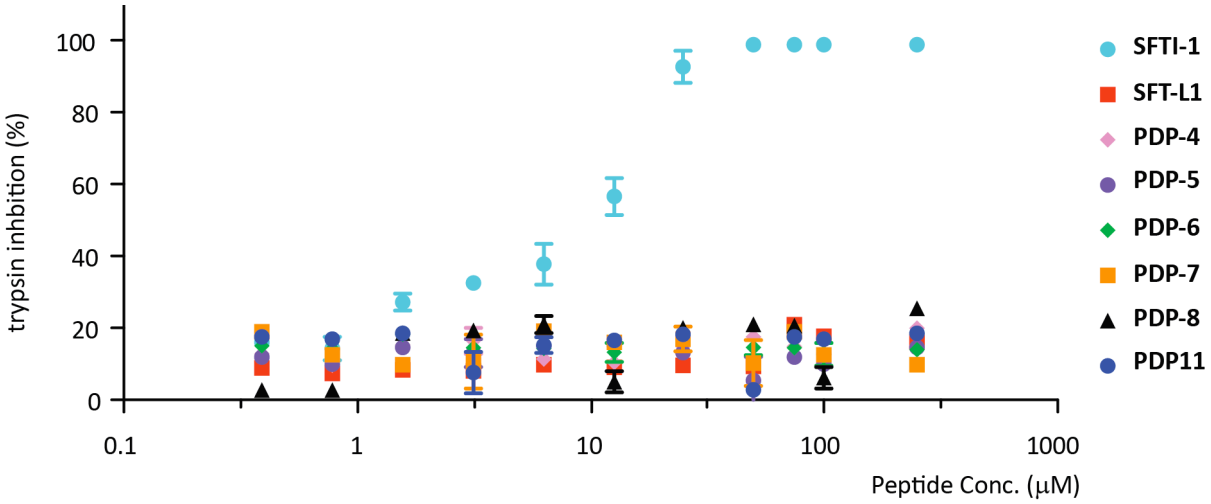


Supplemental Fig. 17: Stereo view of the family of 20 structures with highest MOLPROBITY (Chen et al., 2010) score for each PDP determined in this study (**A**) PDP-4 (pink); (**B**) PDP-5 (purple); (**C**) PDP-6 (green); (**D**) PDP-7 (orange); (**E**) PDP-11 (blue). The amino acid sequence of each peptide is shown below the peptide and selected residues are labeled for orientation. Like SFTI-1, PDP-5, -6, -7 and -11 all possess distinct β-strand regions, whereas PDP-4 is made up of turns like SFT-L1, as highlighted by the ribbon representation in Fig. 2A.

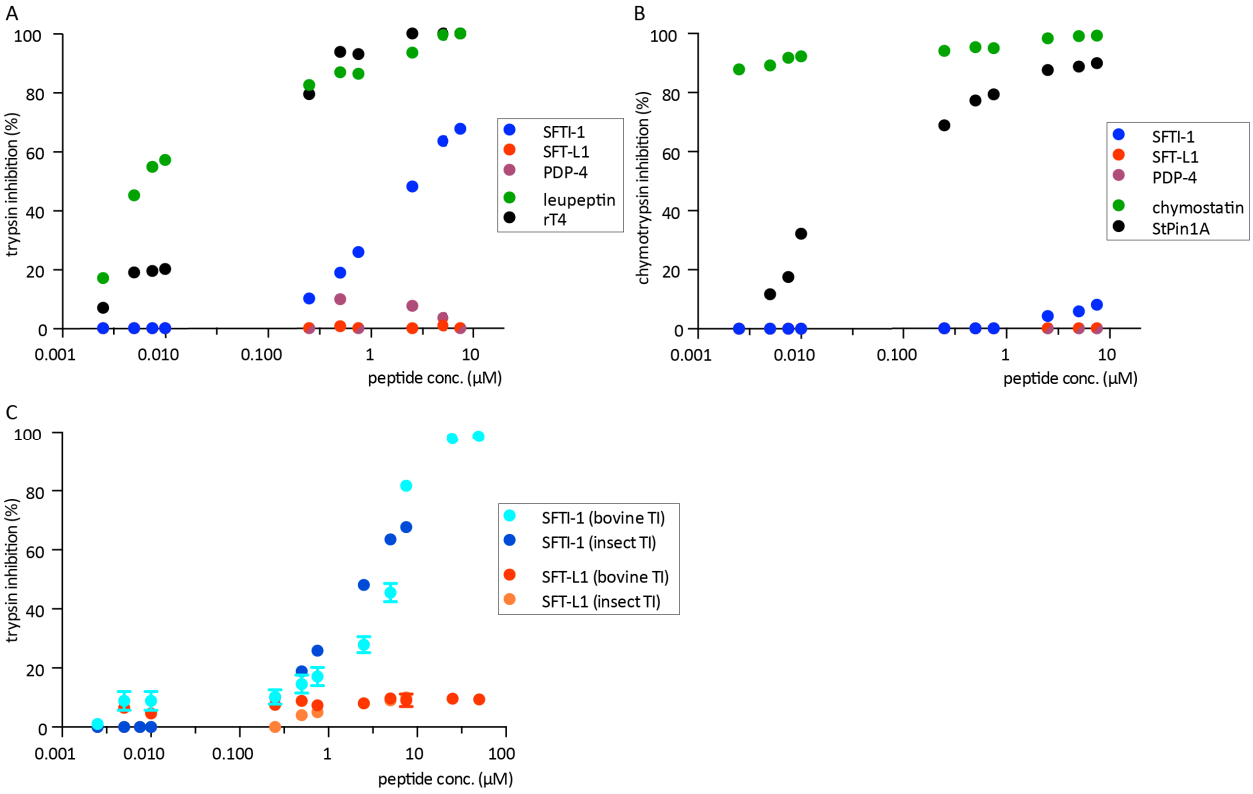
.



Supplemental Fig. 18: Molecular surface representation and comparison of PDPs (**A**) SFTI-1(Korsinczky et al., 2001); (**B**) SFT-L1 (Mylne et al., 2011); (**C**) PDP-5; (**D**) PDP-6; (**E**) PDP-7; (**F**) PDP-4; (**G**) PDP-11. Hydrophobic residues are colored green, polar residues in cyan, neutral residues in white, cysteines in yellow, negatively charged residues in red and positively charged residues in blue. Highly characteristic is the large portion of hydrophobic residues in all PDPs and one to three protruding positively charged residues for all PDPs except SFT-L1 and PDP-6 where the protruding Lys5 in SFTI-1 is responsible for its potent trypsin inhibition. Selected amino acids are labeled by one-letter code for orientation reference.

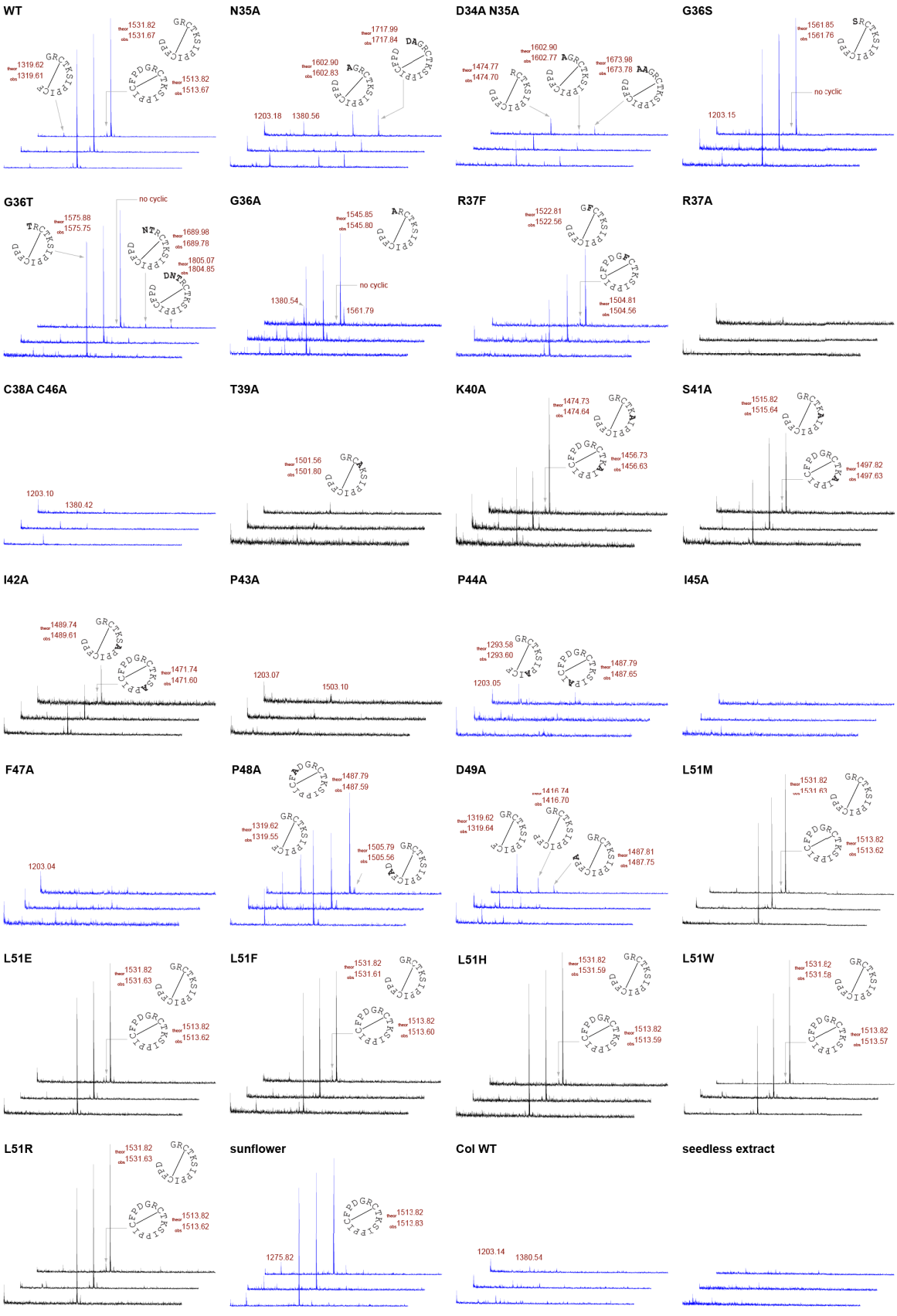


Supplemental Fig. 19: Trypsin inhibitory assays. PDPs were assayed for their ability to inhibit the digestive enzyme trypsin, as SFTI-1 is a potent inhibitor of trypsin (Luckett et al., 1999). SFT-L1 was used as a negative control as it has previously been shown to lack inhibitor ability against trypsin (Mylne et al., 2011). The graph shows the lack of trypsin inhibition up to 0.25 mM for PDP-4, PDP-5, PDP-6, PDP-7 and PDP-11 as predicted based on sequence and structure.

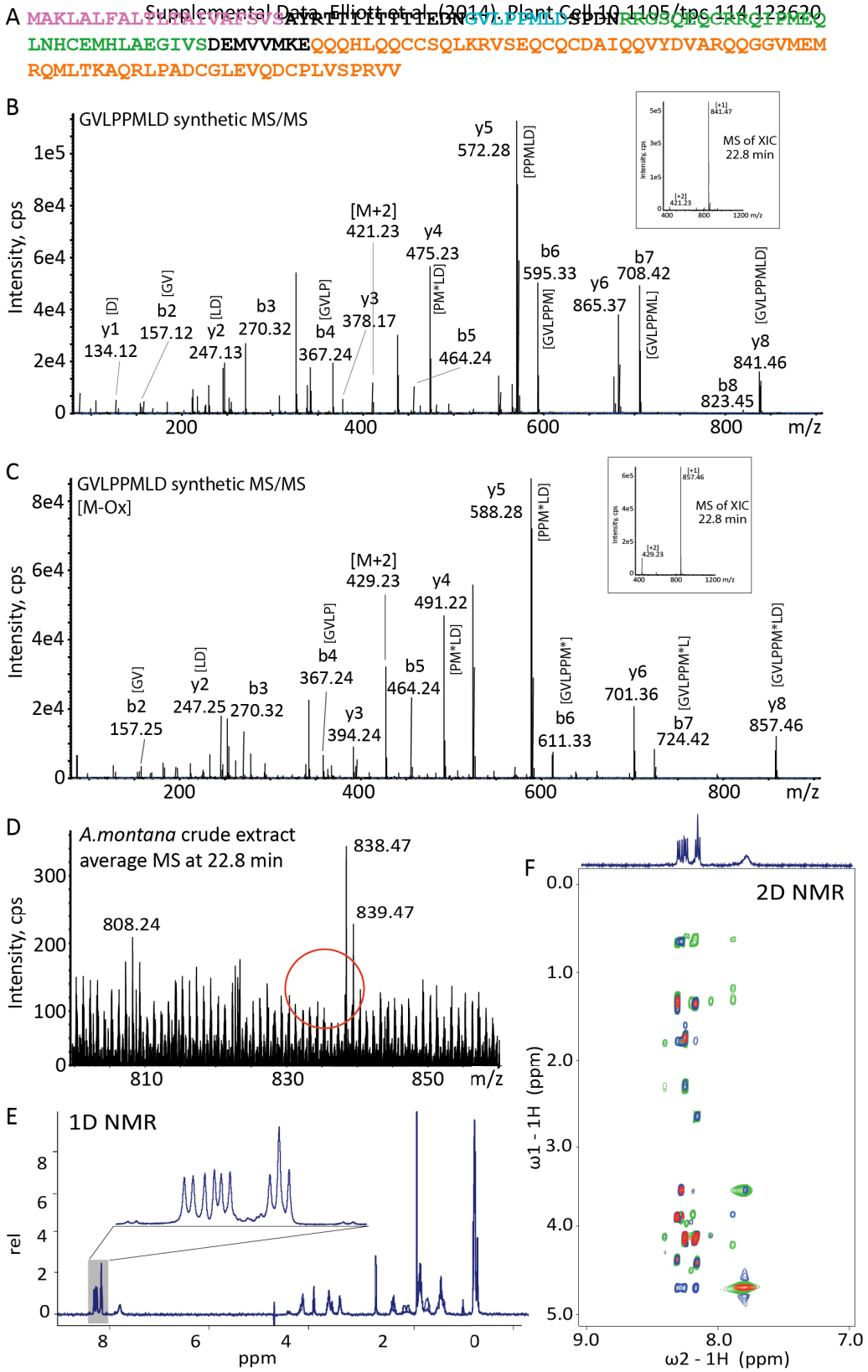


Supplemental Fig. 20: Inhibition of insect proteases. (**A**) Percentage inhibition of trypsin activity in a *Helicoverpa armigera* whole gut extract by various PDPs; (**B**) Percentage inhibition of chymotrypsin activity in a *H. armigera* whole gut extract by various PDPs. In both assays, SFTI-1 (blue), SFT-L1 (red) and PDP-4 (purple) were tested. For the trypsin assay leupeptin (green) and rT4 (black) were included as potent positive controls. In the chymotrypsin assay chymostatin (green) and StPin1A (black) were included as positive controls. In the presence of low concentrations of protease inhibitors there is sometimes an increase in activity compared to control samples that do not contain protease inhibitors. Therefore, values less than 0% were ignored for ease of viewing of the data. In cases where at high concentrations two points were co-located on the graph we used half circles instead of full circles; (c) Comparison of trypsin inhibition (TI) by SFTI-1 (aqua/blue) and the negative control SFT-L1 (red/orange) in assays with either bovine trypsin or insect (*H. armigera*) trypsin.


MAKLIILVVLAILAFVEVSVSGYKTSISTITIEDNGRCTKSIPPICFPDGLDNPRGCQIRIQQLNHCQMHLTSFDYKLRMAVENPKQQQHLSLCCNQLQEVEKQCQCEAIKQVVEQAQKQLQQGQGQQQVQQMVKKAQMLPNQCNLQCSI
36
35
34
38
46
49
51
50
52
53
84

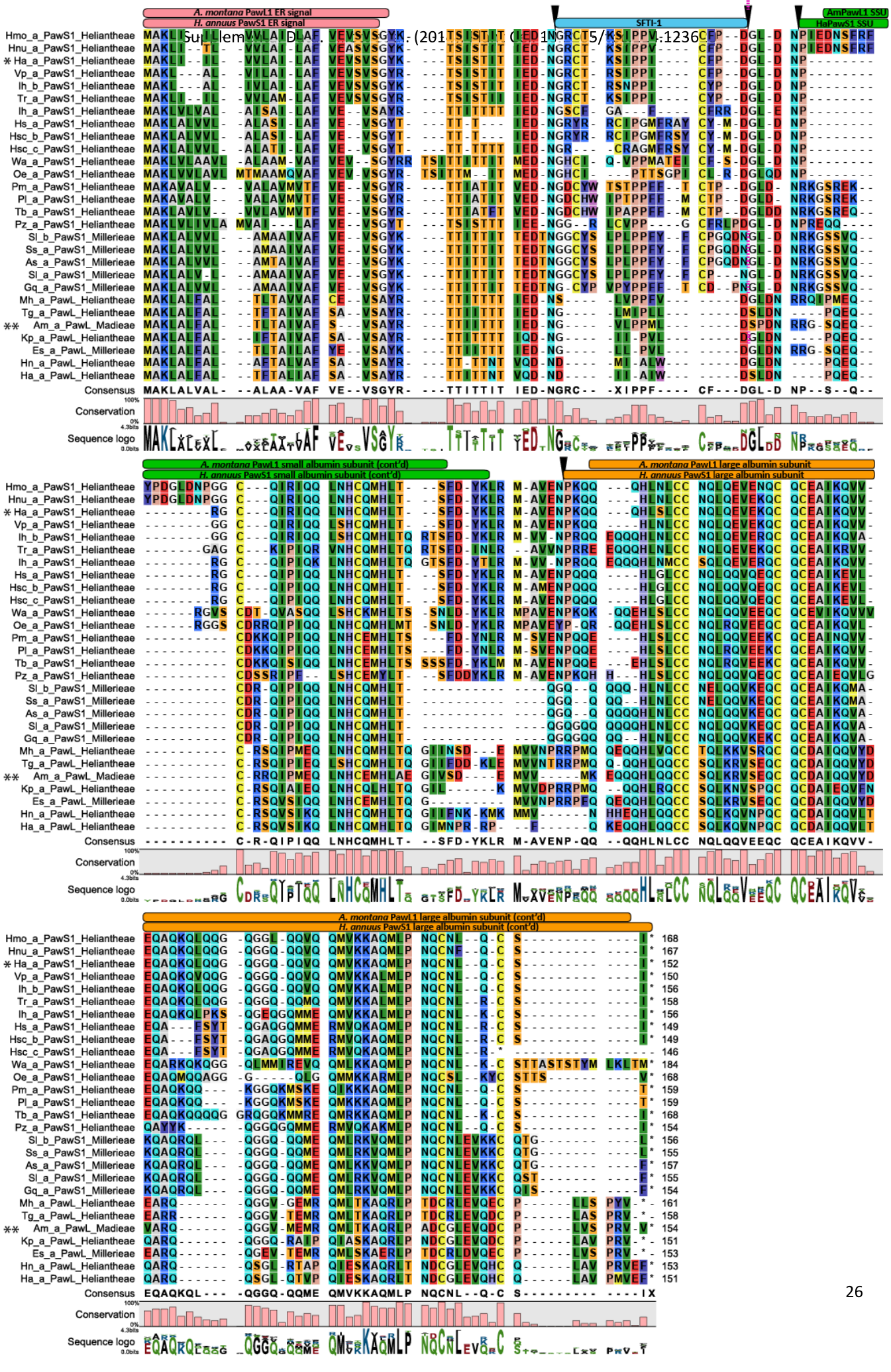

WT
N35A
D34A N35A
G36S
G36T
G36A
R37F
R37A
C38A C46A
T39A
K40A
S41A
I42A
P43A
P44A
I45A
F47A
P48A
D49A
L51M
L51E
L51F
L51H
L51W
L51R
sunflower
Col WT
seedless extract
no cyclic

Supplemental Fig. 21: *In vivo* processing of PawS1 mutants. MALDI analysis of seed extracts from transgenic *Arabidopsis* containing PawS1 mutants. For each construct, eight lines were extracted. The three MALDI spectra with the highest signal:noise were chosen for this Figure. In the event less than three extracts provided good spectra, an additional eight or 16 lines were extracted (i.e. a maximum of 24) before a negative or weak result was accepted. The observed masses (obs) were compared to theoretical masses (theor) of the peptide displayed. In peptide schematics, mutated residues or those not typically a part of SFTI-1 were marked in bold. Suspected sodium adducts (+22) were not labeled. Background masses of 1203 and 1380 from non-transgenic *Arabidopsis* were seen in some of the construct lines indicating low abundance of the transgenic peptides. Blue spectra are previously published data (Mylne et al., 2011). Black spectra are new data. A summary of these results is presented in Fig. 3*B*.

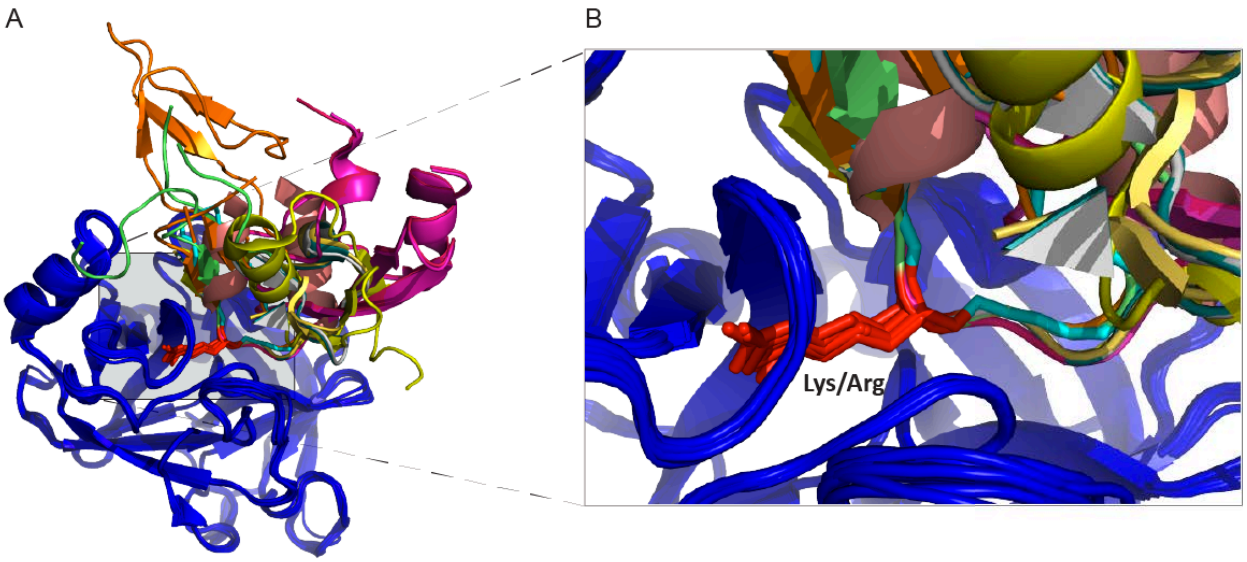


Supplemental Fig. 22: The peptide GVLPPMLD encoded by Am-*PawL1* is not detectable in peptide extracts of *Arnica montana* seeds, when considered as a cyclic (823.5 Da) or linear (841.5 Da) mass and when compared to the properties of a synthetic version of the peptide. (**A**) The protein sequence of Am-*PawL1* with the predicted peptide in blue; (**B**) The MS/MS triggered spectra of the synthetic GVLPPMLD linear peptide, which elutes on the LC at 22.8 min (inset); (**C**) The MS/MS triggered spectra of the synthetic GVLPPMLD linear peptide with an oxidized Met residue (M*) which elutes on the LC at

22.8 min (inset) with the non-modified peptide; (**D**) Averaged mass spectrum at 22.8 min of *A. montana* crude peptide extract displays no ions representative for GVLPPMLD. The linear mass of 841.47 would be found in the area indicated by the red circle. Although not shown, the LC-MS was also queried for the predicted cyclic mass (823.5 Da), cyclic with oxidation of the Met (839.5 Da) and amidation of cyclic and linear GVLPPMLD (822.5 and 840.5 Da); (**E**) 1D NMR spectrum of synthetic linear GVLPPMLD. The inset NH proton chemical shift region shows an averaging of peaks indicating GVLPPMLD has either no structure, or no specific structural conformation; (**F**) An overlay of the NH "fingerprint" region of the 2D NMR spectra TOCSY (green), NOESY (red) and ROESY (blue). The absence of six clear spin systems (one for each amino acid other than the two proline residues) in the TOCSY and the lack of expected inter-residue peaks support the fact that GVLPPMLD is unstructured or does not have a specific structure in the linear form, indicating that it would be susceptible to breakdown in the lytic biological environment of a seed.


A. montana PawL1 ER signal
H. annuus PawS1 ER signal
SFTI-1
AmPawL1 SSU
HePawS1 SSU
A. montana PawL1 small albumin subunit (cont'd)
H. annuus PawS1 small albumin subunit (cont'd)
A. montana PawL1 large albumin subunit
H. annuus PawS1 large albumin subunit
A. montana PawL1 large albumin subunit (cont'd)
H. annuus PawS1 large albumin subunit (cont'd)
Hmo_a_PawS1_Heliantheae
Hnu_a_PawS1_Heliantheae
* Ha_a_PawS1_Heliantheae
Vp_a_PawS1_Heliantheae
Ih_b_PawS1_Heliantheae
Tr_a_PawS1_Heliantheae
Ih_a_PawS1_Heliantheae
Hs_a_PawS1_Heliantheae
Hsc_b_PawS1_Heliantheae
Hsc_c_PawS1_Heliantheae
Wa_a_PawS1_Heliantheae
Oe_a_PawS1_Heliantheae
Pm_a_PawS1_Heliantheae
Pl_a_PawS1_Heliantheae
Tb_a_PawS1_Heliantheae
Pz_a_PawS1_Heliantheae
Sl_b_PawS1_Millerieae
Ss_a_PawS1_Millerieae
As_a_PawS1_Millerieae
Sl_a_PawS1_Millerieae
Gq_a_PawS1_Millerieae
Mh_a_PawL_Heliantheae
Tg_a_PawL_Heliantheae
** Am_a_PawL_Madieae
Kp_a_PawL_Heliantheae
Es_a_PawL_Millerieae
Hn_a_PawL_Heliantheae
Ha_a_PawL_Heliantheae
Consensus
Conservation
Sequence logo

Supplemental Fig. 23: Alignment of predicted protein sequences for *PawL1* and selected *PawS1* genes. Sequences were aligned by CLC Genomics Workbench 6.0.5 using its default settings and were not manually altered. The sequences were ordered by CLC Genomics based on similarity. Above the alignment, the domains for *H. annuus* PawS1 (*) and *Arnica montana* PawL1 (**) are displayed. Black arrows denote AEP-dependent cleavage sites for *H. annuus* PawS1 (experimental evidence is only lacking for the site preceding the characterised large albumin subunit). The expanded PDP region seen in some *PawS1* genes and *PawL1* genes was removed so the sequences could be aligned. The region where sequence was removed is marked (dotted magenta line) with the length of removed region as follows: Kp_a_PawL1 22 residues; Gq_a_PawS1 100 residues; As_a_PawS1 22 residues; Sl_b_PawS1 22 residues; Ss_a_PawS1 22 residues. This alignment shows the regions conserved between PawS1 and PawL1, the great sequence variation in the peptide region of all genes as well as the clear groupings of Heliantheae PawS1, Millereae PawS1 and the PawL1s from Heliantheae, Millereae and Madieae.



C

PDB	Active site residue	Family
1TAB	CSCTKSMPPKCRCSDIR...	BBI
2G81	CECTKSIPPQCRCSQVR...	BBI
1F2S	RICPRIWMECKRDSDCM...	Squash TI
1PPE	RVCPRILMECKKDSDCL...	Squash TI
2BTC	RVCPKILMECKKDSDCL...	Squash TI
1P2K	...PCVARIIRYFYNAKAGL...	Kazal
2FTL	...PCKARIIRYFYNAKAGL...	Kunitz
2F3C	...CACPRVLHRVCGSDGNT...	Kunitz
2CMY	...CYAQRSPELLRRCLDNC	*V hederifolia*
1SFI	GRCTKSIPPICFPD	SFTI-1

Supplemental Fig. 24: Structural similarity between trypsin inhibitors from a variety of inhibitor families (**A**) The X-ray crystallography structures of 10 trypsin inhibitors in complex with trypsin; comprising Bowman-Birk Inhibitors (PDB codes 1TAB in green, 2G81 in orange), squash trypsin inhibitors (PDB codes 1F2S in yellow,1PPE in aqua, 2BTC in white), Kunitz inhibitors (PDB codes 1P2K in magenta, 2FTL in pink), a Kazal-type serine protease (PDB code 2F3C in yellow), an inhibitor from *Veronica hederifolia* (PDB code 2CMY in peach) and SFTI-1 (PDB code 1SFI in cyan) are overlaid in ribbon form. The complexes are aligned based on the homology of trypsin and this also leads to a direct alignment of the active residue of the trypsin inhibitors, showing the characteristic similarity at the active Lys or Arg residue and the surrounding P3-P1 residues. Trypsin in each complex is blue, the active site residues are red for all inhibitors; (**B**) A zoomed in region surrounding the active site residues clearly shows the similarity of these diverse trypsin inhibitors, yet as seen in (**A**) apart from this two to three residue region there is very little similarity between the proteins; (**C**) an alignment of the sequences of the trypsin inhibitors further illustrates the similarity of the P1 residue and highlights the dis-similarity elsewhere in the sequences. Figure prepared with Pymol. Only the main chain is shown in cartoon format, except in the case of the (red) active inhibitory residue in each case where the side chain was shown in stick format.

Supplemental Tables

Supplemental Table 1: Range of predicted and confirmed peptides from *PawS1* genes. Information contained includes: name assigned to the peptide; predicted peptide sequence; species originally discovered in; subtribe and tribe of the species; predicted mass (Da); the observed LC retention time (min) of those confirmed and whether the peptide has been confirmed in-planta, gene abbreviation code and the GenBank code. The *in planta* peptide confirmation is based on retention time, mass, MS/MS ion fragmentation and where possible MS/MS sequencing of the peptide and/or comparison with a synthetically synthesized version of the peptide. (*) Denotes that *Helianthus mollis* was the species from which the gene was amplified but the peptide was confirmed in *Helianthus schweinitzii.* (#) Denotes that *Heliopsis scabra* was the species from which the gene was amplified but the peptide was confirmed in *Heliopsis helianthoides* (*Heliopsis scabra* is listed as a variety of *Heliopsis helianthoides* in the Flora North America treatment of the genus). Abbreviations: Conf. = confirmed *in planta*; Tribe H = Heliantheae, M = Millerieae; Subtribe H = Helianthinae, Z = Zinniinae, E = Ecliptinae, G = Galinsoginae. Backbone indicators, C = cyclic and A = acyclic, these are either confirmed (Supplemental Fig. 3 and Supplemental Fig. 7-14) or predicted based on sequence similarity to those that are confirmed.

Peptide	Sequence	Species Binomial	Tribe	Sub Tribe	Mass (Da)	Back-bone	Conf.	Gene code	GenBank
SFTI-1	GRCTKSIPPICFPD	*Helianthus annuus*	H	H	1512.7	C	Y	Ha	FJ469150
SFTI-1	GRCTKSIPPICFPD	*Helianthus exilis*	H	H	1512.7	C	Y	Hex	FJ749263
SFTI-1	GRCTKSIPPICFPD	*Helianthus porteri*	H	H	1512.7	C	Y	Hpo	JX262717
SFTI-1	GRCTKSIPPICFPD	*Helianthus praecox subsp. praecox*	H	H	1512.7	C	N	Hpr	JX262718
PDP-16	GRCTRSNPPICYPD	*Iostephane heterophylla*	H	H	1557.7	C	N	Ih_b	JX262741
PDP-3	GRCTKSIPPICYPD	*Tithonia rotundifolia*	H	H	1528.7	C	Y	Tr	JX262722
PDP-12	GRCTKSIPPVCFPD	*Helianthus mollis/ Helianthus schweinitzii* *	H	H	1498.7	C	Y	Hmo	JX262723
PDP-12	GRCTKSIPPVCFPD	*Helianthus nuttallii*	H	H	1498.7	C	Y	Hnu	JX262724
SFTI-1	GRCTKSIPPICFPD	*Helianthus tuberosus*	H	H	1512.7	C	N	Htu	FJ749265
SFTI-1	GRCTKSIPPICFPD	*Helianthus schweinitzii*	H	H	1512.7	C	Y	Hs_a	JX262719
SFTI-1	GRCTKSIPPICFPD	*Helianthus schweinitzii*	H	H	1512.7	C	Y	Hs_b	JX262720
SFTI-1	GRCTKSIPPICFPD	*Helianthus schweinitzii*	H	H	1512.7	C	Y	Hs_c	JX262721
PDP-15	GRCTRSIPPICFPD	*Aldama phenax*	H	H	1540.7	C	Y	Vp	JX262740
PDP-4	GSCFGAFCFRRD	*Iostephane heterophylla*	H	H	1344.6	C	Y	Ih_a	JX262725
PDP-5	GRYRRCIPGMFRAYCYMD	*Heliopsis scabra/Heliopsis helianthoides#*	H	Z	2237.0	C	Y	Hsc_a	JX262726
PDP-13	GRYRRCIPGMFRSYCYMD	*Heliopsis scabra/Heliopsis helianthoides#*	H	Z	2253.0	C	Y	Hsc_b	JX262727
PDP-14	GRCRAGMFRSYCYMD	*Heliopsis scabra/Heliopsis helianthoides#*	H	Z	1794.7	C	Y	Hsc_c	JX262728
PDP-7	GHCIPTTSGPICLRD	*Otopappus epaleaceus*	H	E	1548.7	C	N	Oe	JX262730
PDP-8	GGRLCVPPGCFRLPD	*Philactis zinnioides*	H	Z	1565.8	C	Y	Pz	JX262731
	GDCHWIPTPPFFMCTPD	*Perymenium jelskii*	H	E	1942.8	C	N	Pl	JX262733
PDP-17	GDCHWIPAPPFFMCTPD	*Tilesia baccata*	H	E	1912.8	C	N	Tb	JX262734
PDP-9	GDCYWTSTPPFFTCTPD	*Perymenium macranthus*	H	E	1916.8	C	N	Pm	JX262732
PDP-6	GHCIQVPPMATEICFSD	*Wamalchitamia aurantiaca*	H	E	1826.8	C	N	Wa	JX262729
PDP-19	GGCYSLPLPPFYFCPN	*Sabazia liebmannii*	M	G	1753.8	A	N	Sl_a	JX262735
PDP-10	GCYPVPYPPFFTCDPN	*Galinsoga quadriradiata*	M	G	1813.8	A	Y	Gq	JX262736
PDP-11	GCWPVPYPPFFDCKPN	*Galinsoga quadriradiata*	M	G	1863.8	A	Y	Gq	JX262736

	GGCYSLPLPPFYFCPGQD(N)	*Alloispermum scabrifolium*	M	G	1939.8	C	N	As	JX262737
		Sabazia liebmannii			or	or	N	Sl_b	JX262738
		Sabazia sarmentosa			2053.9	A	N	Ss	JX262739
PDP-18	GRCYPVPYPPFYTCTPD	*Sabazia liebmannii*	M	G	1954.9	C	N	Sl_b	JX262738
		Sabazia sarmentosa						Ss	JX262739
	GRCYPVPYPPFYTCTPH	*Alloispermum scabrifolium*	M	G	1994.9	A	N	As	JX262737

Supplemental Table 2: NMR structure statistics. Summary of restraints included in the CYANA calculations. Root-mean-square deviation values were calculated over the entire structure. Stereochemical quality was assessed via MolProbity (Chen et al., 2010). Clashscore is the number of steric overlaps > 0.4 Å per 10^3 atoms. The MolProbity score is defined as the following: 0.42574*log(1+clashscore) + 0.32996*log(1+max(0,pctRotOut-1)) + 0.24979*log(1+max(0,100-pctRamaFavored-2)) + 0.5. [a] Two restraints were used per hydrogen bond

	PDP-4	**PDP-5**	**PDP-6**	**PDP-7**	**PDP-11**
Experimental restraints					
Distance restraints					
Total NOE	101	263	174	118	184
Intra-residue	31	74	37	41	19
Inter-residue	70	189	137	77	165
Sequential (|*i-j*| = 1)	50	101	83	45	66
Medium range (|*i-j*| < 4)	15	32	21	14	56
Long range (|*i-j*| > 5)	5	56	33	18	43
Hydrogen-bond restraints[a]	4	5	5	5	5
Total Dihedral-angle restraints	20	31	33	27	28
ϕ	10	13	13	9	7
ψ	4	9	15	11	9
χ_1	6	9	3	7	12
Total number of restraints per residue	10.1	16.6	13.3	10.0	13.6
Target function, (Å)	0.07 ± 0.01	0.28 ± 0.005	0.57 ± 0.006	0.12± 0.03	0.73 ± 0.01
Root-mean-square deviation to mean coordinate structure[b], (Å)					
Backbone atoms	0.25 ± 0.17	0.09 ± 0.06	0.48 ± 0.20	0.33 ± 0.15	0.59 ± 0.21
All heavy atoms	1.46 ± 0.31	1.22 ± 0.19	1.18 ± 0.19	1.17 ± 0.28	1.12 ± 0.23
Stereochemical quality					
Residues in most favoured Ramachandran region, %	55.50	93.75	91.00	92.31	85.71
Residues in allowed Ramachandran region, %	44.50	6.25	9.00	7.69	14.29
Ramachandran outliers, %	0	0	0	0	0
Unfavourable sidechain rotamers, %	15.00	9.67	7.67	10.77	1.00
Cβ deviations > 0.25 Å	0	0	0	0	0
Clashscore, all atoms	0	0	5.37	0	4.07

	(100th percentile)	(100th percentile)	(91st percentile)	(100th percentile)	(96th percentile)
Overall MolProbity score	2.29 (59th percentile)	1.60 (91st percentile)	2.44 (51st percentile)	1.74 (87th percentile)	1.84 (84th percentile)

[a] Two restraints were used per hydrogen bond

[b] root-mean-square deviation was calculated among 20 refined structures

Supplemental Table 3: PDP physicochemical properties. PDPs are compared to one another and the previously described SFTI-1 (Korsinczky et al., 2001) and SFT-L1 (Mylne et al., 2011). From left to right; peptide name, number of residues, molecular weight (Daltons), the disulfide bond conformation, secondary structure, flexibility indicated by deuterium exchange rates in NMR studies (in PDP-5 and PDP-7 all amides are exchanged within ~1 hour), the number of hydrogen bonds in the structure, pI and the hydrophobicity indicated by the calculated GRAVY score. The bottom row indicates the range of each variable across the family of peptides to show the diversity. Abbreviations: SRH = Short right-handed hook; A1 = Anti-parallel β-sheet.

PDP	# residues	MW (Da)	S-S Type	SS	Flexible? (Y/N)	H-bonds	pI	GRAVY score
SFTI-1	14	1512.7	SRH	1A	N	6	8.06	-0.1
SFT-L1	12	1202.5	SRH	Turns	N	4	3.67	0.4
4	12	1344.6	SRH	Turns	N	4	8.07	0.1
5	18	2237.0	SRH	1A	Y	5	9.39	-0.6
6	17	1826.8	SRH	1A	N	5	4.35	0.3
7	15	1548.7	SRH	1A	Y	5	6.73	0.03
11	16	1863.8	SRH	1A	N	5	5.82	-0.4
Range	**12-18**	**1.20-2.24 kDa**	**SRH**	**1A 5/7**	**2/7**	**4-6**	**3.7-9.4**	**-0.6 - 0.4**

Supplemental Table 4: Summary of findings by Konarev et al. (2002) who screened various Asteraceae family members for trypsin inhibitors using in-gel trypsin inhibition assays. They found low molecular weight (MW) trypsin inhibitors (TIs) only in *Helianthus* species and *Tithonia*, but did not find low molecular weight trypsin inhibitors in the following:

Subfamily	Tribe	**Tested negative for low MW TIs**
Carduoideae		9 genera, 26 species
Cichorioideae		10 genera, 17 species
Asteroideae	Astereae	6 genera, 8 species
Asteroideae	Anthemideae	7 genera, 11 species
Asteroideae	Senecioneae	5 genera, 9 species
Asteroideae	Eupatorieae	2 genera, 4 species
Asteroideae	Tageteae	1 genus, 2 species
Asteroideae	Calendulae	2 genera, 2 species
Asteroideae	Inuleae	2 genera, 3 species

Supplemental Table 5: Output statistics of sequencing and assembly quality for the *de novo* transcriptome of *Helianthus annuus and Arnica montana* provided by Beijing Genomic Institute. Total RNA was extracted from mature dry seeds as described (Mylne et al., 2012).

Sequencing	***H. annuus***	***A. montana***
Raw reads	40,742,686	28,916,086
Clean reads	40,742,686	27,516,042
Clean nt	3,666,841,740	2,476,443,780
Av. read length	90	90
Q20	98.36%	98.37%

Assembly	***H. annuus***	***A. montana***
Contig number	161,498	137,471
Contig av. length	297	308
Contig N50	425	444
Unigene	81,344	73,281
Unigene av. length	649	630
Unigene N50	933	902

Supplemental Table 6: Models and output statistics of positive selection analysis. Column 1 lists the regions of PawS1 protein analysed; column 2 are the model types tested; column 3, the likelihood ratio to test the goodness-of-fit between models; column 4 are the p-values of each test, indicating significant results; column 5 lists the amino acid residues (by number and single letter identifier) that are under positive selection based on the NEB analysis; and column 6 lists the amino acid residues (by number and single letter identifier) that are under positive selection based on the BEB analysis. *No single rapidly evolving residue is contained within the peptide region.

	Models compared	**2 ΔlnL**	**p-value**	**Sites under selection (NEB)**	**Sites under selection (BEB)**
Full sequence*	M0 vs M3	128.13	0	5I 6I 7L 11I 100 Q	NA
	M1a vs M2a	15.25	0.0004	6I 11I	6I 11I
	M7 vs M8	24.18	0	5I 6I 7L 11I	5I 6I 7L 11I
Peptide Region	M0 vs M3	8.94	0.06	0	NA
	M1a vs M2a	0	1	0	0
	M7 vs M8	0.0001	0.9999	0	0
ER	M0 vs M3	71.25	0	5I 6I 7L 9L 11I	NA
	M1a vs M2a	20.91	0	5I 6I 7L 9L 11I	6I 11I
	M7 vs M8	19.87	0	5I 6I 7L 9L 11I	6I 7L 9L11I
Small Subunit	M0 vs M3	35.16	0.0001	2R	NA
	M1a vs M2a	11.28	0.004	2R	2R
	M7 vs M8	9.06	0.01	2R	2R
Large Subunit	M0 vs M3	21.73	0.0002	6S 49Q	NA
	M1a vs M2a	3.43	0.18	0	0
	M7 vs M8	5.26	0.06	49Q	49Q
Full Sequence	M0 vs M3	125.63	0	5I 6I 7L 11I 33R 96Q	NA

excluding peptide region	M1a vs M2a	18.51	0	5I 6I 7L 11I	6I 11I
	M7 vs M8	27.46	0	5I 6I 7L 11I	5I 6I 7L 11I
Spacer regions	M0 vs M3	24.62	0	4S 7I	NA
	M1a vs M2a	0.28	0.87	NO	NO
	M7 vs M8	6.15	0.05	4S 7I	NO

Supplemental Table 7: Newick trees used in PAML analyses of PawS1 regions. Numbers after colons are branch lengths (number of synonymous nucleotide substitutions).

PawS1 region	Newick tree file
Full sequence	(((((Ih_a: 0.406202, Ih_b: 0.035337): 0.040482, ((Vp: 0.088577, Tr: 0.189654): 0.000004, (((((Ha: 0.030722, Hex: 0.109555): 0.010260, Hpr: 0.072607): 0.000004, (Hs_a: 0.031106, Hs_b: 0.022479, Hs_c: 0.062257): 0.018391): 0.000004, ((Hmo: 0.030571, Hnu: 0.051848): 0.000004, Htu: 0.041194): 0.000004): 0.000004, Hpo: 0.061876): 0.104454): 0.000004): 0.190977, ((Hsc_a: 0.000004, Hsc_b: 0.000004, Hsc_c: 0.043805): 0.188723, Pz: 0.312587): 0.096077): 0.083100, ((((Pl: 0.048494, Pm: 0.038610): 0.188231, Wa: 0.427407): 0.000004, Tb: 0.155036): 0.119386, Oe: 0.286767): 0.089433): 0.069606, ((As: 0.051801, (Gq: 0.060487, ((Sl_a: 0.126999, Sl_b: 0.000004): 0.000004, Ss: 0.000004): 0.052074): 0.000004): 0.120550): 0.107068);
Peptide region	(((((Ih_a: 16.104820, Ih_b: 0.000004): 0.000004, ((Vp: 0.000004, Tr: 0.000004): 0.000004, ((((Ha: 0.000004, Hex: 0.000004): 0.000004, Hpr: 0.000004): 0.000004, (Hs_a: 0.000004, Hs_b: 0.000004, Hs_c: 0.000004): 0.000004): 0.000004, ((Hmo: 0.000004, Hnu: 0.000004): 0.000004, Htu: 0.000004): 0.000004): 0.000004, Hpo: 0.000004): 0.000004): 0.000004): 0.684550, ((Hsc_a: 0.000004, Hsc_b: 0.000004, Hsc_c: 0.000004): 10.716322, Pz: 0.000004): 0.000004): 1.173082, ((((Pl: 0.000004, Pm: 0.000004): 3.126557, Wa: 4.447389): 0.000004, Tb: 0.000004): 0.000004, Oe: 5.579020): 0.000004): 0.000004, ((As: 0.000004, (Gq: 0.390257, ((Sl_a: 0.000004, Sl_b: 0.000004): 0.000004, Ss: 0.000004): 0.000004): 0.000004): 0.000004);
ER	(((((Ih_a: 1.492485, Ih_b: 0.000004): 0.000004, ((Vp: 0.131209, Tr: 0.637085): 0.000004, (((((Ha: 0.000004, Hex: 0.551460): 0.000004, Hpr: 0.062253): 0.000004, (Hs_a: 0.063852, Hs_b: 0.065682, Hs_c: 0.000004): 0.063852): 0.000004, ((Hmo: 0.000004, Hnu: 0.208104): 0.000004, Htu: 0.199035): 0.000004): 0.000004, Hpo: 0.000004): 0.543857): 0.000004): 0.323236, ((Hsc_a: 0.000004, Hsc_b: 0.000004, Hsc_c: 0.208506): 0.787784, Pz: 0.356500): 0.039338): 0.357229, ((((Pl: 0.056717, Pm: 0.000004): 1.491969, Wa: 0.246640): 0.000004, Tb: 0.739294): 0.000004, Oe: 0.606119): 0.000004): 0.001429, ((As: 0.113873, (Gq: 0.000004, ((Sl_a: 0.189022, Sl_b: 0.000004): 0.000004, Ss: 0.000004): 0.000004): 0.000004): 0.078649): 0.134450);
Small Subunit	(((((Ih_a: 0.529281, Ih_b: 0.106789): 0.146907, ((Vp: 0.247533, Tr: 0.854691): 0.000004, (((((Ha: 0.000004, Hex: 0.000004): 0.110753, Hpr: 0.117336): 0.000004, (Hs_a: 0.000004, Hs_b: 0.123007, Hs_c: 0.110752): 0.000004): 0.000004, ((Hmo: 0.000004, Hnu: 0.000004): 0.000004, Htu: 0.000004): 0.000004): 0.000004, Hpo: 0.253427): 0.000004): 0.000004): 0.247923, ((Hsc_a: 0.000004, Hsc_b: 0.000004, Hsc_c: 0.000004): 0.000004, Pz: 0.966390): 0.000004): 0.000004, ((((Pl: 0.000004, Pm: 0.000004): 0.000004, Wa: 1.606497): 0.118210, Tb: 0.222670): 0.563123, Oe: 0.522274): 0.000004): 0.200966, ((As: 0.101113, (Gq: 0.097787, ((Sl_a: 0.341542, Sl_b: 0.000004): 0.000004, Ss: 0.000004): 0.236457): 0.000004): 0.154796): 0.132288);
Large Subunit	(((((Ih_a: 0.173866, Ih_b: 0.079020): 0.041295, ((Vp: 0.119536, Tr: 0.081655): 0.000004, (((((Ha: 0.057509, Hex: 0.056514): 0.000004, Hpr: 0.098793): 0.000004, (Hs_a: 0.038204, Hs_b: 0.000004, Hs_c: 0.098768): 0.037902): 0.000004, ((Hmo:

Full sequence excluding peptide region	0.057675, Hnu: 0.018905): 0.000004, Htu: 0.000004): 0.018758): 0.000004, Hpo: 0.037557): 0.076868): 0.000004): 0.167377, ((Hsc_a: 0.000004, Hsc_b: 0.000004, Hsc_c: 0.020670): 0.150708, Pz: 0.225497): 0.192773): 0.015575, ((((Pl: 0.084621, Pm: 0.061911): 0.085162, Wa: 0.175625): 0.000004, Tb: 0.148411): 0.055683, Oe: 0.255990): 0.161573): 0.079755, ((As: 0.036961, (Gq: 0.075327, ((Sl_a: 0.115023, Sl_b: 0.000004): 0.000004, Ss: 0.000004): 0.056356): 0.000004): 0.095884): 0.086637);
Spacer regions	(((((Ih_a: 0.406202, Ih_b: 0.035337): 0.040482, ((Vp: 0.088577, Tr: 0.189654): 0.000004, (((((Ha: 0.030722, Hex: 0.109555): 0.010260, Hpr: 0.072607): 0.000004, (Hs_a: 0.031106, Hs_b:0.022479, Hs_c: 0.062257): 0.018391): 0.000004, ((Hmo: 0.030571, Hnu: 0.051848): 0.000004, Htu: 0.041194): 0.000004): 0.000004, Hpo: 0.061876): 0.104454): 0.000004): 0.190977, ((Hsc_a: 0.000004, Hsc_b: 0.000004, Hsc_c: 0.043805): 0.188723, Pz: 0.312587): 0.096077): 0.083100, ((((Pl: 0.048494, Pm: 0.038610): 0.188231, Wa: 0.427407): 0.000004, Tb: 0.155036): 0.119386, Oe: 0.286767): 0.089433): 0.069606, ((As: 0.051801, (Gq: 0.060487, ((Sl_a: 0.126999, Sl_b: 0.000004): 0.000004, Ss: 0.000004): 0.052074): 0.000004): 0.120550): 0.107068);
	(((((Ih_a: 1.519348, Ih_b: 0.000004): 0.000004, ((Vp: 0.000004, Tr: 0.111322): 0.000004, (((((Ha: 0.000004, Hex: 0.108576): 0.000004, Hpr: 0.000004): 0.000004, (Hs_a: 0.000004, Hs_b: 0.000004, Hs_c: 0.000004): 0.000004): 0.000004, ((Hmo: 0.000004, Hnu: 0.000004): 0.000004, Htu: 0.000004): 0.000004): 0.000004, Hpo: 0.112750): 0.109180): 0.000004): 0.000004, ((Hsc_a: 0.000004, Hsc_b: 0.000004, Hsc_c: 0.000004): 0.331285, Pz: 0.239073): 0.000004): 0.102554, ((((Pl: 0.000004, Pm: 0.000004): 0.088704, Wa: 0.811711): 0.000004, Tb: 0.000004): 0.638492, Oe: 0.124041): 0.000004): 0.696506, ((As: 0.000004, (Gq: 0.060629, ((Sl_a: 0.000004, Sl_b: 0.000004): 0.000004, Ss: 0.000004): 0.000004): 0.000004): 0.759012): 0.725242);

Supplemental Table 8: Exact indel rates and extension probabilities of the different albumin protein regions for alignment in Supplemental Fig. 5 and using trees in Supplemental Table 7. Insertion deletion rates were calculated using ProtPal (Westesson et al., 2012) with fixed alignment and species tree and normalised by the synonymous substitution rates as computed with PAML.

	Alignment (a.a)	**Insertion rates**	**Deletion rates**
Full sequence	312	0.13	0.02
Peptide region	126	0.25	0.08
ER signal	26	0.06	0.02
Small Subunit	36	0.14	0.01
Large Subunit	80	0.14	0.004
Full sequence (excluding peptide region)	186	0.13	0.02
Spacer	44	0.09	0.03

Supplemental Table 9: Primers used in this study. Note that lower case in the sequence column denotes mismatches, either the mutagenic codon (JM508- JM531) or unintentional due to their design based on a *de novo* transcriptome assembly (AJ1, AJ3).

Primer	Sequence	Purpose
JM204	AGG TTT CTG TTT CTG GTT AC	BAC library screening
JM192	GAG CAT TGC AAG TTG CAT TGG	BAC library screening
JM218	CGG CCT TCC TAG CAT TTG T	BAC library screening

JM262	GCT TTT CCC GAT GGC CTG GAC AAC	Probe for BAC fingerprinting
JM73	AAA AGC GGC TTC CCA TCT	Probe for BAC fingerprinting
AE51	CCT CTT CCA CTT GTT GCA TTC	Clone *PawS1* from Asteraceae
AE54	AGT CAC ACG AGT GTG TGT TTT	Clone *PawS1* from Asteraceae
JM508	TTT CCC GAT GGC gag GAC AAC CCC CGA	PawS1 mutagenesis
JM509	TCG GGG GTT GTC ctc GCC ATC GGG AAA	PawS1 mutagenesis
JM510	TTT CCC GAT GGC atg GAC AAC CCC CGA	PawS1 mutagenesis
JM511	TCG GGG GTT GTC cat GCC ATC GGG AAA	PawS1 mutagenesis
JM512	TTT CCC GAT GGC agg GAC AAC CCC CGA	PawS1 mutagenesis
JM513	TCG GGG GTT GTC cct GCC ATC GGG AAA	PawS1 mutagenesis
JM514	TTT CCC GAT GGC tgg GAC AAC CCC CGA	PawS1 mutagenesis
JM515	TCG GGG GTT GTC cca GCC ATC GGG AAA	PawS1 mutagenesis
JM516	TTT CCC GAT GGC cat GAC AAC CCC CGA	PawS1 mutagenesis
JM517	TCG GGG GTT GTC atg GCC ATC GGG AAA	PawS1 mutagenesis
JM518	TTT CCC GAT GGC ttt GAC AAC CCC CGA	PawS1 mutagenesis
JM519	TCG GGG GTT GTC aaa GCC ATC GGG AAA	PawS1 mutagenesis
JM520	GAG GAC AAT GGC gct TGT ACT AAG TCG	PawS1 mutagenesis
JM521	CGA CTT AGT ACA agc GCC ATT GTC CTC	PawS1 mutagenesis
JM522	AAT GGC AGG TGT gct AAG TCG ATT CCC	PawS1 mutagenesis
JM523	GGG AAT CGA CTT agc ACA CCT GCC ATT	PawS1 mutagenesis
JM524	GGC AGG TGT ACT gct TCG ATT CCC CCG	PawS1 mutagenesis
JM525	CGG GGG AAT CGA agc AGT ACA CCT GCC	PawS1 mutagenesis
JM526	AGG TGT ACT AAG gct ATT CCC CCG ATT	PawS1 mutagenesis
JM527	AAT CGG GGG AAT agc CTT AGT ACA CCT	PawS1 mutagenesis
JM528	TGT ACT AAG TCG gct CCC CCG ATT TGT	PawS1 mutagenesis
JM529	ACA AAT CGG GGG agc CGA CTT AGT ACA	PawS1 mutagenesis
JM530	ACT AAG TCG ATT gct CCG ATT TGT TTT	PawS1 mutagenesis
JM531	AAA ACA AAT CGG agc AAT CGA CTT AGT	PawS1 mutagenesis
CD3	ATG GCA AAA CTT GCA CTT	*A. montana PawL1* RT-PCR
CD4	TTA AAC GAC CCT TGG GCT	*A. montana PawL1* RT-PCR
AJ1	AgT TGC ACT TTT TGC CaT CA	*H. annuus PawL1* 3′ RACE
AJ3	GgG AGC CTC TGA GCC TTA CT	*H. annuus PawL1* 5′ RACE
AJ5	AAG CAG TGG TAT CAA CGC AGA G	Clone *H. annuus PawL1*
AJ6	AAA GGA AGC ATA ATG AGA TCA ATA CA	Clone *H. annuus PawL1*

Supplemental Table 10: ESI-ToF-MS/MS product ions for endo-GluC and trypsin fragments which correspond to AmPawL1. The table lists the b and y ions for the 1+, 2+ and 3+ ions. Observed masses are highlighted in green and marked on Fig. 4*C*. (**a**) endo-GluC Frag 1; (**b**) endo-GluC Frag 2; (**c**) endo-GluC Frag 3; (**d**) trypsin Frag 2; (**e**) trypsin Frag 1.

a endo-GluC Frag 1

Sequence	#	b^{1+}	b^{2+}	b^{3+}	y^{1+}	y^{2+}	y^{3+}	#
E (CAM)	1	187.1	94.0	63.0	2343.0	1172.0	781.7	18
Q	2	315.1	158.1	105.7	2157.0	1079.0	719.7	17
Q	3	443.2	222.1	148.4	2028.9	1015.0	677.0	16
Q	4	571.2	286.1	191.1	1900.9	950.9	634.3	15
H	5	708.3	354.6	236.8	1772.8	886.9	591.6	14
L	6	821.4	411.2	274.5	1635.8	818.4	545.9	13
Q	7	949.4	475.2	317.1	1522.7	761.8	508.2	12
Q	8	1077.5	539.2	359.8	1394.6	697.8	465.5	11
C (CAM)	9	1237.5	619.3	413.2	1266.6	633.8	422.9	10
C (CAM)	10	1397.5	699.3	466.5	1106.5	553.8	369.5	9
S	11	1484.5	742.8	495.5	946.5	473.8	316.2	8
Q	12	1612.6	806.8	538.2	859.5	430.3	287.2	7
L	13	1725.7	863.3	575.9	731.4	366.2	244.5	6
K	14	1853.8	927.4	618.6	618.4	309.7	206.8	5
R	15	2009.9	1005.4	670.6	490.3	245.6	164.1	4
V	16	2108.9	1055.0	703.7	334.2	167.6	112.1	3
S	17	2196.0	1098.5	732.7	235.1	118.1	79.0	2
E	18	2325.0	1163.0	775.7	148.1	74.5	50.0	1

b endo-GluC Frag 2

Sequence	#	b^{1+}	b^{2+}	b^{3+}	y^{1+}	y^{2+}	y^{3+}	#
Q	1	129.1	65.0	43.7	1527.6	764.3	509.9	12
C (CAM)	2	289.1	145.0	97.0	1399.5	700.3	467.2	11
Q	3	417.1	209.1	139.7	1239.5	620.3	413.8	10
C (CAM)	4	577.1	289.1	193.1	1111.5	556.2	371.2	9
D	5	692.2	346.6	231.4	951.4	476.2	317.8	8
A	6	763.2	382.1	255.1	836.4	418.7	279.5	7
I	7	876.3	438.6	292.8	765.4	383.2	255.8	6
Q	8	1004.4	502.7	335.5	652.3	326.7	218.1	5
Q	9	1132.4	566.7	378.1	524.2	262.6	175.4	4
V	10	1231.5	616.2	411.2	396.2	198.6	132.7	3
Y	11	1394.5	697.8	465.5	297.1	149.1	99.7	2
D	12	1509.6	755.3	503.9	134.0	67.5	45.4	1

c endo-GluC Frag 3

Sequence	#	b^{1+}	b^{2+}	b^{3+}	y^{1+}	y^{2+}	y^{3+}	#
V	1	100.1	50.5	34.0	1170.6	585.8	390.9	10
Q	2	228.1	114.6	76.7	1071.5	536.3	357.8	9
D	3	343.2	172.1	115.1	943.4	472.2	315.2	8
C (CAM)	4	503.2	252.1	168.4	828.4	414.7	276.8	7
P	5	600.2	300.6	200.7	668.4	334.7	223.5	6
L	6	713.3	357.2	238.4	571.4	286.2	191.1	5
V	7	812.4	406.7	271.5	458.3	229.6	153.4	4
S	8	899.4	450.2	300.5	359.2	180.1	120.4	3
P	9	996.5	498.7	332.8	272.2	136.6	91.4	2
R	10	1152.6	576.8	384.9	175.1	88.1	59.0	1

d trypsin Frag 2

Sequence	#	b^{1+}	b^{2+}	b^{3+}	y^{1+}	y^{2+}	y^{3+}	#
L	1	114.1	57.5	38.7	2025.9	1013.5	676.0	18
P	2	211.1	106.1	71.1	1912.8	956.9	638.3	17
A	3	282.2	141.6	94.7	1815.8	908.4	605.9	16
D	4	397.2	199.1	133.1	1744.8	872.9	582.3	15
C (CAM)	5	557.2	279.1	186.4	1629.7	815.4	543.9	14
G	6	614.2	307.6	205.4	1469.7	735.4	490.6	13
L	7	727.3	364.2	243.1	1412.7	706.9	471.6	12
E	8	856.4	428.7	286.1	1299.6	650.3	433.9	11
V	9	955.4	478.2	319.2	1170.6	585.8	390.9	10
Q	10	1083.5	542.3	361.8	1071.5	536.3	357.8	9
D	11	1198.5	599.8	400.2	943.4	472.2	315.2	8
C (CAM)	12	1358.5	679.8	453.5	828.4	414.7	276.8	7
P	13	1455.6	728.3	485.9	668.4	334.7	223.5	6
L	14	1568.7	784.8	523.6	571.4	286.2	191.1	5
V	15	1667.7	834.4	556.6	458.3	229.6	153.4	4
S	16	1754.8	877.9	585.6	359.2	180.1	120.4	3
P	17	1851.8	926.4	617.9	272.2	136.6	91.4	2
R	18	2007.9	1004.5	670.0	175.1	88.1	59.0	1

e trypsin Frag 1

Sequence	#	b^{1+}	b^{2+}		y^{1+}	y^{2+}		#
E	1	130.0	65.5		1814.8	907.9		14
Q	2	258.1	129.6		1685.8	843.4		13
Q	3	386.2	193.6		1557.7	779.4		12
Q	4	514.2	257.6		1429.7	715.3		11
H	5	651.3	326.1		1301.6	651.3		10
L	6	764.4	382.7		1164.6	582.8		9
Q	7	892.4	446.7		1051.5	526.2		8
Q	8	1020.5	510.7		923.4	462.2		7
C (CAM)	9	1180.5	590.8		795.3	398.2		6
C (CAM)	10	1340.5	670.8		635.3	318.2		5
S	11	1427.6	714.3		475.3	238.1		4
Q	12	1555.6	778.3		388.3	194.6		3
L	13	1668.7	834.9		260.2	130.6		2
K	14	1796.8	898.9		147.1	74.1		1

a

Sequence	#	b^{1+}	b^{2+}	b^{3+}	y^{1+}	y^{2+}	y^{3+}	#
S	1	88.0	44.5	30.0	3280.5	1640.7	1094.2	27
Q	2	216.1	108.6	72.7	3193.4	1597.2	1065.1	26
E	3	345.1	173.1	115.7	3065.4	1533.2	1022.5	25
Q	4	473.2	237.1	158.4	2936.3	1468.7	979.4	24
C (CAM)	5	633.2	317.1	211.7	2808.3	1404.6	936.8	23
R	6	789.3	395.2	263.8	2648.3	1324.6	883.4	22
R	7	945.4	473.2	315.8	2492.2	1246.6	831.4	21
Q	8	1073.5	537.2	358.5	2336.1	1168.5	779.4	20
I	9	1186.6	593.8	396.2	2208.0	1104.5	736.7	19
P	10	1283.6	642.3	428.5	2094.9	1048.0	699.0	18
M	11	1414.6	707.8	472.2	1997.9	999.4	666.6	17
E	12	1543.7	772.3	515.2	1866.8	933.9	622.9	16
Q	13	1671.7	836.4	557.9	1737.8	869.4	579.9	15
L	14	1784.8	892.9	595.6	1609.7	805.4	537.2	14
N	15	1898.9	949.9	633.6	1496.6	748.8	499.6	13
H	16	2035.9	1018.5	679.3	1382.6	691.8	461.5	12
C (CAM)	17	2195.9	1098.5	732.7	1245.5	623.3	415.9	11
E	18	2325.0	1163.0	775.7	1085.5	543.3	362.5	10
M	19	2456.0	1228.5	819.3	956.5	478.7	319.5	9
H	20	2593.1	1297.0	865.0	825.4	413.2	275.8	8
L	21	2706.2	1353.6	902.7	688.4	344.7	230.1	7
A	22	2777.2	1389.1	926.4	575.3	288.2	192.4	6
E	23	2906.2	1453.6	969.4	504.3	252.6	168.8	5
G	24	2963.3	1482.1	988.4	375.2	188.1	125.7	4
I	25	3076.4	1538.7	1026.1	318.2	159.6	106.7	3
V	26	3175.4	1588.2	1059.1	205.1	103.1	69.0	2
S	27	3262.5	1631.7	1088.2	106.1	53.5	36.0	1

b

Seq	#	b^{1+}	b^{2+}	b^{3+}	y^{1+}	y^{2+}	y^{3+}	#
G	1	58.0	29.5	20.0	3337.5	1669.2	1113.2	28
S	2	145.1	73.0	49.0	3280.5	1640.7	1094.2	27
Q	3	273.1	137.1	91.7	3193.4	1597.2	1065.1	26
E	4	402.2	201.6	134.7	3065.4	1533.2	1022.5	25
Q	5	530.2	265.6	177.4	2936.3	1468.7	979.4	24
C (CAM)	6	690.2	345.6	230.7	2808.3	1404.6	936.8	23
R	7	846.3	423.7	282.8	2648.3	1324.6	883.4	22
R	8	1002.4	501.7	334.8	2492.2	1246.6	831.4	21
Q	9	1130.5	565.7	377.5	2336.1	1168.5	779.4	20
I	10	1243.6	622.3	415.2	2208.0	1104.5	736.7	19
P	11	1340.6	670.8	447.5	2094.9	1048.0	699.0	18
M	12	1471.7	736.3	491.2	1997.9	999.4	666.6	17
E	13	1600.7	800.9	534.2	1866.8	933.9	622.9	16
Q	14	1728.8	864.9	576.9	1737.8	869.4	579.9	15
L	15	1841.9	921.4	614.6	1609.7	805.4	537.2	14
N	16	1955.9	978.5	652.6	1496.6	748.8	499.6	13
H	17	2093.0	1047.0	698.3	1382.6	691.8	461.5	12
C (CAM)	18	2253.0	1127.0	751.7	1245.5	623.3	415.9	11
E	19	2382.0	1191.5	794.7	1085.5	543.3	362.5	10
M	20	2513.0	1257.0	838.4	956.5	478.7	319.5	9
H	21	2650.1	1325.6	884.0	825.4	413.2	275.8	8
L	22	2763.2	1382.1	921.7	688.4	344.7	230.1	7
A	23	2834.2	1417.6	945.4	575.3	288.2	192.4	6
E	24	2963.3	1482.1	988.4	504.3	252.6	168.8	5
G	25	3020.3	1510.6	1007.4	375.2	188.1	125.7	4
I	26	3133.4	1567.2	1045.1	318.2	159.6	106.7	3
V	27	3232.4	1616.7	1078.2	205.1	103.1	69.0	2
S	28	3319.5	1660.2	1107.2	106.1	53.5	36.0	1

Supplemental Table 11: ESI-ToF-MS/MS product ions for the AmPawL1 small sub-unit (SSU). The table lists the b and y ions for the 1+ 2+ and 3+ ions. Observed masses are highlighted in green, see Supplemental Fig. 16 for labeled MS/MS spectra. (**a**) SSU1; (**b**) SSU2.

Supplemental Methods

Searching Heliantheae GenBank entries for BBIs

To test whether BBIs or BBI-like gene had been discovered in sunflower or its close relatives we BLAST searched GenBank (20-May-2013, nr database, limited to Heliantheae taxid:102814) with the amino acid sequence of a soyabean BBI (XP_003533609) and the only hits were to PawS1 as shown in (a) below. Similar use of rice BBI (EAZ10360) also produced hits to PawS1 as shown in (b) below with the next most significant alignment being a weak, non-BBI hit as shown in (c) below. BLAST analysis of GenBank (nr database, limited to Heliantheae taxid:102814) with CTKSIPPIC and excluding PawS1/SFTI-1 hits could only find a low significance hit in *Echinacea pallida* cellulose synthase shown in (d) below.

```
(a)    Sbjct Sequence ID: gb|ACS74805.1| preproalbumin PawS1 [Helianthus annuus]
       Query  39  IAGDNYNLKSTTSACCDACACTKSIPPIC  67
                  ++   Y    +T    D   CTKSIPPIC
       Sbjct  18  VSVSGYKTSISTITIEDNGRCTKSIPPIC  46

(b)    Sbjct Sequence ID: gb|ACS74805.1| preproalbumin PawS1 [Helianthus annuus]
       Query  101  DNTTCTKSIPPIC  113
                   DN  CTKSIPPIC
       Sbjct  34   DNGRCTKSIPPIC  46

(c)    Sbjct Sequence ID: emb|CBK62698.1| Art v 1 precursor [Ambrosia artemisiifolia]
       Query  6   ILLFLLAVGGLAAAHGDTIRLPSEGDAPPQPAKPWDCCD  44
                  +L+F+LA+  +A+  G     PS   +     K  D CD
       Sbjct  5   LLVFVLAISEIASVKGKLCEKPSVTWSGKCKVKQTDKCD  43

(d)    Sbjct Sequence ID: gb|ACA05345.1| cellulose synthase [Echinacea pallida]
       Query  1    CTKSIPPIC  9
                   CT   PPIC
       Sbjct  187  CT--LPPIC  193
```

References

Bouzidi, M., Franchel, J., Tao, Q., Stormo, K., Mraz, A., Nicolas, P., and Mouzeyar, S. (2006). A sunflower BAC library suitable for PCR screening and physical mapping of targeted genomic regions. Theor. Appl. Genet. **113,** 81-89.

Chen, V.B., Arendall, W.B., III, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D Biol. Crystallogr. **66,** 12-21.

Konarev, A.V., Anisimova, I.N., Gavrilova, V.A., Vachrusheva, T.E., Konechnaya, G.Y., Lewis, M., and Shewry, P.R. (2002). Serine proteinase inhibitors in the Compositae: distribution, polymorphism and properties. Phytochemistry **59,** 279-291.

Korsinczky, M.L., Schirra, H.J., Rosengren, K.J., West, J., Condie, B.A., Otvos, L., Anderson, M.A., and Craik, D.J. (2001). Solution structures by ^{1}H NMR of the novel cyclic trypsin inhibitor SFTI-1 from sunflower seeds and an acyclic permutant. J. Mol. Biol. **311,** 579-591.

Luckett, S., Garcia, R.S., Barker, J.J., Konarev, A.V., Shewry, P.R., Clarke, A.R., and Brady, R.L. (1999). High-resolution structure of a potent, cyclic proteinase inhibitor from sunflower seeds. J. Mol. Biol. **290,** 525-533.

Mylne, J.S., Colgrave, M.L., Daly, N.L., Chanson, A.H., Elliott, A.G., McCallum, E.J., Jones, A., and Craik, D.J. (2011). Albumins and their processing machinery are hijacked for cyclic peptides in sunflower. Nat. Chem. Biol. **7,** 257-259.

Mylne, J.S., Chan, L.Y., Chanson, A.H., Daly, N.L., Schaefer, H., Bailey, T.L., Nguyencong, P., Cascales, L., and Craik, D.J. (2012). Cyclic peptides arising by evolutionary parallelism via asparaginyl-endopeptidase-mediated biosynthesis. Plant Cell **24,** 2765-2778.

Westesson, O., Lunter, G., Paten, B., and Holmes, I. (2012). Accurate Reconstruction of Insertion-Deletion Histories by Statistical Phylogenetics. PLoS ONE **7,** e34572.

Wishart, D.S., Bigam, C.G., Holm, A., Hodges, R.S., and Sykes, B.D. (1995). ^{1}H, ^{13}C and ^{15}N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. J. Biomol. NMR **5,** 67-81.