Planta (2022) 256:93
https://doi.org/10.1007/s00425-022-03998-w

ORIGINAL ARTICLE

Genetic variation and structural diversity in major seed proteins among and within *Camelina* species

Dwayne Hegedus[1,2] · Cathy Coutu[1] · Branimir Gjetvaj[1] · Abdelali Hannoufa[3] · Myrtle Harrington[1] · Sara Martin[3] · Isobel A. P. Parkin[1] · Suneru Perera[1,2] · Janitha Wanasundara[1,2]

Received: 8 April 2022 / Accepted: 12 September 2022 / Published online: 6 October 2022

Abstract

Main conclusion **Genetic variation in seed protein composition, seed protein gene expression and predictions of seed protein physiochemical properties were documented in *C. sativa* and other *Camelina* species.**

Abstract Seed protein diversity was examined in six *Camelina* species (*C. hispida, C. laxa*, *C. microcarpa*, *C. neglecta*, *C. rumelica* and *C. sativa*). Differences were observed in seed protein electrophoretic profiles, total seed protein content and amino acid composition between the species. Genes encoding major seed proteins (cruciferins, napins, oleosins and vicilins) were catalogued for *C. sativa* and RNA-Seq analysis established the expression patterns of these and other genes in developing seed from anthesis through to maturation. Examination of 187 *C. sativa* accessions revealed limited variation in seed protein electrophoretic profiles, though sufficient to group the majority into classes based on high MW protein profiles corresponding to the cruciferin region. *C. sativa* possessed four distinct types of cruciferins, named CsCRA, CsCRB, CsCRC and CsCRD, which corresponded to orthologues in *Arabidopsis thaliana* with members of each type encoded by homeologous genes on the three *C. sativa* sub-genomes. Total protein content and amino acid composition varied only slightly; however, RNA-Seq analysis revealed that *CsCRA* and *CsCRB* genes contributed > 95% of the cruciferin transcripts in most lines, whereas *CsCRC* genes were the most highly expressed cruciferin genes in others, including the type cultivar DH55. This was confirmed by proteomics analyses. Cruciferin is the most abundant seed protein and contributes the most to functionality. Modelling of the *C. sativa* cruciferins indicated that each type possesses different physiochemical attributes that were predicted to impart unique functional properties. As such, opportunities exist to create *C. sativa* cultivars with seed protein profiles tailored to specific technical applications.

Keywords *Camelina sativa* · Cruciferin · Gene expression · Protein functionality · Protein modelling

Abbreviations

daa	Days after anthesis
G1	Sub-genome I
G2	Sub-genome II
G3	Sub-genome III
HVR	Hypervariable region
IA	Intrachain disulphide bond-containing
IE	Interchain disulphide bond-containing
PGRC	Plant Gene Resources Center
RMSD	Root mean square difference

Communicated by Dorothea Bartels.

✉ Dwayne Hegedus
Dwayne.Hegedus@agr.gc.ca

[1] Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK S7N 0X2, Canada

[2] Department of Food and Bioproduct Sciences, University of Saskatchewan, Saskatoon, SK, Canada

[3] Agriculture and Agri-Food Canada, London, ON, Canada

Introduction

Interest in *Camelina sativa* (camelina), grown in Europe in medieval times for food and fuel, stems from the need to diversify annual crop rotation portfolios with those that have smaller environmental footprints and the potential to produce valuable secondary products. It is compatible with practices used to produce contemporary oilseed crops, such as canola/oilseed rape and soybean, can be grown on marginal lands with fewer inputs and has higher tolerance to

drought and cold (Vollman et al. 1996). It is also naturally resistant to several diseases (Sharma et al. 2002; Eynck et al. 2012) and insects (Deng et al. 2002; Henderson et al. 2004; Soroka et al. 2015) that afflict canola.

Camelina sativa seed comprises approximately 36–47% oil (Moser 2012) and 43% protein (Zubr 2003). While it is being aggressively marketed as a diesel and aviation fuel feedstock (Li and Mupondwa 2014), the high levels of polyunsaturated fatty acids, in particular α-linolenic acid (38% total fatty acid), make it an attractive source of ω3 fatty acids in food and feed. α-linolenic acid is the precursor for the essential long chain polyunsaturated fatty acids eicosapentanoic acid (20:5 ω3) and docosahexanoic acid (22:6 ω3) that have human health benefits. Farmed fish species, such as salmon and cod, can convert α-linolenic acid to these longer chain polyunsaturated fatty acids when camelina oil is substituted for fish oil in their diet (Hixson et al. 2014; Hixson and Parrish 2014). This was attributed to the induction of two genes encoding fatty acyl elongases in the livers of fish fed diets containing only camelina oil (Xue et al. 2014, 2015). Other studies have reported increased ω3 fatty acid levels in chicken meat (Ariza et al. 2010) and eggs (Kakani et al. 2012), as well as in milk (Szumacher-Strabel et al. 2011) when camelina meal is incorporated into the diet at fairly low levels. Complete replacement of fish oil with camelina oil in farmed fish diets seems possible as this has no impact on weight gain, fillet sensory quality (Hixson et al. 2014) or the ability to mount an immune response (Booman et al. 2014), though some differences in tissue lipid composition (Hixson et al. 2014) and intestinal function (Morias et al. 2012) have been noted.

Camelina meal is also being considered as a protein source in farmed fish, poultry and livestock. Atlantic cod (*Gadus morhua*) tolerated up to 24% inclusion of camelina meal in place of fish meal in their diets without affecting weight gain (Hixson et al. 2016a). Salmonids were more sensitive to fish meal replacement and tolerated up to 5% (Atlantic salmon, *Salmo salar*) (Hixson et al. 2016b) and 14% (rainbow trout, *Oncorhynchus mykiss*) (Ye et al. 2016) camelina meal in their diets without ill effects. In cod, high inclusion rates were associated with increased expression of appetite-stimulating hormones and decreased expression of appetite-suppressing hormones indicating that the meal is affecting nutritional quality or palatability (Tuziak et al. 2014). In broiler chickens, low energy and nitrogen utilisation from camelina-based meals was attributed to high jejunal digesta viscosity, likely due to high levels of seed coat mucilage remaining in the meal, and to the presence of glucosinolates which can affect palatability (Pekel et al. 2015). Conversely, in growing pigs the ileal digestibility of crude protein from camelina expeller cake was only slightly less than the comparable canola product and was recommended for use in swine diets (Almeida et al. 2013). In cattle, the amount of undegraded protein in the rumen differed among meals from ten camelina genotypes (Colombini et al. 2014), but was generally higher than for canola meal. With the exception of glucosinolates, the levels of anti-nutritional factors including phytic acid, condensed tannins and sinapine, were lower in the camelina meals than canola meal. The essential amino acid composition of camelina meal is comparable to that from canola, soybean and flax meals (Zubr 2003); however, differences in amino acid profiles among camelina lines have been reported (Colombini et al. 2014). Of particular significance are the essential amino acids lysine and methionine as they cannot be synthesised de novo by animals and must be provided in the diet, though methionine can be converted to cysteine. Both are limiting in plant-based diets, most notably in cereals and some legumes (Ufaz and Galili 2008), and are added as supplements to feeds at a significant cost to fish (Wilson and Halver 1986), poultry (Kidd et al. 1998) and swine (Brinegar et al. 1950) production.

Camelina breeding is still in its infancy, but release of the camelina genome (Kagale et al. 2014) and transcriptome data (Liang et al. 2013; Nguyen et al. 2013; Mudalkar et al. 2014; Kagale et al. 2016) will facilitate rapid advances in crop improvement. As in other Brassicaceae, the major seed proteins in camelina are of the 2S albumins (napin) and 12S globulins (cruciferin), with transcript data indicating that there are 8 and 17 expressed members of these gene families, respectively, in *C. sativa* cv. Sunesson (Nguyen et al. 2013). The napin dimer possesses four disulphide bonds and consequently these proteins are rich in cysteine, while cruciferins tend to have higher levels of lysine. Oleosins are amphiphilic proteins with well-separated hydrophilic and hydrophobic domains; an attribute that allows them to interact with both lipid and water. While less abundant than cruciferin or napin, they play a major role in seed lipid accumulation and stabilisation of oil bodies, as well as other aspects of plant development (D'Andrea 2016). Manipulation of amino acid levels is possible through mutation (Kita et al. 2010; Marsolais et al. 2010) or down-regulation (Schmidt et al. 2011) of the major seed storage protein genes.

To date, there has been no broad examination of *C. sativa* seed protein or seed amino acid content diversity. To this end, we established seed protein profiles for six *Camelina* species and 187 *C. sativa* accessions from a global diversity collection held at the Plant Gene Resources Center for Canada (pgrc.agr.gc.ca). Amino acid content was determined for representatives from each major seed protein profile group and transcriptomic analysis was conducted to catalogue the expressed seed protein genes from the most diverse lines. These studies established that there is potential to select or engineer *C. sativa* lines with altered seed protein and/or amino acid profiles that may be more useful in food/feed or technical applications.

Materials and methods

Plant materials

A list of Camelina species, accessions and their source is provided in Suppl. Table S1a. Another 187 *C. sativa* accessions were obtained from PGRC (Agriculture & Agri-Food Canada, Saskatoon) (Suppl. Table S1b). *C. sativa* DH55 is a doubled haploid line for which the genome sequence is available (Kagale et al. 2014).

Seed protein extraction and separation

Seeds of *C. hispida* var. *hispida*, *C. hispida* var. *grandiflora, C. sativa*, *C. laxa*, *C. neglecta*, *C. microcarpa* (4x and 6x) and *C. rumelica* were generated at the Agriculture and Agri-Food Canada, Ottawa Research and Development Centre under controlled conditions within a growth chamber with randomised individual position and re-randomisation of position every two weeks. Self-incompatible taxa were hand pollinated to induce seed set. Seeds of *C. sativa* lines obtained from PGRC were generated at the Agriculture and Agri-Food Canada, Saskatoon Research and Development Centre. Plants were grown in 6-inch pots in a soilless medium (Stringam 1971) in a growth chamber with a photoperiod of 16 h and light/dark temperatures of 20 °C/16 °C. At maturity, water was withheld and plants allowed to dry, at which point seed was collected from the entire plant and seed from each plant kept separate.

Seeds (30 mg) from individual plants grown at the same time and under the same conditions, each representing one biological replicate, were ground under liquid nitrogen using a Helix grinder (Helix Technologies Inc., French Lick, IN, USA). The material was suspended in 1.2 ml of lysis buffer (7 M urea, 2 M thiourea, 19 mM Tris–HCl, 14 mM Tris-base, 0.2% Triton X-100) with 8% Complete mini EDTA-free protease inhibitor (Roche Diagnostics, Laval, Canada), 1.5 mg/ml DNase I (Roche Diagnostics) in dilution buffer (10 mM Tris–Cl pH 7.5, 150 mM NaCl, 1 mM $MgCl_2$), and 0.01 mg/ml bovine pancreas RNase A (Sigma-Aldrich, Oakville, Canada) added just prior to use (Withana-Gamage et al. 2013a). Soluble proteins were isolated by centrifugation at 10,000 g for 20 min. Disulfide bonds were reduced by incubation for 30 min at 4 °C with 1.0 mM DTT when required. Protein concentrations were determined using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Nepean, Canada).

An Experion Pro260 analysis kit (Bio-Rad Laboratories, Mississauga, Canada) was used to determine the relative proportion of each protein based on size from the seed extracts. Fresh, not frozen, protein samples were adjusted to 0.5 µg/µl and treated according to the manufacturer's protocol (Experion Pro260 Analysis kit, Bio-Rad Laboratories). In brief, gel solution, gel-stain solution, Pro260 ladder and sample buffer were prepared with Experion Pro260 analysis kit reagents. Note only the Pro260 ladder was heated to 100 °C; the samples were heated to 65 °C to prevent thiourea in the buffer from denaturing the proteins. Experion Pro260 chip micro-channels were used to separate proteins on an Experion automated electrophoresis station (Bio-Rad Laboratories). The resulting electropherograms were analysed using the percentage determination function in the Experion software which calculates each protein peak as a percent of the total protein within the sample.

Amino acid analysis

Seeds (3 g) from individual plants grown at the same time and under the same conditions, each representing one biological replicate, were defatted using hexane based on the methods of Troeng (1955) and Barthet and Daun (2004). Seeds were placed in sealed, steel tubes with 3 ball bearings and 25 ml of hexane (Sigma-Aldrich). Samples were ground for 45 min using an Eberbach shaker followed by filtration to remove oils and hexanes. Defatted meal was air-dried overnight followed by storage at − 20 °C. Total nitrogen content of the defatted meal was determined using a Flash EA 112 Series N/Protein 2000 Organic Elemental Analyzer (Thermo Fischer Scientific). This system uses a dynamic flash combustion system coupled with a gas chromatographic separation system based on the AOAC Method 972.43 (1999). Approximately, 15 mg of defatted meal from each sample (biological replicate) was analysed in triplicate (technical replicates). The nitrogen to protein conversion factor used was 6.25 (Mariotti 2008; AACC Method 46–18.01 1999). Moisture levels in the defatted meal were determined as weight loss upon drying to stability at 105 °C for 24 h in a forced air oven (AACC Method 44–01.01 1999). Approximately, 700 mg of defatted camelina meal was dried for each sample.

Amino acid profiles were analysed following the procedure of AOAC Method 994.12 (2005) and Tuan and Phillips (1997). Tryptophan was quantified following method of Nielsen and Hurrell (1985). For *C. sativa* lines from the PGRC repository, protein hydrolysis was conducted using a microwave, acid hydrolysis method modified from Lill et al. (2007) and Kabaha et al. (2011). Acid hydrolysis converts asparagine and glutamine into aspartic acid and glutamic acid, respectively; therefore, these amino acids are quantified together. Separation and quantification of amino acids was performed using a high-performance liquid chromatography (HPLC) system (Waters Alliance

2695) equipped with a Waters 2475 fluorescence detector with excitation wavelength of 250 nm, emission wavelength of 395 nm. Amino acids were resolved using a multistep gradient elution with an injection volume of 5 µl. Response peaks were recorded with the software Empower (Waters Corporation, Brossard, Canada). Pre-column derivatization using AccQ-Fluor (Waters Corporation) was done for all samples, except tryptophan which was diluted prior to application. For all amino acids except cysteine, methionine and tryptophan, 5 mg of protein basis was hydrolysed with 6 M HCL (Optima grade, Thermo Fisher Scientific) with 1% phenol using a CEM Discover SPD Microwave digester (ramp time 5.5 min, hold at 195 °C for 10 min, maximum pressure at 140 psi and maximum power at 300 W). Hydrolysates were neutralised with sodium hydroxide, filtered through a 0.45 µm Phenex RC syringe filter and applied to a Waters Oasis HLB C18 Cartridge. Flow through and washes were collected. Cysteine and methionine were determined as cystic acid and methionine sulfone after oxidation with performic acid followed by microwave hydrolysis with 6 M HCl, then neutralised and filtered as described. Tryptophan was determined by hydrolysing 10 mg of protein in 4.2 M NaOH in a 10 ml quartz hydrolysis tube with a teflon liner using a CEM Discover SPD Microwave digester (ramp time 6.0 min, hold at 215 °C for 20 min, with maximum pressure set at 140 psi and maximum power at 300 W). Hydrolysed samples were neutralised with HCl and filtered prior to application on a Waters Oasis HLB C18 Cartridge. The flow-through and washes were collected. Samples were stored at -20 °C prior to dilution and HPLC analysis. DL 2-aminobutyric acid and DL 5-methyl-tryptophan (Sigma-Aldrich) were used as internal standards. For experiments with *Camelina* species, amino acid analysis was conducted as described above, except the hydrolysis was performed as follows. Defatted meal was placed into 10 ml Pyrex screw cap vials with protein equivalents of 5 mg (nitrogen to protein conversion factor of 6.25). Hydrolysis was done in 2 ml of 6 M HCl (Optima grade, Thermo Fisher Scientific) with 1% (w/v) phenol for 24 h at 110 °C, with the exception of cysteine and methionine which were oxidised to cystic acid and methionine sulfone prior to hydrolysis in 6 M HCl. Tryptophan was not assessed.

Amino acids were reported as % w/w (weight of the specific amino acid/weight of all amino acids recovered X-100). For samples from each biological replicate, representing single plants grown at the same time and under the same conditions, amino acid and nitrogen analysis were performed in triplicate (technical replicates) and moisture determination as a single reading. Technical replications of the same sample presenting a large coefficient of variation (> 10) were repeated. Statistical differences between biological replicates were identified using JMP 13 software. A one-way analysis of variance (ANOVA) and the multiple comparison Tukey honestly significant difference (HSD) test were used to identify and rank significant differences ($P \leq 0.05$).

RNA-Seq analysis

C. sativa DH55 flower buds along the main raceme were marked at anthesis and developing bolls taken every 4 days from anthesis to seed maturity (40 days). RNA was isolated separately from samples from each time point. Buds from lines identified as belonging to one of three protein profile groups, either Group 1 (CN113733 and CN30476), Group 2 (CN30477and CN45816), or Group 3 (CN111331 and CN114265), were also marked at anthesis and bolls sampled similarly; however, prior to RNA isolation, equal amounts (by weight) of material from each time point were pooled into a single sample representing an average developmental profile for each line. This allowed the suite of seed protein genes expressed in each line to be compared, although it was not possible to determine when they were expressed. RNA isolation was performed similar to Suzuki et al. (2004) with volumes modified to allow extraction in 1.5 ml tubes. RNA was quantified on a Qbit using the BR RNA kit (Invitrogen/Thermo Fisher Scientific), and library generation (Truseq stranded mRNA kit) and Illumina sequencing (800,000–1,000,000 reads per sample) were performed by the National Research Council of Canada DNA Services Lab (Saskatoon, Canada). Reads were trimmed for adapters and quality using Trimmomatic 0.30, with a phred 33 quality score cutoff of 15 used for leading, trailing, and sliding window (4 bp) trimming, discarding any reads with under 55 bp remaining after trimming. CLC Genomics Workbench 11.0.1 was used to run RNAseq Analysis (version 2.1), which mapped the reads to the genome and calculated the transcripts per million (TPM). Quantile normalisation was applied to improve between-sample comparisons.

Proteomics analysis

Seed protein was solubilised in non-reducing protein loading buffer (2% SDS, 10% glycerol, 0.01% bromophenol blue in 60 mM Tris–HCl buffer, pH 6.8) and separated by electrophoresis on 12% SDS-PAGE gels. A high molecular weight region (49–54 kDa) was cut from the gel and subjected to LS-MS/MS analysis at the Genome BC Proteomics Centre, University of Victoria, Canada, as per the following procedure. Trypsin digests were performed as previously described (Loiselle et al. 2005). Briefly, the gel slice was cut into 1 mm cubes and transferred to a Genomics Solutions Progest (DigiLab Inc., Holliston, MA, USA) perforated digestion tray. The gel pieces were de-stained (methanol/water/acetic acid, 50/45/5, by vol.) prior to reduction with 10 mM dithiothreitol and alkylation with 100 mM

iodoacetamide. Modified sequencing-grade porcine trypsin solution (20 ng/µl) (Promega, Madison, WI, USA) was added at an enzyme/protein ratio of 1:50. Proteins were then digested for 5 h at 37 °C prior to collection of the tryptic digests and acid extraction of the gel slices (acetonitrile/water/formic acid, 50/40/10, by vol.). The samples were then lyophilised and stored at − 80 °C prior to analysis.

The peptide digest was separated by on-line reverse-phase chromatography using an EASY-nLC II system (Thermo Fisher Scientific) with a reverse-phase Magic C-18AQ pre-column (100 µm I.D., 2 cm length, 5 µm, 100 Å) and reverse-phase nano-analytical column Magic C-18AQ (75 µm I.D., 15 cm length, 5 µm, 100 Å) (Michrom BioResources Inc., Auburn, AL, USA) both prepared in-house, at a flow rate of 300 nl/min. The chromatography system was coupled on-line with an LTQ Orbitrap Velos mass spectrometer equipped with a Nanospray II source (Thermo Fisher Scientific). Solvents were A: 2% acetonitrile, 0.1% formic acid; B: 90% acetonitrile, 0.1% formic acid. After pre-column (~ 10 µl, 249 bar) and nanocolumn (~ 6 µl, 249 bar) equilibration, samples were separated by gradient elution (0 min: 5% B; 45 min: 45% B; 2 min: 80% B; hold 8 min: 80% B). The LTQ Orbitrap Fusion (Thermo Fisher Scientific) parameters were as follows: nano-electrospray ion source with spray voltage 2.1 kV, capillary temperature 225 °C. Survey MS1 scan m/z range 400–2,000 profile mode, resolution 60,000 FWHM at 400 m/z with AGC target 1E6, and one microscan with maximum inject time of 500 ms. Lock mass Siloxane 445.120024 for internal calibration with preview mode for FTMS master scans: on, injection waveforms: on, monoisotopic precursor selection: on; rejection of charge state: 1. The samples were analysed by the following methods: (1) top 15 FTMS/IT-CID method with the fifteen most intense ions charge state 2–4 exceeding 5000 counts were selected for CID ion trap MS/MS fragmentation (ITMS scans 2–16) with detection in centroid mode. Dynamic exclusion settings were: repeat count: 2; repeat duration: 15 s; exclusion list size: 500; exclusion duration: 60 s with a 10 ppm mass window. The CID activation isolation window was: 2 Da; AGC target: 1E4; maximum inject time: 100 ms; activation time: 10 ms; activation Q: 0.250; and normalised collision energy 35%.

A database was generated based on the published proteome of *C. sativa* (Kagale et al. 2014, 2016) and common contaminant sequences (human keratin and porcine trypsin) added. All cruciferin, napin, vicilin, and oleosin sequences were manually curated prior to inclusion in the database. The following sequences were corrected: napins (Csa11g017000, Csa12g024720, Csa12g024730), cruciferins (Csa14g004960, Csa03g005050, Csa11g015240), vicilins (Csa19g031870, Csa01g025880, Csa01g025890, Csa16g016660, Csa05g038120) and oleosin (Csa12g079570). All seed protein sequences were deposited in Genbank (accessions OL404969-OL405008). Tandem mass spectra were extracted, charge state deconvoluted and deisotoped by Proteome Discoverer version 1.4. All MS/MS samples were analysed using Mascot version 1.4.1.14 (Matrix Science, London, UK). Mascot was set up to search with a fragment ion mass tolerance of 0.60 Da and a parent ion tolerance of 8.0 PPM. Carbamidomethyl of cysteine was specified as a fixed modification. Deamidation of asparagine and glutamine, oxidation of methionine and propionamide of cysteine were specified as variable modifications. Scaffold (version Scaffold_4.8.4, Proteome Software Inc., Portland, OR, USA) was used to validate MS/MS based peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 95.0% probability by the Scaffold Local FDR algorithm. Protein identifications were accepted if they could be established at greater than 95.0% probability and contained at least 2 identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm (Nesvizhskii et al. 2003). Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony. Proteins sharing significant peptide evidence were grouped into clusters.

Phylogenetic analysis

Phylogenetic analysis was conducted using MEGA version 6.06 (Tamura et al. 2013). Sequences were aligned using MUSCLE with parameters set at gap opening penalty 10, gap extension penalty 0.2 and gap separation distance 4 for protein alignments and gap opening penalty 15, gap extension penalty 6.66, transition weight 0.5 for DNA alignments. Maximum likelihood trees were constructed using the best substitution model for each data set with 500 bootstrap iterations.

Protein modelling

The Swiss Model First Approach (Waterhouse et al. 2018) was used to identify the best template and to generate an initial structure for each cruciferin. The SWISS-MODEL template library (SMTL version 2020-05-20, PDB release 2020-05-15) (Bienert et al. 2017) was searched for evolutionary-related structures matching the target sequence using default settings (http://swissmodel.expasy.org). The best template, PDB 3KGL.1.A, was found with HHblits and identified as a homotrimer. The template structure was obtained from X-ray crystallography with a resolution of 2.98 angstroms. A structural alignment was calculated and the fit adjusted to the template using Swiss PDB Viewer, SPDBV (https://spdbv.vital-it.ch). The resultant structurally aligned SPDBV project files were submitted to Swiss Model workspace. Loops were constructed for untemplated regions and adjacent residues with low root

mean square differences (RMSD) using the Scan Loop Data Base for realistic loop options. When an acceptable loop was not identified, the residues associated with the loop were submitted for modelling to the DaReUS-Loop server (https://biserv.rpbs.univ-paris-diderot.fr/services/DaReUS-Loop). Energy minimization of the structure was done after loop selection. Energy minimization computations (bonds, angles, torsion, improper, non-bonded and electrostatic) were conducted with the GROMOS96 module in Swiss PDB Viewer. Model quality was reviewed using QMEAN and GMQE from Swiss Model, Ramachandran plot statistics were calculated using ProCheck (https://servicesn.mbi.ucla.edu/PROCHECK) and Z-Score from ProSA (https://prosa.services.came.sbg.ac.at/prosa.php). RMSD of the final structure was calculated for the structurally-aligned residues against the template 3KGL.1.A using Swiss PDB Viewer (van Gunstern 1996).

Electrostatic surface potentials of the molecules were calculated using the default settings in the APBS electrostatic plugin (Dolinsky et al. 2007). The molecule was prepared using PDB2PQR workflow to add missing side chains and hydrogen atoms, to assign partial charges and radii, and to remove ligands. The electrostatic map was calculated with the grid spacing set to 0.5 with molecular surface visualisation set at ± 5 on the solvent-excluded surface (Connolly surface). The protein dielectric constant was set at 2, the solvent dielectric constant at 78, and the temperature at 310 K. Hydrophobicity was ranked using the Eisenberg scale (Eisenberg et al. 1984). Models were coloured using the color_h pyMol script (https://pymolwiki.org/index.php/Color_h).

ClustalW was used for multiple sequence alignments. Evolutionary sequence conservation was determined using the ConSurf server (https://consurf.tau.ac.il/) (Landau et al. 2005). Phosphorylation sites were identified using Net Phos 2.0 (http://www.cbs.dtu.dk/services/NetPhos-2.0). PyMol (https://pymolwiki.org/index.php/FindSurfaceResidues) was used to colour each of the identified sites. Surface accessible phosphorylation sites on the trimer were identified using the find surface residues feature in PyMol. The cutoff to define exposed or not exposed residues was set at 2.0 squared Angstroms. CAST-P (computed atlas of surface topography of proteins) was used to calculate the main pocket of the trimer. Pocket volume, area, circumference, openings and sum of mouth areas were reported using Connolly solvent-excluded surface area, which is the contact surface created when a sphere of size 1.4 angstroms is rolled over the model.

Results

Seed protein profile diversity in Camelina species

Total seed protein from lines representing the spectrum of *Camelina* species (Suppl. Table S1) was separated by capillary electrophoresis under reducing (with β-ME) and non-reducing conditions (without β-ME) (Fig. 1; Suppl. Table S2). While many of the major peaks were in common between the species, a scheme to differentiate them based on unique peaks and patterns specific to each was developed (Suppl. Fig. S1). The *C. sativa*/*C. microcarpa* 4X/*C. microcarpa* 6X/*C. rumelica rumelica*/*C. rumelica transcapida* group could be differentiated from the *C. neglecta*, *C. laxa*/*C. hispida hispida*/*C. hispida grandiflora* group by the presence or absence of a 17 kDa peak under reducing conditions. *C. sativa* could then be differentiated by the presence of a 14 kDa peak and *C. rumelica rumelica*/*C. rumelica transcapida* differentiated from *C. microcarpa* 4X/*C. microcarpa* 6X by the presence or absence of a 33 kDa peak. *C. microcarpa* 4X exhibited a 54 kDa peak under non-reducing conditions, while *C. microcarpa* 6X did not. *C. neglecta* could be differentiated from *C. laxa*/*C. hispida hispida*/*C. hispida grandiflora* by a 12 kDa peak under reducing conditions and the latter further differentiated by 33 and 29 kDa peaks.

Protein and amino acid content in meal from Camelina species

The percent protein of defatted meal varied considerably between species, but generally less so between accessions of the same species (Table 1). Meal from the *C. microcarpa* 4X lines exhibited the lowest protein content, approximately 31%, while meal from *C. hispida hispida*, *C. laxa*, *C. rumelica transcapida* and lines within the *C. rumelica rumelica* and *C. sativa* groups approached or exceeded 40%. Amino acid content in the meal also varied significantly within and between species (Table 2). Of the essential amino acids most often added as supplements to feeds, lysine levels varied from a low of 4.77% (w/w) in meal from *C. rumelica rumelica* 609 to a high of 5.74% in *C. sativa* 1063 meal. Meal from the *C. sativa* lines generally had higher levels of lysine. Of the sulphur-containing amino acids, methionine was highest in meal from *C. rumelica rumelica* 609 and lowest in *C. microcarpa* 6X 198 meal, while cysteine was highest in *C. rumelica rumelica* 247 meal, but lowest in *C. rumelica rumelica* 1034 meal. Interestingly, histidine levels were significantly higher in the meal from *C. rumelica rumelica* 1034 (4.77%), which was almost twice that found in meal from

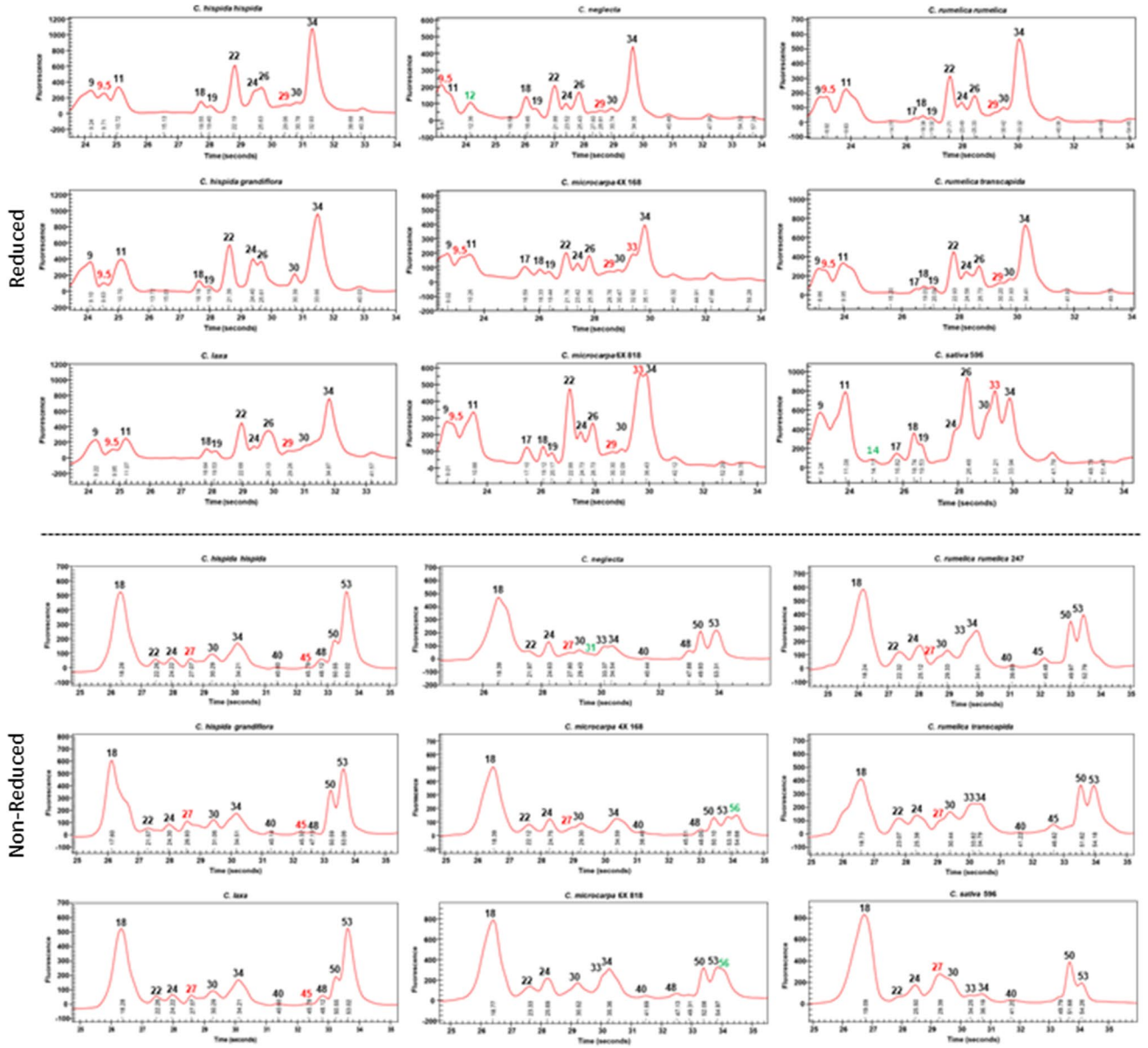

Fig. 1 Seed protein profiles from various *Camelina* species. Traces were generated by capillary electrophoretic separation of total seed protein under reducing (upper panel) and non-reducing (lower panel) conditions. Commons peaks (black numbers), peaks differing between species (red numbers) and peaks unique to a species (green numbers)

the other species. Serine content was highest (5.39%) in *C. sativa* 605 meal, but lowest (4.43%) in meal from another *C. sativa* line, 252. Threonine was also lowest (3.83%) in meal from *C. sativa* line 1662, but exceeded 4.5% in other *C. sativa* lines and other *Camelina* species.

Seed protein profile diversity in C. sativa

As variation in seed protein profile was observed with the nine *C. sativa* accessions examined above, the analysis was extended to include a global collection of 187 *C. sativa* lines from the PGRC. Lines could be classified based on the similarity of seed protein profiles under reducing or non-reducing conditions. It should be noted that while classification of the lines based on protein profiles generated under the two conditions was generally in agreement, some lines were placed into different groups dependent upon the condition under which the seed protein was separated. This allowed for an even finer level of discrimination when both data sets were considered. A complete list of the lines tested with accompanying capillary electrophoresis electropherograms can be found in Suppl. Table S3.

Under reducing conditions, seven different profiles were noted with the majority of the lines exhibiting one of three

Table 1 Protein content in meal from various *Camelina* species

Species	Line	Protein[1] (%)	SD	Significance category[2]
C. hispida grandiflora	248	36.17	3.71	>BCDEFGHI
C. hispida hispida	240	39.40	1.42	ABCD>>>>>
C. laxa	612	37.39	0.68	ABCDEFG>>
C. neglecta	246	34.06	2.66	>>>DEFGHI
C. microcarpa 4X	168	31.54	0.12	>>>>>>>HI
	718	31.42	0.98	>>>>>>>>I
	965	31.19	1.86	>>>>>>GHI
C. microcarpa 6X	198	32.94	0.89	>>>>>FGHI
	818	33.69	0.47	>>>>EFGHI
C. rumelica rumelica	247	39.24	0.84	ABCD>>>>>
	609	37.96	1.32	ABCDEF>>>
	1022	37.01	1.37	ABCDEFGH>
	1034	36.97	1.90	ABCDEFGH>
	1255	37.42	0.66	ABCDEFG>>
C. rumelica transcapida	245	40.03	2.99	ABC>>>>>>
C. sativa	239	41.70	0.49	AB>>>>>>>
	252	40.43	3.24	ABC>>>>>>
	596	35.78	1.07	>>CDEFGHI
	605	39.16	1.17	ABCDE>>>>
	621	41.68	0.87	AB>>>>>>>
	1044	40.72	1.21	ABC>>>>>>
	1062	39.91	0.86	ABC>>>>>>
	1063	41.83	2.03	A>>>>>>>>
	1662	42.49	0.20	A>>>>>>>>

[1]Mean ± SD ($n = 3$ biological replicates each with 3 technical replicates)

[2]Letters denote significant differences ($P = 0.05$). Tukey–Kramer comparison for least squares means

profiles as exemplified by lines CN113733, CN111311 and CN30477 (Fig. 2). Lines with these profiles exhibited several unique protein peaks or patterns between 22 and 36 kDa (Fig. 2a). Three distinct profiles were observed under non-reducing conditions with the pattern of proteins ranging from 49 to 54 kDa being one of the more distinguishing features (Fig. 2b). Profile 1 (e.g. CN113733) had a single peak ca. 51. kDa with a small higher molecular weight (MW) shoulder. Profile 2 (e.g. CN30477) was distinguishable by a unique peak at ca. 23 kDa, by a peak at ca. 36 kDa appearing as a shoulder on a common higher MW peak at ca. 39 kDa, and by two smaller, broad peaks of relatively equal abundance at ca. 52 and 55 kDa. Lines exhibiting Profile 3 (e.g. CN111331) were similar to Profile 1, but had two large peaks at ca. 51 and 54 kDa. Lines in the same category often showed slight differences in the ratio of proteins, but the profiles were very similar (Fig. 2c).

Protein and amino acid content in meal from diverse C. sativa accessions

Percent protein in defatted meal was found to vary considerably among the *C. sativa* lines; however, this did not correlate with protein profiles (Table 3). Meal from line CN113733 had the highest protein content (53.71%), while meal from line CN111331 had the lowest (43.26%). It should be noted that the average meal protein content among these lines (49.49%) was higher than in the nine accessions examined above (40.41%). This likely reflects the different locations and conditions under which the plants were propagated for these experiments.

The amino acid content in meal from the lines representing the three seed protein profiles was also examined to estimate the extent of diversity for this trait among the lines in the PGRC collection (Table 4). While a correlation between seed protein profile and amino acid content was not observed, the lines examined exhibited significant differences in meal amino acid content. Of the essential amino acids required by monogastric animals, methionine (converted to cysteine), threonine and lysine are often lacking in plant-based diets. In this regard, meal from lines CN113733, CN30476 and CN111331 had significantly higher levels (ca. 7–8% more) of lysine, while less variation was found for methionine and threonine levels. Meal from line CN30477 had generally higher levels of essential aliphatic amino acids, namely leucine, isoleucine and valine, than the other lines, while meal from line CN114265 had significantly higher levels of cysteine. Meal from line CN45816 had significantly higher levels of glutamic acid (ca. 4–7% more), but lower levels of hydroxyl amino acids (serine, threonine and tyrosine) as did meal from line CN114265. Meal from line CN30476 had the highest levels of serine and threonine.

Genes encoding major seed storage proteins in C. sativa

Examination of the *C. sativa* DH55 genome sequence (Kagale et al. 2014) identified genes encoding major seed proteins, namely cruciferin, napin, vicilin and oleosin, which were then annotated according to their relationship to the presumed *A. thaliana* orthologues and location of the gene on a specific *C. sativa* sub-genome (Suppl. Table S4). Twelve genes encoded the main Brassicaceae seed storage protein, cruciferin, of which five were located on sub-genome I (G1), four on sub-genome II (G2) and three on sub-genome III (G3). Phylogenetic comparison to the four genes encoding cruciferin in *A. thaliana* (*AtCRA*, *AtCRB*, *AtCRC* and At1g03890) revealed that two tandemly linked genes on G1 (Csa11g070580 and Csa11g070590) and one of the genes on G2 (Csa18g009670) were most similar to *AtCRA* and were named accordingly (Fig. 3; Suppl. Fig. S2a). A *CRA*

Table 2 Amino acid content in meal from various *Camelina* species

Species	ID #	Amino Acid Content (% w/w)[1,2,3]																
		Alanine	Arginine	Aspartate/ Asparagine	Cysteic Acid	Glutamate/ Glutamine	Glycine	Histidine	Isoleucine	Leucine	Lysine	Methionine	Phenyl-alanine	Proline	Serine	Threonine	Tyrosine	Valine
C. hispida grandiflora	248	4.77 ± 0.16 ABCD	9.51 ± 0.30 AB	8.58 ± 0.33 DE	7.98 ± 0.34 CDEFGHI	17.23 ± 0.52 ABC	6.22 ± 0.13 ABCD	2.41 ± 0.08 BC	3.48 ± 0.04 BCDE	5.70 ± 0.04 HIJK	5.47 ± 0.30 ABCDEF	2.57 ± 0.08 CDEFG	3.66 ± 0.06 HI	4.97 ± 0.16 ABCDE	5.18 ± 0.09 ABCD	4.09 ± 0.11 BCDEF	3.17 ± 0.07 AB	4.99 ± 0.08 FGHI
C. hispida hispida	240	4.50 ± 0.09 D	9.29 ± 0.21 AB	8.44 ± 0.13 E	8.22 ± 0.44 BCDEFGH	17.51 ± 0.38 ABC	6.09 ± 0.11 BCDE	2.28 ± 0.05 BC	3.69 ± 0.07 ABCD	5.92 ± 0.04 FGHIJK	5.46 ± 0.24 ABCDEF	2.81 ± 0.16 ABCD	3.69 ± 0.04 GHI	4.96 ± 0.04 ABCDE	4.87 ± 0.10 CDEFGHIJ	4.18 ± 0.14 ABCDEF	3.20 ± 0.04 AB	4.89 ± 0.13 HI
C. laxa	612	4.48 ± 0.09 D	9.50 ± 0.34 AB	8.84 ± 0.17 BCDE	7.80 ± 0.22 FGHI	17.83 ± 0.29 A	5.16 ± 0.06 IJKL	2.57 ± 0.06 B	3.59 ± 0.17 ABCDE	6.35 ± 0.07 ABCD	5.27 ± 0.10 ABCDEF	2.67 ± 0.13 ABCDEF	4.09 ± 0.05 AB	5.05 ± 0.07 ABCDE	4.60 ± 0.20 GHIJ	4.04 ± 0.10 DEF	2.93 ± 0.07 B	5.25 ± 0.24 CDEFG
C. neglecta	246	4.47 ± 0.17 D	9.90 ± 0.16 A	9.15 ± 0.53 BCDE	7.90 ± 0.34 DEFGHI	17.44 ± 0.38 ABC	6.09 ± 0.20 BCDE	2.52 ± 0.08 BC	3.64 ± 0.06 ABCD	5.73 ± 0.13 HIJK	4.88 ± 0.18 EF	2.50 ± 0.14 FGH	3.90 ± 0.05 CDEF	4.81 ± 0.06 CDEF	4.64 ± 0.07 GHIJ	4.02 ± 0.07 DEF	3.20 ± 0.02 AB	5.23 ± 0.08 CDEFGH
C. microcarpa 4X	168	4.72 ± 0.13 ABCD	9.26 ± 0.41 AB	9.15 ± 0.12 BCDE	6.93 ± 0.18 J	16.94 ± 0.12 ABC	6.02 ± 0.13 CDEFG	2.43 ± 0.04 BC	3.52 ± 0.04 BCDE	6.10 ± 0.08 BCDEFG	5.31 ± 0.25 ABCDEF	2.60 ± 0.08 BCDEF	3.91 ± 0.06 CDEF	5.24 ± 0.08 AB	5.03 ± 0.18 ABCDEFG	4.57 ± 0.14 A	3.11 ± 0.08 AB	5.15 ± 0.04 CDEFGHI
C. microcarpa 4X	718	4.79 ± 0.06 ABCD	10.13 ± 0.82 A	9.28 ± 0.91 BCDE	8.43 ± 0.30 BCDEFG	15.66 ± 0.52 DE	5.68 ± 0.29 EFGHI	2.54 ± 0.08 B	3.57 ± 0.14 ABCDE	5.60 ± 0.12 K	5.58 ± 0.21 ABCDE	2.51 ± 0.08 FGH	3.81 ± 0.05 DEFGH	5.02 ± 0.19 ABCDE	5.09 ± 0.14 ABCDEF	4.27 ± 0.14 ABCDE	3.03 ± 0.04 AB	5.00 ± 0.11 EFGHI
C. microcarpa 4X	965	4.68 ± 0.07 ABCD	9.52 ± 0.25 AB	11.24 ± 0.51 A	7.72 ± 0.34 FGHIJ	16.71 ± 0.69 ABCDE	5.38 ± 0.16 HIJK	2.30 ± 0.06 BC	3.49 ± 0.17 BCDE	5.96 ± 0.11 FGHIJ	5.32 ± 0.10 ABCDEF	2.62 ± 0.10 BCDEF	3.78 ± 0.04 EFGHI	4.80 ± 0.10 DEF	4.58 ± 0.13 GHIJ	3.97 ± 0.17 DEF	2.92 ± 0.03 B	5.01 ± 0.08 DEFGHI
C. microcarpa 6X	198	4.67 ± 0.11 ABCD	9.52 ± 0.48 AB	10.11 ± 0.29 AB	7.08 ± 0.43 IJ	17.49 ± 0.20 ABC	6.41 ± 0.17 ABCD	2.49 ± 0.03 BC	3.49 ± 0.03 BCDE	5.66 ± 0.06 HIJK	4.89 ± 0.13 CDEF	2.25 ± 0.18 H	3.86 ± 0.04 DEFG	4.83 ± 0.03 BCDEF	5.03 ± 0.08 ABCDEFGH	3.95 ± 0.05 DEF	3.06 ± 0.06 AB	5.22 ± 0.07 CDEFGHI
C. microcarpa 6X	818	4.57 ± 0.12 BCD	9.77 ± 0.42 AB	8.64 ± 0.21 CDE	8.48 ± 0.38 ABCDEF	17.26 ± 0.19 ABC	4.91 ± 0.12 KL	2.44 ± 0.06 BC	3.76 ± 0.06 AB	6.32 ± 0.20 ABCDE	5.11 ± 0.40 BCDEF	2.73 ± 0.08 ABCDEF	3.97 ± 0.12 ABCD	5.12 ± 0.09 ABCD	4.57 ± 0.20 HIJ	4.04 ± 0.11 DEF	2.97 ± 0.05 AB	5.35 ± 0.12 BCDE
C. rumelica rumelica	247	4.48 ± 0.07 D	10.04 ± 0.12 A	8.15 ± 0.08 E	9.32 ± 0.40 A	17.25 ± 0.35 ABC	5.50 ± 0.06 GHIJ	2.40 ± 0.02 BC	3.51 ± 0.04 BCDE	5.96 ± 0.08 FGHIJ	4.92 ± 0.09 DEF	2.72 ± 0.14 ABCDEF	3.82 ± 0.04 DEFGH	5.12 ± 0.04 ABCD	4.72 ± 0.08 EFGHIJ	4.08 ± 0.09 CDEF	2.98 ± 0.03 AB	5.03 ± 0.07 DEFGHI
C. rumelica rumelica	609	4.40 ± 0.13 D	9.83 ± 0.25 AB	8.43 ± 0.08 E	8.92 ± 0.29 AB	17.26 ± 0.43 ABC	4.96 ± 0.17 KL	2.51 ± 0.01 BC	3.76 ± 0.08 ABC	6.40 ± 0.11 AB	4.77 ± 0.21 F	2.89 ± 0.07 A	4.07 ± 0.08 ABC	5.06 ± 0.12 ABCDE	4.42 ± 0.21 J	4.02 ± 0.14 DEF	2.97 ± 0.09 AB	5.32 ± 0.08 BCDEF
C. rumelica rumelica	1022	4.79 ± 0.23 ABCD	9.16 ± 0.42 AB	9.75 ± 0.69 BC	8.18 ± 0.22 BCDEFGH	16.45 ± 0.60 BCDE	5.54 ± 0.20 FGHI	2.48 ± 0.17 BC	3.44 ± 0.16 CDE	5.98 ± 0.19 EFGHIJ	5.65 ± 0.35 ABC	2.68 ± 0.18 ABCDEF	3.81 ± 0.05 DEFGH	5.02 ± 0.25 ABCDE	4.78 ± 0.05 DEFGHIJ	4.50 ± 0.28 AB	2.92 ± 0.24 B	4.87 ± 0.04 I
C. rumelica rumelica	1034	5.05 ± 0.17 A	8.85 ± 0.40 B	8.96 ± 0.13 BCDE	6.91 ± 0.21 J	15.51 ± 0.26 E	6.50 ± 0.17 ABC	4.70 ± 0.38 A	3.46 ± 0.08 BCDE	5.63 ± 0.12 JK	5.84 ± 0.32 A	2.32 ± 0.06 GH	3.82 ± 0.04 DEFGH	4.51 ± 0.08 F	5.28 ± 0.24 ABC	4.37 ± 0.15 ABCD	2.95 ± 0.23 AB	5.34 ± 0.06 BCDEF
C. rumelica rumelica	1255	4.55 ± 0.05 CD	9.96 ± 0.18 A	8.82 ± 0.11 CDE	7.94 ± 0.39 DEFGHI	16.81 ± 0.14 ABCD	5.19 ± 0.16 IJKL	2.54 ± 0.02 BC	3.87 ± 0.03 A	6.46 ± 0.04 A	4.80 ± 0.12 F	2.84 ± 0.10 AB	4.11 ± 0.02 A	4.74 ± 0.07 EF	4.62 ± 0.03 GHIJ	4.26 ± 0.06 ABCDE	3.12 ± 0.04 AB	5.35 ± 0.07 BCD
C. rumelica transcapida	245	4.74 ± 0.05 ABCD	9.45 ± 0.24 AB	8.68 ± 0.06 CDE	7.57 ± 0.25 HIJ	17.06 ± 0.19 ABC	6.59 ± 0.13 AB	2.41 ± 0.03 BC	3.51 ± 0.08 BCDE	5.68 ± 0.03 IJK	5.61 ± 0.21 ABCD	2.53 ± 0.10 FGH	3.62 ± 0.02 I	4.81 ± 0.03 CDEF	5.34 ± 0.11 AB	4.33 ± 0.09 ABCD	3.14 ± 0.04 AB	4.92 ± 0.05 GHI
C. sativa	239	4.46 ± 0.06 D	9.64 ± 0.10 AB	8.50 ± 0.09 DE	7.88 ± 0.28 EFGHI	17.70 ± 0.36 AB	6.19 ± 0.20 ABCDE	2.21 ± 0.06 C	3.72 ± 0.05 ABCD	5.99 ± 0.12 EFGHI	5.36 ± 0.12 ABCDEF	2.81 ± 0.09 ABCDE	3.74 ± 0.06 FGHI	4.86 ± 0.06 CDEF	4.52 ± 0.05 IJ	4.09 ± 0.10 BCDEF	3.18 ± 0.11 AB	5.14 ± 0.08 CDEFGHI
C. sativa	252	4.79 ± 0.05 ABCD	9.52 ± 0.23 AB	8.71 ± 0.05 CDE	7.58 ± 0.13 GHIJ	17.02 ± 0.18 ABC	6.71 ± 0.05 A	2.39 ± 0.02 BC	3.50 ± 0.05 BCDE	5.68 ± 0.04 IJK	5.39 ± 0.20 ABCDEF	2.55 ± 0.10 EFG	3.68 ± 0.02 GHI	4.73 ± 0.04 EF	5.39 ± 0.05 A	4.27 ± 0.10 ABCDE	3.22 ± 0.04 A	4.90 ± 0.10 HI
C. sativa	596	4.90 ± 0.08 ABC	9.60 ± 0.33 AB	9.20 ± 0.23 BCDE	7.71 ± 0.23 FGHIJ	16.71 ± 0.45 ABCDE	6.04 ± 0.19 CDEF	2.29 ± 0.11 BC	3.28 ± 0.11 E	6.01 ± 0.12 DEFGHI	5.31 ± 0.25 ABCDEF	2.59 ± 0.10 BCDEFG	3.93 ± 0.05 BCDE	5.17 ± 0.16 ABC	4.90 ± 0.06 BCDEFGHI	4.50 ± 0.16 AB	2.96 ± 0.18 AB	4.91 ± 0.20 GHI
C. sativa	605	4.58 ± 0.05 BCD	9.81 ± 0.30 AB	8.32 ± 0.28 E	8.12 ± 0.45 BCDEFGH	17.10 ± 0.33 ABC	5.81 ± 0.12 DEFGH	2.55 ± 0.03 B	3.47 ± 0.06 BCDE	6.04 ± 0.08 CDEFGH	5.22 ± 0.30 ABCDEF	2.67 ± 0.16 ABCDEF	3.85 ± 0.05 DEFG	5.30 ± 0.07 A	4.74 ± 0.11 DEFGHIJ	4.28 ± 0.09 ABCDE	3.05 ± 0.05 AB	5.09 ± 0.09 DEFGHI
C. sativa	621	4.50 ± 0.15 D	9.94 ± 0.29 A	8.59 ± 0.20 CDE	8.51 ± 0.36 ABCDEF	17.03 ± 0.22 ABC	5.23 ± 0.27 IJKL	2.45 ± 0.05 BC	3.62 ± 0.09 ABCD	6.42 ± 0.14 AB	4.88 ± 0.31 EF	2.83 ± 0.11 ABC	4.11 ± 0.09 AB	4.92 ± 0.02 BCDE	4.63 ± 0.22 GHIJ	4.14 ± 0.09 BCDEF	3.02 ± 0.04 AB	5.18 ± 0.09 CDEFGHI
C. sativa	1044	4.54 ± 0.08 CD	9.49 ± 0.13 AB	8.87 ± 0.10 BCDE	8.67 ± 0.55 ABCDE	17.31 ± 0.46 ABC	4.98 ± 0.09 JKL	2.46 ± 0.04 BC	3.53 ± 0.17 BCDE	6.10 ± 0.08 BCDEFG	5.32 ± 0.23 ABCDEF	2.67 ± 0.16 ABCDEF	3.89 ± 0.05 DEF	5.04 ± 0.15 ABCDE	4.70 ± 0.13 EFGHIJ	4.03 ± 0.06 DEF	2.93 ± 0.09 B	5.48 ± 0.24 ABC
C. sativa	1062	4.46 ± 0.06 D	9.76 ± 0.26 AB	8.67 ± 0.09 CDE	8.74 ± 0.75 ABCD	17.26 ± 0.18 ABC	4.75 ± 0.06 L	2.43 ± 0.06 BC	3.75 ± 0.04 ABC	6.38 ± 0.07 ABC	4.96 ± 0.13 CDEF	2.67 ± 0.17 ABCDEF	3.95 ± 0.07 ABCDE	5.13 ± 0.08 ABCD	4.43 ± 0.06 J	3.91 ± 0.09 EF	2.97 ± 0.06 AB	5.80 ± 0.11 A
C. sativa	1063	4.95 ± 0.19 AB	9.60 ± 0.44 AB	9.61 ± 0.29 BCD	7.18 ± 0.34 IJ	16.32 ± 0.25 CDE	5.98 ± 0.29 CDEFG	2.44 ± 0.05 BC	3.42 ± 0.08 DE	5.82 ± 0.08 GHIJK	5.74 ± 0.21 AB	2.55 ± 0.07 DEFG	3.81 ± 0.02 DEFGH	4.89 ± 0.15 BCDE	5.15 ± 0.15 ABCDE	4.46 ± 0.18 ABC	3.03 ± 0.06 AB	5.05 ± 0.06 DEFGHI
C. sativa	1662	4.53 ± 0.10 CD	9.55 ± 0.18 AB	8.85 ± 0.09 BCDE	8.83 ± 0.73 ABC	17.17 ± 0.31 ABC	4.97 ± 0.09 JKL	2.37 ± 0.07 BC	3.55 ± 0.03 BCDE	6.23 ± 0.11 ABCDEF	5.32 ± 0.17 ABCDEF	2.69 ± 0.18 ABCDEF	3.91 ± 0.06 CDEF	4.96 ± 0.08 ABCDE	4.68 ± 0.10 FGHIJ	3.84 ± 0.14 F	2.92 ± 0.06 B	5.62 ± 0.10 AB

[1] %AA (w/w) = mg of specific amino acid divided by the total recovered mg (sum of 19 recovered amino acids - tryptophan not determined) multiplied by 100.

[2] Mean ± SD (n=3). Means were calculated using a nested mixed model using the Maximum Likelihood (REML) method. Letters that differ within a column were significantly different using Tukeys HSD test (P<0.05).
Letters which differ within a column are significantly different P<0,05. Tukeys HSD test.

[3] Highest value, Lowest Value

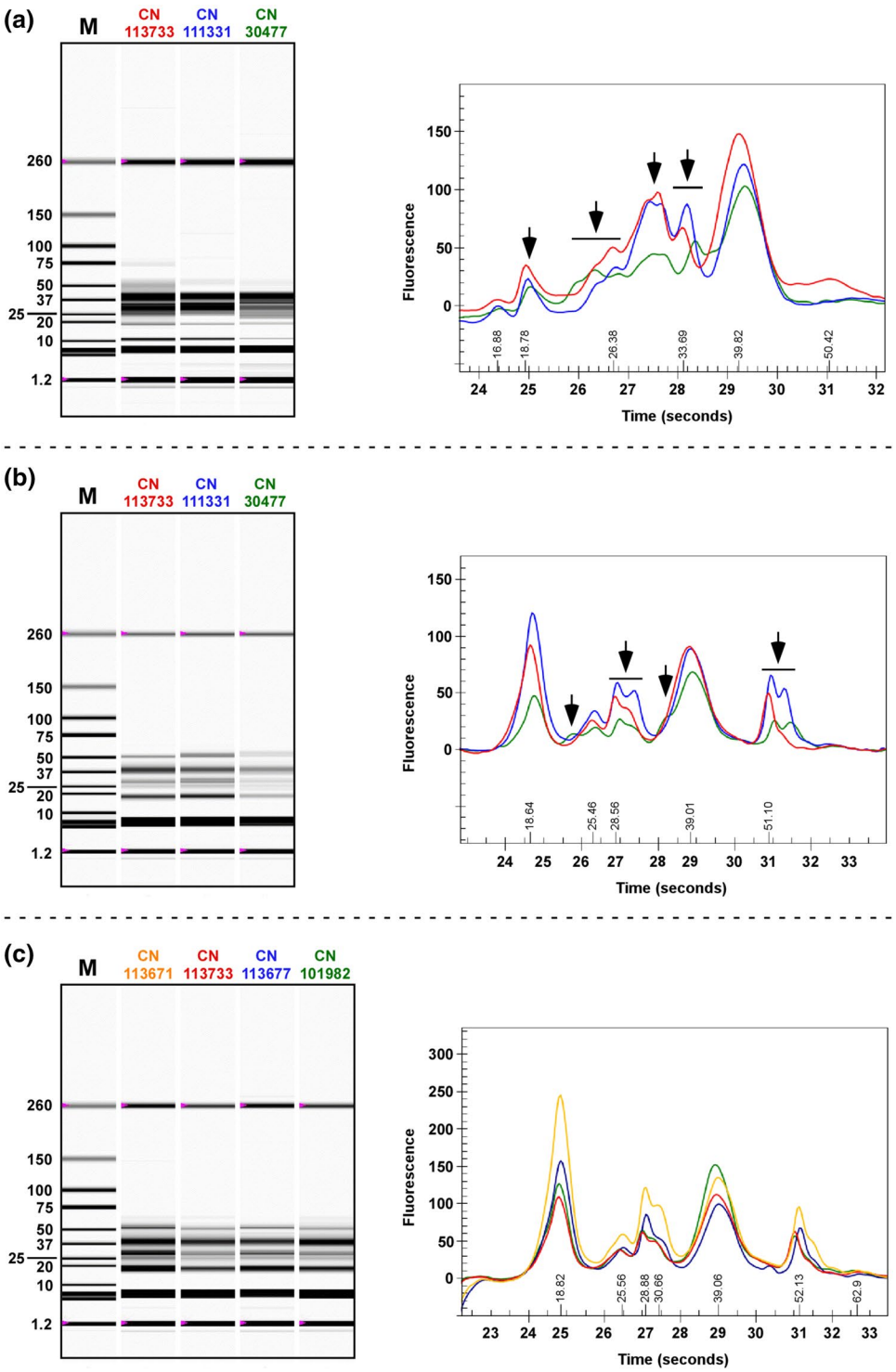

Fig. 2 Seed protein profiles from *C. sativa* accessions. Virtual digital gels (left-hand side) and traces (right-hand side) were generated by capillary electrophoretic separation of total seed protein under reducing (**a**) and non-reducing (**b** and **c**) conditions. **a** and **b** *C. sativa* lines representing the three main profiles (profile 1—CN113733, profile 2—CN30477 and profile 3—CN111331). Arrows denote differences between profiles. **c** Variation among four lines exhibiting seed protein profile 1

orthologue was not found on any of the *C. sativa* G3 chromosomes, while single orthologues of *AtCRB* and *AtCRC* were found on each of the three sub-genomes. The phylogenetic analysis also revealed genes encoding a fourth type of cruciferin in *C. sativa*, hereafter referred as *CsCruD*, which was most similar to the cruciferin encoded by the *A. thaliana* At1g03890 locus. Single *CsCruD* orthologues were found on each of the *C. sativa* sub-genomes, each linked in tandem to a *CsCruB* gene, which is similar to the arrangement in the *A. thaliana* genome.

Vicilin is a cupin-domain protein similar in structure to cruciferin. In total, eight genes encoding vicilin-like proteins were identified in the *C. sativa* DH55 genome (Suppl. Table S4). Phylogenetic analysis revealed that five of the *C.*

Table 3 Protein content in meal from *C. sativa* lines with various seed protein profiles

Species	Protein Profile	Line	Protein[1] (%)	SE	Significance category[2]
C. sativa	1	CN113733	53.71	0.24	A
		CN30476	47.27	0.66	C
	2	CN30477	49.55	0.74	BC
		CN45816	51.77	0.25	AB
	3	CN111331	43.26	0.39	D
		CN114265	51.44	0.97	AB

[1]Mean ± SE ($n = 4$, except for CN111331 where $n = 3$)

[2]Letters denote significant differences ($P = 0.05$). Tukey–Kramer comparison for least squares means

sativa vicilins formed two related subgroups that were most similar to the *A. thaliana* vicilin AtPAP85 (also known as vicilin 1); accordingly, these vicilins were denoted CsVic1A and CsVic1B (Fig. 3; Suppl. Fig. S2b). The *CsVic1A* subgroup contained homeologues from all three sub-genomes (Csa19g031870, Csa1g025880 and Csa15g039290), while the *CsVic1B* subgroup included a gene on G3 (Csa15g039300) and a gene on G2 (Csa01g025890), but was missing a G1 homeologue. The two tandem *Vic1* genes on G2 represent both subgroups, as did the two tandemly linked genes on G3. The remaining vicilin genes (Csa07g016060, Csa16g016660 and Csa05g038120) were most similar to *A. thaliana* vicilin *AtVCL22* (denoted herein as vicilin 2) with homeologues present on each of the three *C. sativa* sub-genomes.

The original annotation of the *C. sativa* DH55 genome identified five genes encoding the 2S albumin, napin (Kagale et al. 2014); however, a transcriptomic study indicated that as many as eight genes might exist (Nguyen et al. 2013). As this did not correspond with the expectation of gene number based on the genomic prediction, the assembly of the genomic regions containing the napin genes was re-examined. This revealed that three of the genes that had been previously annotated as single genes by Kagale et al. (2014) were in fact closely related genes linked in tandem and had been misassembled. In agreement with the previous transcriptomic study, eight genes encoding napin were identified after separation of the tandem genes, four of which were in

Table 4 Amino acid content in meal from *C. sativa* lines with various seed protein profiles

Amino Acid	Amino acid content (% w/w) per accession[1,2]						Average
	Seed protein profile 1		Seed protein profile 2		Seed protein profile 3		
	CN113733	CN30476	CN30477	CN45816	CN111331	CN114265	
Alanine	4.74 ± 0.12 B	4.89 ± 0.08 A	4.72 ± 0.06 BC	4.61 ± 0.11 C	5.01 ± 0.11 A	4.63 ± 0.11 BC	4.76 ± 0.16
Arginine	9.82 ± 0.29 AB	9.46 ± 0.32 C	9.59 ± 0.16 BC	9.98 ± 0.31 A	9.56 ± 0.24 BC	9.78 ± 0.31 ABC	9.69 ± 0.32
Aspartate/ Asparagine	9.45 ± 0.14 AB	9.4 ± 0.24 AB	9.59 ± 0.11 A	9.26 ± 0.45 BC	9.49 ± 0.21 AB	9.09 ± 0.22 C	9.38 ± 0.29
Cysteic Acid	3.46 ± 0.21 B	3.38 ± 0.28 B	3.13 ± 0.27 B	3.37 ± 0.62 B	3.27 ± 0.32 B	3.95 ± 0.58 A	3.44 ± 0.48
Glutamate/ Glutamine	17.68 ± 0.33 BC	17.93 ± 0.21 B	17.89 ± 0.12 B	18.63 ± 0.52 A	17.45 ± 0.47 C	17.98 ± 0.3 B	17.93 ± 0.46
Glycine	5.17 ± 0.03 C	5.41 ± 0.05 B	5.5 ± 0.05 B	5.49 ± 0.06 AB	5.64 ± 0.18 A	5.53 ± 0.18 AB	5.45 ± 0.17
Histidine	2.73 ± 0.06 A	2.69 ± 0.07 AB	2.61 ± 0.06 BC	2.67 ± 0.04 AB	2.55 ± 0.1 C	2.66 ± 0.11 AB	2.66 ± 0.09
Isoleucine	3.77 ± 0.09 B	3.71 ± 0.09 B	4.05 ± 0.09 A	3.81 ± 0.16 B	3.77 ± 0.08 B	3.72 ± 0.11 B	3.81 ± 0.16
Leucine	6.93 ± 0.14 AB	6.83 ± 0.11 B	7.04 ± 0.12 A	6.85 ± 0.12 B	6.85 ± 0.14 B	6.85 ± 0.13 B	6.9 ± 0.14
Lysine	5.81 ± 0.08 A	5.86 ± 0.09 A	5.42 ± 0.1 B	5.55 ± 0.07 B	5.8 ± 0.14 A	5.52 ± 0.16 B	5.66 ± 0.21
Methionine	1.84 ± 0.16 AB	1.77 ± 0.16 B	1.86 ± 0.18 AB	1.75 ± 0.2 B	1.85 ± 0.2 AB	2.02 ± 0.27 A	1.85 ± 0.21
Phenylalanine	4.36 ± 0.07 AB	4.37 ± 0.15 AB	4.42 ± 0.05 A	4.26 ± 0.15 B	4.36 ± 0.16 AB	4.33 ± 0.13 AB	4.36 ± 0.13
Proline	5.53 ± 0.09 A	5.39 ± 0.04 B	5.26 ± 0.06 C	5.46 ± 0.16 AB	5.55 ± 0.13 A	5.49 ± 0.11 AB	5.44 ± 0.14
Serine	4.57 ± 0.09 C	4.78 ± 0.09 A	4.59 ± 0.09 BC	4.52 ± 0.13 C	4.71 ± 0.07 AB	4.54 ± 0.09 C	4.62 ± 0.13
Threonine	3.89 ± 0.05 ABC	3.98 ± 0.11 A	3.94 ± 0.06 AB	3.81 ± 0.13 BC	3.95 ± 0.07 AB	3.81 ± 0.1 C	3.9 ± 0.11
Tryptophan	1.38 ± 0.07 A	1.23 ± 0.08 B	1.32 ± 0.08 AB	1.25 ± 0.13 B	1.25 ± 0.09 AB	1.31 ± 0.14 AB	1.29 ± 0.11
Tyrosine	3.2 ± 0.04 C	3.28 ± 0.04 AB	3.35 ± 0.02 A	3.18 ± 0.12 C	3.23 ± 0.07 BC	3.21 ± 0.06 C	3.25 ± 0.08
Valine	5.67 ± 0.11 AB	5.64 ± 0.16 AB	5.74 ± 0.1 A	5.55 ± 0.19 B	5.69 ± 0.1 AB	5.55 ± 0.18 B	5.64 ± 0.15

[1]%AA (w/w) = mg of specific amino acid divided by the total recovered mg (sum of 19 recovered amino acids–tryptophan not determined) multiplied by 100

[2] Mean ± SD ($n = 4$ except for CN111331 where $n = 3$). Letters within a row denote significant differences ($P = 0.05$). Tukey–Kramer comparison for least squares means

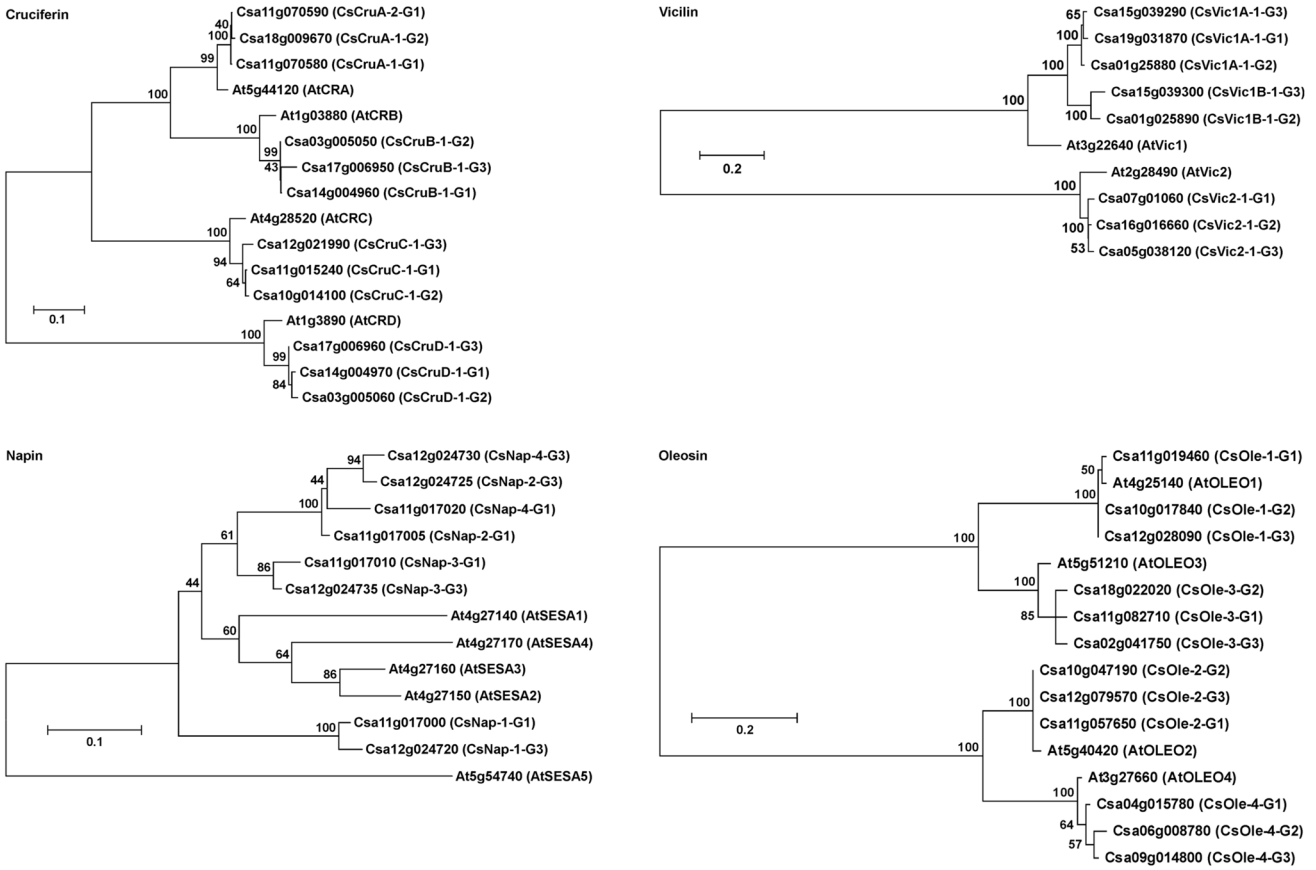

Fig. 3 Phylogenetic analysis of major *C. sativa* seed proteins. Maximum likelihood trees were constructed using the best substitution model for each data set with 500 bootstrap iterations. Numbers beside nodes indicate percentage of trees agreeing with the consensus

a cluster on G1 and four in a cluster on G3. Phylogenetic analysis revealed that the *C. sativa* napins were most similar to AtSESA1, AtSESA2, AtSESA3, and AtSESA4, which are also closely related and linked in tandem on *A. thaliana* chromosome 4, and distinct from the other *A. thaliana* napin, AtSESA5, which is encoded by a gene on chromosome 5 (Fig. 3; Suppl. Fig. S2c). Each of the eight *C. sativa* proteins could be paired to one of the eight napins reported in the earlier transcriptomic study (Nguyen et al. 2013) (Suppl. Fig. S3); however, the *C. sativa* proteins were renamed according to their genomic locations as per cruciferin and vicilin (Suppl. Table S4). No genes encoding napin were found on G2. The first two genes in the tandem series on G1 (Csa11g017000) and G3 (Csa12g024720) appear to be homeologues based on phylogenetic analysis; however, the other paralogues in each tandem series appear to have arisen through separate gene duplication events (Suppl. Fig. S2c) and the fact that there are four in each cluster appears to be coincidental.

Oleosins possess hydrophilic and hydrophobic domains that allow them to organise storage triglycerides into the oil bodies commonly found in cells of oilseed embryos. In total, 12 *C. sativa* genes were found to encode oleosins comprising three homeologues related to each of the genes encoding the four major *A. thaliana* oil body-associated oleosins (OLEO1 to 4) (Fig. 3; Suppl. Fig. S2d). *C. sativa* orthologues of members of the extended oleosin-like family were also identified.

Temporal expression of C. sativa seed storage protein genes through seed development

RNA-Seq analysis was conducted with *C. sativa* DH55 developing bolls from anthesis to seed maturity (40 days) to ascertain the expression profile of genes encoding seed proteins (Table 5). Transcripts derived from all of the genes encoding the two major seed storage proteins, namely cruciferin (12) and napin (8), were identified. Both sets exhibited similar expression patterns with a sharp increase in expression detected between 8–12 days after anthesis (daa) and a sharp decline between 28–32 daa. Members of the tandem napin clusters on G1 and G3 differed greatly in their levels of expression, but not in their temporal patterns. The three homeologous genes

encoding cruciferin CsCruD isoforms were expressed at lower levels than those encoding CsCruA, CsCruB or CsCruC suggesting that CsCruD may contribute less to overall seed protein composition. There was some evidence for genome partitioning with respect to the level of expression of homeologous genes encoding CsCruA, CsCruB or CsCruC.

The expression of the homeologous genes encoding CsVic1A on G1 (Csa19g031870), G2 (Csa01g025880) and G3 (Csa15g039290) increased sharply at 12 daa and high levels of transcripts were detected throughout seed development. The expression of the gene encoding CsVic1B on G3 (Csa15g039300) increased more gradually until 28 daa before declining sharply, while few transcripts were detected from its homeologous partner on G2 (Csa01g025890). Temporal patterns were also apparent in the expression of genes encoding oleosins. In general, the expression of genes encoding oleosins increased between 8 and 12 daa, though those encoding CsOle-1 were induced slightly earlier. Transcript levels from homeologous genes encoding CsOle-3 declined after 20 daa, while the expression of genes encoding CsOle-1, CsOle-2 and CsOle-4 remained elevated or continued to increase until the seeds were mature (40 daa). Of the other proteins known to contribute to seed protein composition, many genes encoding dehydrins or members of various late embryo abundant (LEA) protein families were also expressed at high levels during the later stages of seed development as expected (Suppl. Table S5).

Comparison of the high molecular weight proteome in diverse C. sativa accessions

The feature that most distinguished the *C. sativa* accessions was a high molecular weight region (49–55 kDa) appearing under non-reducing conditions, therefore, proteomics analysis of this region was conducted with two lines representing each of the three major seed protein profiles observed under non-reducing conditions, namely Profile 1 (CN113733 and CN30476), Profile 2 (CN30477 and CN45816) and Profile 3 (CN111331 and CN114265) (Suppl. Fig. S4). As expected, the most abundant proteins within this fraction were cruciferins (Suppl. Table S6) of which all four types were represented. Across all lines, CsCruA (MW 52 kDa) was the most abundant cruciferin and approximately three times more so than CsCruB (MW 51 kDa). The level of CsCruD (MW 50 kDa) was low, but relatively similar among the lines, while the amount of CsCruC (MW 55 kDa) varied extensively. Higher levels of CsCruC were present in line CN45816, while lines CN113733 and CN111331 had 10–12 times less. The relative abundance of the cruciferin isoforms did not fully explain the differences in protein profiles in this region; however, other proteins of similar MW were found in this fraction, including a group of nitrile specifier proteins that were even more abundant than CsCruD.

Comparison of the seed transcriptome in diverse C. sativa accessions

To examine the genetic basis underlying the different seed protein profiles among the *C. sativa* accessions, RNA-Seq analysis (Suppl. Table S7) was also conducted with these lines. Lines from the same seed protein profile groups did not exhibit seed protein gene expression patterns that were indicative of a specific group, although differences in temporal patterns could not be evaluated since bolls from all stages of development were pooled in this experiment. Genetic variation existed in the overall patterns between the lines and in comparison to the collective profile for *C. sativa* DH55 which has an electrophoretic protein profile similar to Profile 3 (Table 5).

The napin genes encoding CsNap-1, CsNap-3, CsNap-4 on the G1 and G3 sub-genomes were expressed at the highest levels, while genes encoding CsNap-2 were expressed at appreciably lower levels (ca. 10–50%) than the other *CsNap* genes in all of the lines. This pattern was similar to that observed with *C. sativa* DH55, although in this line *CsNap-1-G1* was expressed at a lower level and *CsNap-2-G1* at high levels. Notably, in DH55 *CsNap-2-G3* was induced much later and for a shorter period of time than the other napin genes (Table 5), which may have also contributed to the lower overall transcript levels in the other *C. sativa* lines.

The expression pattern of genes encoding CsCruA and CsCruB was similar in all *C. sativa* lines, including DH55. *CsCruA-2-G1* and *CsCruA-1-G2* were expressed at comparable levels and approximately twice that of *CsCruA-1-G1*, while the expression of the *CruB* genes was in the following order, *CsCruB-1-G3* > *CsCruB-1-G1* > *CsCruB-1-G2*. The pattern of *CruC* expression was markedly different between the lines. CN45816 and DH55 (Table 5; Suppl. Table S7) exhibited very high levels of *CsCruC-1-G3* expression (in fact, the highest of all of the cruciferin genes), high levels of *CsCruC-1-G1* expression and lower levels of *CsCruC-1*-G2 expression. Conversely, CN30476 and CN114265 expressed mainly *CsCruC-1-G3* and only at lower levels, while the CN113733, CN30477 and CN111331 possessed few or no *CruC* transcripts. As in DH55, the expression of genes encoding CsCruD was also low in the other *C. sativa* lines when compared to genes encoding CsCruA and CsCruB. Proteomic analysis of the high MW protein region, of which cruciferin was the most abundant member, confirmed these patterns (Suppl. Table S6).

The expression of vicilin genes was similar to DH55 with higher levels of expression detected from genes encoding CsVic1A and with comparatively little contribution from those encoding CsVic1B. The genes encoding CsVic2 on

sub-genomes G1 and G3 were expressed at approximately 30% the level of the genes encoding CsVic1A, with the *CsVic2* gene on G2 contributing few transcripts. Genes encoding oleosins CsOle-1, CsOle-2 and CsOle-4 were expressed at higher levels than those encoding CsOle-3. This was similar to the pattern in DH55, though it should be noted that expression of genes encoding CsOle-3 declined as seed development progressed, while the expression of genes encoding the other oleosins continued to increase throughout (Table 5).

Structural diversity of *C. sativa* cruciferins

In its natural form, cruciferin exists as a hexamer with a stochastic composition dependent on the availability of individual protomers (subunits). The functional properties of cruciferin are, therefore, an average of the functional properties of the subunits contributing to the whole. As variation was observed in the expression of genes encoding CruC and in actual cruciferin composition in the meal, the structure and potential functional properties of *C. sativa* cruciferins were examined.

Homology models of *C. sativa* cruciferins representing each of the four main classes (CsCruA, CsCruB, CsCruC and CsCruD) were constructed using the *B. napus* procruciferin (Cru2/3a, PDB 3KGL) as a template (Fig. 4: Suppl. Fig. S5). The *C. sativa* cruciferins had a reasonable degree of sequence identity with the *B. napus* template: 86.9% (CsCruA), 74.3% (CsCruB), 61.6% (CsCruC) and 51% (CsCruD). The difference between CsCruC and the template was largely attributed to an extended hypervariable region (HVR) II (Fig. 5; Suppl. Fig. S6), while CsCruD is phylogenetically distinct from the other cruciferins. Nonetheless, each of the *C. sativa* cruciferins possessed a highly conserved core structure consisting of two jelly roll β-barrels and two extended helix regions comprised of 27 β-sheets, six α-helices and three 3_{10}-helices, which is typical of cupin domains associated with 11S and 7S globulins (Tandang-Silvas et al. 2010). The HVR regions cannot be resolved by crystallography as they do not possess ordered secondary structures, such as β-sheets or α-helices, and likely form loops protruding from the core (Adachi et al. 2001; Tandang-Silvas et al. 2010). To account for this, the energy minimization approach used by Withana-Gamage et al. (2011) to model *A. thaliana* cruciferin loops was employed; however, models were first constructed for those loops that had a similar modelled loop in the Scan Loop Data Base. The DaReUS-Loop server was used to construct loops for those without an acceptable template in the database. Only then were stereochemical alterations made to minimise energy based on the GROMOS 96 force field calculations. Several parameters indicated that the *C. sativa* cruciferin models were of high quality and geometrically correct (Suppl. Table S8). *G*-factor scores based on torsion angles and covalent bond geometry ranged from − 0.09 to − 0.16 which was well within the generally regarding acceptable value range of 0 to − 0.5. Ramachandran plots showed that the sum of the percentage of residues in the core, allowed and additionally-allowed regions was 100% for CsCruA, CsCruB and CsCruD and 99.96% for CsCruB. Qualitative Model Energy ANalysis (QMEAN) scores, a composite measure of several geometric parameters (Benkert et al. 2008) with 0 considered as a good model and values < − 0.4 generally considered poor, ranged from − 0.98 to − 0.27. *Z*-scores, a measure of overall model quality based on the deviation of the total energy of the structure with respect to an energy distribution derived from random conformations (Benkert et al. 2011), ranged from 6.73 to 7.11. These scores were similar to models of *A. thaliana* cruciferins (Withana-Gamage et al. 2011) and within the range observed for models of proteins of similar size. RMSD derived by superimposing the *C. sativa* cruciferin models on the template indicated close alignment of the backbone with RMSD values all below 0.5 Å.

Alignment of the *C. sativa* cruciferins indicated a high degree of variability between CsCruA, CsCruB, CsCruC and CsCruD in each of the five HVRs (Fig. 5; Suppl. Fig. S6); these are also referred to as disordered regions due to their inability to be modelled or resolved by crystallographic methods (Adachi et al. 2001, 2003). HVR-I and HVR-V reside at the amino- and carboxy-terminus of the mature cruciferin, respectively, with the differences between the *C. sativa* cruciferin types attributed to amino acids with various properties. The three major solvent-exposed loops are represented by HVR-I, HVR-III (a.k.a. the extended loop region) and HVR-IV and were replete with charged (glutamate, arginine and lysine) and polar (asparagine, glutamine and serine) amino acids. The CsCruC paralogues had the longest HVR-II regions, although this was much shorter than that found within *A. thaliana* CruC (Suppl. Fig. S6). Hexamer formation proceeds by interaction of the interchain disulphide bond-containing (IE) faces of two trimers after proteolytic processing at the β-cleavage site (Fig. 5) which permits movement of HVR-IV to the periphery of the protein and exposes the trimer-interacting regions. The four trimer-interacting regions were highly conserved in CsCruA, CsCruB and CsCruC (Fig. 5); however, several differences were noted in CsCruD, in particular in polar and charged residues important for hydrogen bond and ionic interactions between the trimers (Adachi et al. 2001, 2003; Tandang-Silvas et al. 2010). This suggests that while CsCruD may form trimers, its participation may lead to hexamers with less stable structures. HVR-II and HVR-V remain on the IE face and their high degree of variability contributes to the lower degree of evolutionary conservation, as well as variation in electrostatic potential and hydrophobicity (Fig. 4) which

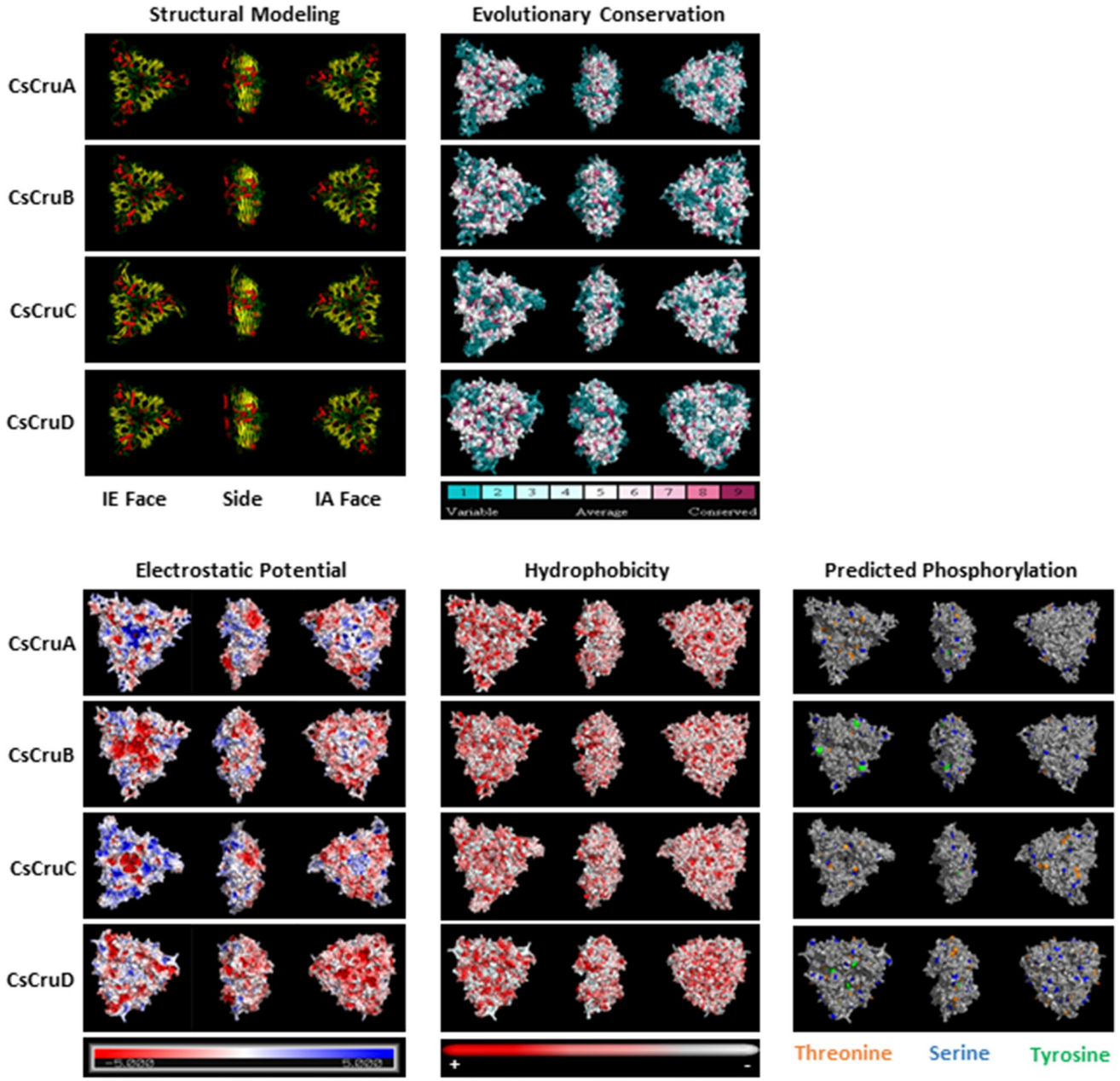

Fig. 4 Structural modelling, evolutionary conservation, surface hydrostatic potential, surface hydrophobicity and predicted phosphorylation of *C. sativa* cruciferins. Structural modelling panel: yellow = β-sheet, red = α-helix and green = loops. IE–interchain interacting face. IA–intrachain interacting face. One representative from each cruciferin type is shown: CsCRA (CRA-1-G1), CsCRB (CRB-1-G1), CsCRC (CRC-1-G1) and CsCRD (CsCRD-1-G1)

may also influence the stability of trimer-trimer interactions. Additional cysteine residues not predicted to be involved in inter- or intrachain disulphide bond formation were present in CsCruB and CsCruC (Fig. 5; Suppl. Fig. S6), which could promote interactions with other proteins/molecules or intersubunit disulphide bond exchanges (Shimada et al. 1980; Inquello et al. 1993).

In the context of functional properties (i.e. the properties that proteins confer in multi-component systems), the physicochemical properties of native cruciferin are directly related to the nature of the surface-exposed residues (Withana-Gamage et al. 2013a, 2013b, 2015, 2020). CsCruD had the highest percentage of negatively charged amino acids (11.1%; total net charge − 14) and the lowest isoelectric point (4.99) of the *C. sativa*, *A. thaliana* and *B. napus* cruciferins (Table 6). CsCruD had a grand average hydropathicity (GRAVY) value of − 0.375, making it the least hydrophilic of all the cruciferins examined,

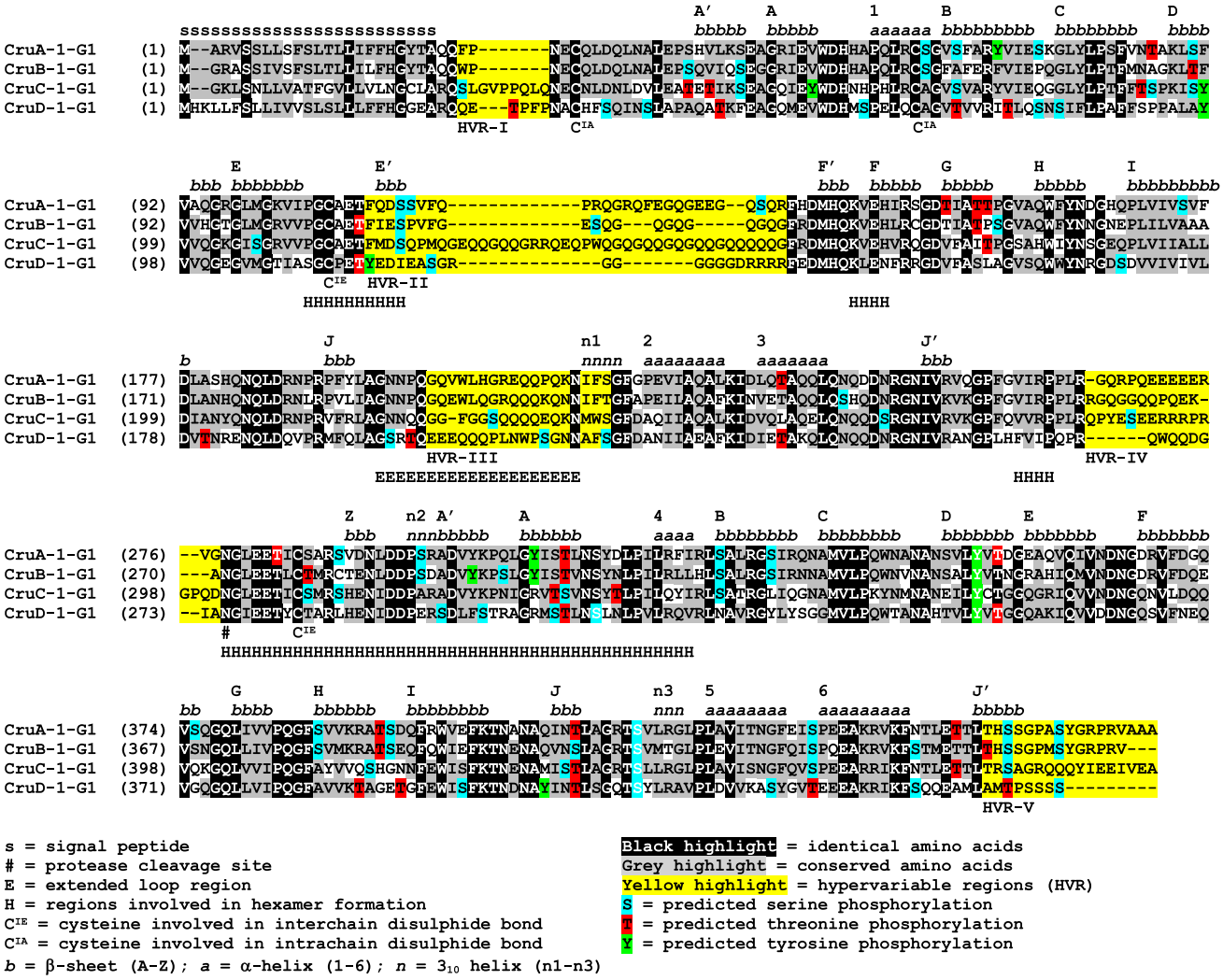

Fig. 5 Alignment and features associated with *C. sativa* cruciferins

while CsCruC was the most hydrophilic cruciferin (GRAVY = − 0.627) and was comparable to *A. thaliana* CruC. This suggests that CsCruC would be the most soluble in aqueous solution, while CsCruD would be the least soluble. The spatial arrangement of hydrophilic and hydrophobic residues on the exposed surfaces was also markedly different for the cruciferin types. The intrachain disulphide bond-containing (IA) faces of CsCruB and CsCruC had negatively charged peripheries with a positively charged central region, while the IA face of CsCruD was dominated by negatively charged amino acids (Fig. 4). As expected, the IA face of all cruciferins were generally hydrophilic; however, in CsCruA, CsCruB and CsCruC, hydrophobic residues tended to occur in small clusters, while those in CsCruD were more evenly distributed across its surface (Fig. 4).

Phosphorylation of cruciferin was first noted in *A. thaliana* (Wan et al. 2007) and now appears to be a general occurrence in seed and vegetative storage proteins (Mouzo et al. 2018). Phosphorylation of serine, threonine or tyrosine was predicted to occur on 23–37 residues in the *C. sativa* cruciferin forms (Fig. 5; Suppl. Fig. S6; Suppl. Table S9). These were predicted to occur within the core structure, on the IE face and on the surface (IA face, periphery and in solvent accessible cavities) (Fig. 4) indicating that this post-translational modification may influence protein folding, subunit interactions, as well as surface-active properties.

An important property for proteins used as food ingredients is their ability to bind/sequester small molecules, such as pigments and flavours. This is related to number, size and chemical properties of pockets in the tertiary and quaternary structure that are accessible to the solvent. The total number of pockets (1.4 Å probe) in the *C. sativa* cruciferin trimers ranged from 214 (CsCruD) to 260 (CsCruC) (Table 6). A larger central pocket forms when the protomers associate to form the trimer and is accessible via an opening on the

Table 5 Expression of *C. sativa* cv. DH55 genes encoding seed storage proteins

Gene	Protein	Gene expression (normalized transcripts per million) at various days post-anthesis									
		4	8	12	16	20	24	28	32	36	40
Csa11g017000	CsNap-1-G1	20	73	4,357	4,672	5,844	2,932	4,672	70	52	127
Csa11g017005	CsNap-2-G1	51	157	14,291	14,291	10,198	24,333	17,574	509	565	711
Csa11g017010	CsNap-3-G1	104	109	25,510	24,333	20,372	14,291	24,333	90	155	154
Csa11g017020	CsNap-4-G1	80	63	20,372	20,372	17,574	17,574	20,372	51	93	97
Csa12g024720	CsNap-1-G3	43	509	17,574	11,281	11,281	11,281	14,291	199	264	308
Csa12g024725	CsNap-2-G3	4	44	2,540	3,248	4,672	1,779	12,490	14	32	27
Csa12g024730	CsNap-3-G3	70	167	24,333	25,510	24,333	20,372	25,510	61	68	120
Csa12g024735	CsNap-4-G3	102	156	30,695	30,695	25,510	25,510	30,695	335	402	446
Csa11g070580	CsCruA-1-G1	13	22	3,768	5,844	6,486	3,768	2,642	110	130	110
Csa11g070590	CsCruA-2-G1	31	14	10,507	10,507	10,507	10,507	10,507	601	647	507
Csa18g009670	CsCruA-1-G2	41	17	10,198	12,490	14,291	12,490	10,198	396	424	390
Csa14g004960	CsCruB-1-G1	20	23	3,036	6,486	6,875	4,672	4,357	130	140	126
Csa03g005050	CsCruB-1-G2	3	1	1,110	2,540	3,469	1,261	683	22	19	16
Csa17g006950	CsCruB-1-G3	23	32	4,672	8,181	8,181	8,181	6,486	32	58	45
Csa11g015240	CsCruC-1-G1	34	13	6,875	10,198	12,490	10,198	98	330	354	207
Csa10g014100	CsCruC-1-G2	11	41	2,642	3,469	5,407	1,590	97	53	56	59
Csa12g021990	CsCruC-1-G3	68	69	12,490	17,574	30,695	30,695	678	1,251	1,261	720
Csa14g004970	CsCruD-1-G1	1	11	1,178	2,642	2,138	717	781	3	4	7
Csa03g005060	CsCruD-1-G2	0	1	337	730	533	123	180	2	3	0
Csa17g006960	CsCruD-1-G3	4	20	892	1,752	2,540	1,296	892	9	12	30
Csa11g019460	CsOle1-1-G1	10	251	2,138	3,600	3,348	3,600	3,469	2,389	2,276	2,138
Csa10g017840	CsOle1-1-G2	16	483	3,348	5,407	4,357	4,869	8,181	3,469	4,357	5,844
Csa12g028090	CsOle1-1-G3	22	346	2,932	4,869	4,093	6,486	5,844	4,672	6,486	6,875
Csa11g057650	CsOle2-1-G1	7	28	1,419	2,430	2,199	2,642	1,972	3,348	3,469	2,430
Csa10g047190	CsOle2-1-G2	5	51	1,251	1,694	1,590	1,844	1,694	1,296	1,752	1,538
Csa12g079570	CsOle2-1-G3	13	52	1,844	2,932	2,752	3,248	4,093	11,281	11,281	10,507
Csa11g082710	CsOle3-1-G1	1	69	775	853	507	371	298	87	75	105
Csa18g022020	CsOle3-1-G2	1	110	885	879	439	196	375	69	62	39
Csa02g041750	CsOle3-1-G3	1	70	1,037	1,261	637	361	360	109	94	108
Csa04g015780	CsOle4-1-G1	6	3	853	1,972	2,642	2,138	2,701	2,752	3,036	3,469
Csa06g008780	CsOle4-1-G2	8	1	862	2,389	3,036	2,701	2,932	1,037	1,251	1,086
Csa09g014800	CsOle4-1-G3	8	0	950	2,701	3,114	2,430	3,036	1,261	1,296	1,280
Csa19g031870	CsVic1A-1-G1	5	1	432	755	2,276	2,752	2,540	4,357	3,248	3,768
Csa01g025880	CsVic1A-1-G2	5	2	178	405	1,678	3,036	2,138	10,198	6,875	10,198
Csa15g039290	CsVic1A-1-G3	3	2	165	359	1,186	2,540	1,635	5,407	4,672	4,869
Csa01g025890	CsVic1B-1-G2	23	10	1	0	0	0	2	0	0	2
Csa15g039300	CsVic1B-1-G3	4	15	43	307	747	631	717	29	31	52
Csa07g016060	CsVic2-1-G1	2	18	730	989	1,972	892	561	402	335	445
Csa16g016660	CsVic2-1-G2	0	0	40	43	50	25	17	26	11	8
Csa05g038120	CsVic2-1-G3	2	17	670	910	1,460	760	650	235	239	337

Scale
0
100
500
1,000
5,000
10,000
20,000
30,000

Table 6 Properties of *B. napus*, *A. thaliana* and *C. sativa* cruciferins

Property*	Cruciferin							
	B. napus 3KGL	*A. thaliana*			*C. sativa*			
		CRA	CRB	CRC	CruA	CruB	CruC	CruD
Protomer								
Formula	$C_{2247}H_{3515}N_{671}O_{696}S_8$	$C_{2200}H_{3442}N_{658}O_{670}S_8$	$C_{2118}H_{3322}N_{616}O_{636}S_{15}$	$C_{2436}H_{3814}N_{734}O_{756}S_{12}$	$C_{2178}H_{3408}N_{644}O_{664}S_7$	$C_{2116}H_{3310}N_{618}O_{647}S_{16}$	$C_{2288}H_{3610}N_{688}O_{710}S_{13}$	$C_{1975}H_{3057}N_{565}O_{612}S_{10}$
Amino acids	466	449	432	501	445	435	469	405
M_r (kDa)	51.3	50.1	48.1	55.9	49.5	48.3	52.5	44.8
pI	6.6	7.26	6.36	6.36	6.41	5.96	6.51	4.99
Negative residues	43 (9.2%)	45 (10.0%)	42 (9.7%)	45 (9.0%)	46 (10.3%)	40 (9.2%)	45 (9.6%)	45 (11.1%)
Positive residues	41 (8.8%)	45 (10.0%)	39 (9.0%)	42 (8.4%)	43 (9.7%)	34 (7.8%)	43 (9.2%)	32 (7.9%)
GRAVY	− 0.557	− 0.562	− 0.432	− 0.691	− 0.487	− 0.46	− 0.627	− 0.375
Total charge	0	− 2	− 5	− 2	− 5	− 8	− 1	− 14
Trimer								
Total pockets	–	228	270	283	221	247	260	214
Central pocket volume (Å^3)	–	17,419.4	9959.7	5092.9	8709.5	17,173.2	4178.7	3070.3
Central pocket area (Å)	–	10,024.4	6755.4	3133.1	4799.9	8821.3	2369.6	1741.2
Central pocket circumference (Å)	–	896.1	449.1	218.4	251.9	733.9	86.4	14.4
Central pocket openings	–	28	15	1	6	15	1	1
Central pocket mouth area (Å)	–	1695.1	762.2	577.7	624.8	1856.0	275.5	12.8

*M_r molecular weight, *pI* isoelectric point, total number of negatively charged residues (Asp + Glu), total number of positively charged residues (Arg + Lys), GRAVY–grand average hydropathy value according to Kyte and Doolittle (1982). Negative scores indicate increasing hydrophilicity, positive scores indicate increasing hydrophobicity

IE face. The size of this pocket size is also a measure of packing efficiency. In homomeric form, CsCruB had the largest pocket volume (17,173.2 Å^3), twice that of CsCruA (8709.5 Å^3) and four-five times that of CsCruC (4178.7 Å^3) and CruD (3070.3 Å^3) (Table 6). The CruB central pocket was also the most accessible with a mouth opening area of 1856.0 Å^2 with 15 individual openings (orientations through which a water molecule may pass). CsCruD had the smallest pocket volume and was also the least accessible with a mouth opening area of only 12.8 Å^2 with one opening. CsCruC had a similar pocket volume with a wider mouth area 275.5 Å^2; however, this was accessible by only a single opening.

Discussion

Current interest in *C. sativa* is mainly centred around oil and its use in bio-fuels (Li and Mupondwa 2014) or as a supplement in animal and fish feeds (Hixson et al. 2014; Hixson and Parrish 2014); however, utilisation of its meal protein (Colombini et al. 2014; Pekel et al. 2015; Hixson et al. 2016a, 2016b) will be necessary to achieve maximal commercial exploitation and valorization. *C. sativa* seed comprises about 43% protein (Zubr 2003), but little or nothing is known about other closely related *Camelina* species. The current study established that *Camelina* species exhibit different seed protein profiles and these differences can separate genotypes representing them. The percent protein in defatted meal also varied between species and less so between lines within the same species. Meal from *C. microcarpa* had the lowest protein content, 31%, while meal from *C. hispida hispida*, *C. laxa*, *C. rumelica transcapida* and some *C. rumelica rumelica* and *C. sativa* lines all reached or

exceeded 40%. This is slightly higher than the 38% reported for canola meal, but less than the 46% for soybean meal (So and Duncan 2021), which have been bred for oil and protein content, respectively.

Lysine and methionine are not synthesised de novo by animals and must be obtained from their diets. These are also limiting in wholly plant-based diets and are often added as supplements to feeds used for monogastric animals, such as fish (Wilson and Halver 1986), poultry (Kidd et al. 1998) and swine (Brinegar et al. 1950). Meals derived from Cruciferous oilseeds generally have higher levels of lysine and methionine than cereals, with *C. sativa* exhibiting a reasonably-balanced essential amino acid profile. Like protein content, amino acid content in the meal also varied between *Camelina* species. Lysine levels were lowest in meal from *C. rumelica rumelica* (4.77%) and highest in most *C. sativa* lines (up to 5.74% in line 1063). Histidine was highest in the meal from *C. rumelica rumelica* (4.77% in line 1034), almost twice that found in meals from any of the other *Camelina* species. Interestingly, the amino acid composition of the two major seed proteins, napin and cruciferin, would account for only about one-half of the total lysine and histidine (Suppl. Table S10) indicating that unincorporated/free amino acids or other proteins of lesser abundance are major contributors to the overall meal amino acid profile. Variation in meal amino acid composition was observed between lines within a species. Methionine and cysteine were highest in meal from *C. rumelica rumelica* lines 609 (2.89%) and 247 (9.32%), respectively, but lowest in *C. rumelica rumelica* line 1034. Serine content was highest in meal from *C. sativa* line 605 meal (5.39%), but lowest in line 252 (4.43%). Threonine was also lowest in meal from *C. sativa* line 1662 (3.83%); however, other *C. sativa* lines exceeded 4.5% similar to other *Camelina* species. This analysis clearly demonstrates that variation among *C. sativa* lines and in related species exists, which could be accessed to develop lines producing meals with amino acid compositions that are better suited for monogastric diets. However, it remains to be demonstrated whether adequate levels of several or all limiting essential amino acids can be achieved in the same genetic background as regulatory mechanisms governing carbon/nitrogen partitioning may not permit this. With respect to essential amino acids, canola meal has comparable levels of histidine (3.39%), isoleucine (3.47), leucine (6.19%), phenyalanine (4.06%), and threonine (4.27), slightly lower levels of lysine (5.92%), and lower levels of cysteine (2.29%), methionine (1.94%), tyrosine (2.50%) and valine (4.97) (Wanasundara et al. 2016) than were found in lines from the various *Camelina* species examined here. It should be noted that differences in analytical techniques must be considered in such comparisons and significant variation in protein and amino acid content has been reported in canola meal from different crushing plants (Le Thanh et al. 2019).

For the most part, variation in seed protein profile between *C. sativa* lines was limited in the 187 accessions examined, which is in keeping with genotypic analyses (Singh et al. 2015; Luo et al. 2019; Chaudhary et al. 2020). This may be attributed to the notion that *C. sativa* is a recent allopolyploid where most homeologous genes are expressed and little sub-genome fractionation has occurred (Kagale et al. 2014, 2016). Despite this, most of the lines could be placed into one of three classes based on differences in the electrophoretic profile of high molecular weight proteins consisting mainly of cruciferin. *C. sativa* possesses 12 genes encoding cruciferin, with each of the three sub-genomes having a contingent of homeologues (Kagale et al. 2014). The 12 *C. sativa* cruciferins are phylogenetically related to the four *A. thaliana* cruciferins, namely AtCRA (At5g44120), AtCRB (At1g03880), AtCRC (At4g28520), and AtCRD (At1g03890). A *CRA* orthologue is not present on any of the *C. sativa* sub-genome G3 chromosomes; however, a tandem duplication occurs on G1 chromosome 11 yielding *CsCruA-1-G1* (Csa11g070580) and *CsCruA-2-G1* (Csa11g070590). Interestingly, the *CsCruB* and *CsCruD* paralogues are also closely linked on each of the sub-genomes, similar to that in *A. thaliana*, even though they are the two most distantly related cruciferins. This signature is suggestive of a duplication event that occurred in a progenitor genome with sufficient time for divergence before the original triplication event that gave rise to the ancestor of both *A. thaliana* and *C. sativa*. It is especially interesting that this arrangement has been maintained through subsequent genome polyploidization and fractionation events in *C. sativa*. The situation with the organisation of napin genes is equally compelling. The *A. thaliana* genome contains 5 genes encoding napin, four of these are linked in tandem on chromosome 4 and are closely related, while the fifth is present on chromosome 5. *Camelina sativa* also has two clusters of four napin genes, one on G1 and the other on G3; no napin genes occur on any G2 chromosomes. This arrangement, however, appears to be coincidental as phylogenetic comparisons between the genes within the *A. thaliana* and *C. sativa* napin clusters indicate that each evolved through a different duplicative route. When the napin gene or gene cluster was lost from G2 might be resolved by examination of genomes from other *Camelina* species (Chaudhary et al. 2020). Two genes encoding vicilin 1 lie in tandem on both G2 and G3, while a single gene is present on G1. This genomic arrangement and phylogenetic analysis suggest that these two sub-genomes are more closely related to one another than to G1, a notion which is supported by genotypic data (Chaudhary et al. 2020).

RNA-Seq analysis of seven *C. sativa* lines revealed that the same homeologues/paralogues encoding napins, oleosins and vicilins were expressed and at similar levels; however,

the expression of cruciferin homeologues/paralouges differed widely between lines in some instances. In the *C. sativa* type strain DH55, genes encoding cruciferins were mainly expressed from the 12th to the 28th day post-anthesis. The general pattern of expression according to transcript levels was *CsCruC* > *CsCruA* > *CsCruB* > *CsCruD*. This same relative expression profile is also present in *A. thaliana* (TAIR; https://www.arabidopsis.org/) and, thus, appeared to be evolutionarily conserved and possibly of functional importance. However, upon examination of six additional *C. sativa* lines, only CN45816 shared this pattern with DH55. In the other five lines, genes encoding CsCruA and CsCruB contributed the majority of the transcripts with those encoding CsCruC and CsCruD providing only a minor component. These general patterns were confirmed by proteomic analysis. The differences in the abundance of cruciferin isoforms/types between the lines has significant consequences as cruciferin is the most abundant seed storage protein and, as such, is the principal contributor to the physiochemical and nutritional properties of meal protein. Cruciferin is a hexamer with the degree of heterogeneity determined by the stoichiometry of the various protomers. While this serves to homogenise the physiochemical properties of individual cruciferin types (Withana-Gamage et al. 2011, 2013a, 2013b, 2015, 2020), it is conceivable that *C. sativa* lines could be selected that produce meals or globulin isolates with properties suited to specific applications. Reduction in the expression of the entire napin gene family via RNA interference (Nguyen et al. 2013) and targeted disruption of homeologous genes encoding CsCruC (Lyzenga et al. 2019) have been successful in altering *C. sativa* seed protein composition and, by inference, the physiochemical properties of the meal. Vicilins are similar to cruciferins in that they are bicupin-domain globulins; however, they remain as trimers similar to the 7S globulins in legumes (Shewry et al. 1995). In *A. thaliana*, the genes encoding vicilins 1 and 2 are expressed at low levels during seed development (TAIR; https://www.arabidopsis.org/) and these proteins likely contribute little to seed protein composition. Conversely, genes encoding CsVic1A were expressed at levels comparable to those encoding CsCruB and moreso than those encoding CsCruC in many of the *C. sativa* lines. Interestingly, neither the *A. thaliana* nor the *C. sativa* vicilin 2 proteins were predicted to contain a signal peptide and are, therefore, unlikely to be deposited within protein storage vacuoles.

Given the sequence and structural similarity between *A. thaliana and C. sativa* cruciferin isoforms, it may be assumed that they share similar physiochemical properties. Cruciferins and other 11S/12S globulins contain two conserved β-barrel or cupin domains; however, the five hypervariable regions confer different properties on individual isoforms (Tandang-Silvas et al. 2010). As noted with *A. thaliana* cruciferins (Withana-Gamage et al. 2011), HVR-I and HVR-III are located on the solvent-exposed surface of the IA face in the hexamer, while HVR-IV moves to the periphery after cleavage at the β-site. In both *A. thaliana* and *C. sativa*, CruC possesses an extended, glutamine-rich, HVR-II within the alpha subunit. In specialised *A. thaliana* lines producing homomeric cruciferins, AtCRC was found to form a compact and less hydrophobic hexamer than either homomeric AtCRA or AtCRB. This resulted in increased thermostability and reduced susceptibility to hydrolysis by pepsin, but altered its ability to form heat-induced gels and to stabilise oil-in-water emulsions (Withana-Gamage et al. 2013a, 2013b, 2015, 2020). Furthermore, reduced proteolytic susceptibility is one of several factors that contribute to the antigenic potential of cupin-like proteins (Mills et al. 2002) making elimination of CsCruC in *C. sativa* an attractive goal (Lyzenga et al. 2019). Homomeric AtCRA and AtCRB formed strong heat-induced gels (Withana-Gamage et al. 2015) and possessed good ability to stabilise oil-in-water emulsions over a wide pH range (Withana-Gamage et al. 2020). Structural features that facilitate flavour or small molecule binding, such as the size of the central pocket and mouth opening (Guichard 2006), were most prominent in CsCruB followed by CsCruA. CsCruD has an unusual HVR-IV that is rich in arginine rather than glutamine residues as in other cruciferin types. Its IA face (solvent-exposed) is dominated by negatively charged amino acids with a more even distribution of hydrophobic residues suggesting that it may possess unique properties. CruD also presents an enigma. It is expressed at very low levels compared to genes encoding other cruciferins. It also possesses alterations in polar and charged residues important for interaction between trimers (Adachi et al. 2001, 2003; Tandang-Silvas et al. 2010), suggesting that it may destabilise hexamers when present. While this may seem counter-intuitive, seed storage proteins must be both stable and be rapidly mobilised during seed germination. Following imbibition, globulin mobilisation is achieved through the sequential hydrolysis of a limited number of internal sites by metallo-endopeptidases followed by a more general degradation by cysteine proteases (Muntz et al. 2001; Tan-Wilson and Wilson 2011). Slight structural instability introduced by CruD may assist in this process when this minor isoform is present and may explain why it remains in *A. thaliana* and *C. sativa*, as well as in other Brassicaceae.

In conclusion, the wealth of information on seed protein diversity in *Camelina* species provided in this work will initially be useful in breeding/engineering lines with higher protein content and amino acid profiles suitable for animal and, possibly, human diets. The plant protein industry is already moving in this direction and beyond, with particular interest in purified protein isolates, mainly albumins (napins) and globulins (cruciferins), for specific food applications (So and Duncan 2021). In the future, knowledge of the genes and

their expression patterns that underlie the protein profiles will permit the creation of specialised *C. sativa* lines that, for example, produce homogeneous cruciferins with properties tailored to specific applications. Indeed, targeted disruption of entire cruciferin gene families, notably CsCruC, has already been demonstrated in *C. sativa* (Lyzenga et al. 2019). It is only a matter of time before this is applied to other oilseed species.

Author contribution statement DD, SM, IAP, AH and JW conceived, designed and funded the research. BG, MH and SP conducted experiments. CC analysed data. DD, IAP and JW wrote the manuscript. All authors read and approved the manuscript.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00425-022-03998-w.

Acknowledgements This work was funded by the Agriculture and Agri-Food Canada Canadian Crop Genomics Initiative and the Global Institute for Food Security.

Funding Open Access provided by Agriculture & Agri-Food Canada.

Declarations

Conflict of interest Authors declare that they do not have any conflict of interest.

Data availability statement The datasets generated during and/or analysed during the current study are deposited in publicly available repositories as indicated or available from the corresponding author upon reasonable request.

References

Adachi M, Takenaka Y, Gidamis AB, Mikami B, Utsumi S (2001) Crystal structure of soybean proglycinin A1aB1b homotrimer. J Mol Biol 305:291–305

Adachi M, Kanamori J, Masuda T, Yagasaki K, Kitamura K, Mikami B, Utsumi S (2003) Crystal structure of soybean 11S globulin: glycinin A3B4 homohexamer. Proc Natl Acad Sci USA 100:7395–7400

Almeada FN, Htoo JK, Thomson J, Stein HH (2013) Amino acid digestibility in camelina products fed to growing pigs. Can J Anim Sci 93:335–343

AOAC Method 972.43. (1997) Microchemical determination of carbon, hydrogen, and nitrogen, automated method. In: Official methods of analysis of AOAC International, 16th edn. AOAC International, Arlington, VA, USA

AACC Method 44–01.01. (1999) Calculation of percent moisture. In: Approved methods of analysis, 11th edition. AACC International, St. Paul, MN, USA. https://doi.org/10.1094/AACCIntMethod-44-01.01

AACC Method 46–18.01. (1999) Crude protein, calculated from percentage of total nitrogen, in feeds and feedstuffs. In: Approved methods of analysis, 11th edition. AACC International, St. Paul, MN, USA. https://doi.org/10.1094/AACCIntMethod-46-18.01

AOAC Method 994.12. (2005) Amino acids in feeds: Performic acid oxidation with acid hydrolysis–sodium metabisulfite method. In: Official methods of analysis of AOAC International, 18th edition. AOAC International, Gaithersburg, MD, USA

Ariza AE, Quezada N, Cherian G (2010) Feeding *Camelina sativa* meal to meat-type chickens: effect on production performance and tissue fatty acid composition. J Appl Poult Res 19:157–168

Barthet VJ, Daun JK (2004) Oil content analysis: myths and reality. In: Luthria DL (ed) Oil extraction and analysis: Critical issues and comparative studies. AOCS Press, Arlington, pp 100–117

Benkert P, Tosatto SCE, Schomburg D (2008) QMEAN: a comprehensive scoring function for model quality assessment. Proteins 71:261–277

Benkert P, Biasini M, Schwede T (2011) Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 27:343–350

Bienert S, Waterhouse A, de Beer TAP, Tauriello G, Studer G, Bordoli L, Schwede T (2017) The SWISS-MODEL repository-new features and functionality. Nucleic Acids Res 45:D313–D319

Booman M, Xu Q, Rise ML (2014) Evaluation of the impact of camelina oil-containing diets on the expression of genes involved in the innate anti-viral immune response in Atlantic cod (*Gadus morhua*). Fish Shellfish Immunol 41:52–63

Brinegar MJ, Williams HH, Ferris FH, Loosli JK, Maynard LA (1950) The lysine requirements for the growth of swine. J Nutr 42:129–138

Chaudhary R, Koh CH, Kagale S, Tang L, Wu SW, Lv Z, Mason AS, Sharpe AG, Diederichsen A, Parkin IAP (2020) Assessing diversity in the *Camelina* genus provides insights into the genome structure of *Camelina sativa*. G3: Genes Genomes Genetics. 10(4):1297–1308

Colombini S, Broderick GA, Galasso I, Martinelli T, Rapetti L, Russo R, Reggiani R (2014) Evaluation of *Camelina sativa* (L.) Crantz meal as an alternative protein source in ruminant rations. J Sci Food Agric 94:736–743

D'Andrea S (2016) Lipid droplet mobilization: the different ways to loosen the purse strings. Biochimie 120:17–27

Deng S-D, Yun G-L, Zhang Q-W, Xu H-L, Cai Q-N (2002) Effect of false flax (*Camelina sativa*) on larval feeding and adult behavioral response of the diamondback moth (*Plutella xylostella*). Acta Entomol Sin 47:474–478

Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, Klebe G, Baker NA (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucleic Acids Res 5:W522–W525

Eisenberg D, Schwarz E, Komaromy M, Wall R (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Biol 179:125–142

Eynck C, Seguin-Swartz G, Clarke WE, Parkin IAP (2012) Monolignol biosynthesis in associated with resistance to *Sclerotinia sclerotiorum* in *Camelina sativa*. Mol Plant Pathol 13:887–899

Guichard E (2006) Flavour retention and release from protein solutions. Biotechnol Adv 24:226–229
Henderson AE, Hallett RH, Soroka JJ (2004) Prefeeding behavior of the crucifer flea beetle, *Phyllotreta cruciferae*, on host and nonhost crucifers. J Insect Behav 17:17–39
Hixson SM, Parrish CC (2014) Substitution of fish oil with camelina oil and inclusion of camelina meal in diets fed to Atlantic cod (*Gadus morhua*) and their effects on growth, tissue lipid classes, and fatty acids. J Anim Sci 92:1055–1067
Hixson SM, Parrish CC, Anderson DM (2014) Full substitution of fish oil with camelina (*Camelina sativa*) oil, with partial substitution of fish meal with camelina meal, in diets for farmed Atlantic salmon (*Salmo salar*) and its effect on tissue lipids and sensory quality. Food Chem 157:51–61
Hixson SM, Parrish CC, Wells JS, Winkowski EM, Anderson DM (2016a) Inclusion of camelina meal as a protein source in diets for farmed Atlantic cod *Gadus morhua*. Aquacult Res 47:2607–2622
Hixson SM, Parrish CC, Wells JS, Winkowski EM, Anderson DM, Bullerwell CN (2016b) Inclusion of camelina meal as a protein source in diets for farmed salmonids. Aqua Nutr 22:615–630
Inquello V, Raymond J, Azana JL (1993) Disulfide interchange reactions in 11S globulin subunits of *Cruciferae* seeds. Eur J Biochem 217:891–895
Kabaha K, Taralp A, Cakmak I, Ozturk L (2011) Accelerated hydrolysis method to estimate the amino acid content of wheat (*Triticum durum* Desf.) flour using microwave irradiation. J Agric Food Chem 59:2958–2965
Kagale S, Koh C, Nixon J, Bollina V, Clarke WE, Tuteja R, Spillane C, Robinson SJ, Links MG, Clarke C, Higgins EE, Huebert T, Parkin SAG, IA, (2014) The emerging biofuel crop *Camelina sativa* retains a highly undifferentiated hexaploid genome structure. Nat Commun 5:3706
Kagale S, Nixon J, Khedikar Y, Pasha A, Provart N, Clarke W, Bollina V, Coutu C, Hegedus DD, Sharpe A, Parkin I (2016) The developmental transcriptome atlas of the biofuel crop *Camelina sativa*. Plant J 88:879–894
Kakani R, Fowler J, Haq AU, Murphy EJ, Rosenberger TA, Berhow M, Bailey CA (2012) Camelina meal increases egg n-3 fatty acid content without altering quality or production in laying hens. Lipids 47:519–526
Kidd MT, Kerr BJ, Halpin KM, McWard GW, Quarles CL (1998) Lysine levels in starter and grower-finisher diets affect broiler performance and carcass traits. Appl Poult Sci 7:351–358
Kita Y, Nakamoto Y, Takahashi M, Kitamura K, Wakasa K, Ishimoto M (2010) Manipulation of amino acid composition in soybean seeds by the combination of deregulated tryptophan biosynthesis and storage protein deficiency. Plant Cell Rep 29:87–95
Kyte J, Doolittle RF (1982) A simple method of displaying the hydropathic character of a protein. J Mol Biol 157:105–132
Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res 33:W299–W302
Le Thanh BV, Beltranena E, Zhou X, Wang LF, Zijlstra RT (2019) Amino acid and energy digestibility of *Brassica napus* canola meal from different crushing plants fed to ileal-cannulated grower pigs. Animal Feed Sci Technol 252:83–91
Li X, Mupondwa E (2014) Life cycle assessment of camelina oil derived biodiesel and jet fuel in the Canadian Prairies. Sci Total Environ 481:17–26
Liang C, Liu X, Yiu SM, Lim BL (2013) *De novo* assembly and characterization of *Camelina sativa* transcriptome by paired-end sequencing. BMC Genomics 14:146
Lill JR, Ingle ES, Liu PS, Pham V, Sandoval WN (2007) Microwave-assisted proteomics. Mass Spectrom Rev 26:657–671
Loiselle DR, Thelin WR, Parker CE, Dicheva NN, Kesner BA, Mocanu V, Wang F, Milgram SL, Esteban Warren MR, Borchers CH (2005) Improved protein identification through the use of unstained gels. J Proteome Res 4:992–997
Luo Z, Brock J, Dyer JM, Kutchan TM, Augustin M, Ge Y, Fahlgren N, Abdel-Haleem H (2019) Genetic diversity and population structure of a *Camelina sativa* spring panel. Front Plant Sci 10:184
Lyzenga WJ, Harrington M, Bekkaoui D, Wigness M, Hegedus DD, Rozwadowski KL (2019) CRISPR/Cas9 editing of three CRUCIFERIN C homoeologues alters the seed protein profile in *Camelina sativa*. BMC Plant Biol 19:292
Mariotti F, Tome D (2008) Converting nitrogen into protein – beyond 6.25 and Jones factors. Crit Rev Food Sci Nutr 48:177–184
Marsolais F, Pajak A, Yin F, Taylor M, Gabriel M, Merino DM, Ma V, Kameka A, Vijayan P, Pham H, Huang SZ, Rivoal J, Bett KE, Hernández-Sebastià C, Liu Q, Bertrand A, Chapman RA (2010) Proteomic analysis of common bean seed with storage protein deficiency reveals up-regulation of sulfur-rich proteins and starch and raffinose metabolic enzymes, and down-regulation of the secretory pathway. J Proteomics 73:1587–1600
Mills ENC, Jenkim J, Marigheto N, Belton PS, Gunning AP, Morris VJ (2002) Allergens of the cupin superfamily. Biochem Soc Trans 30:925–929
Morais S, Edvardsen RB, Tocher DR, Bell JG (2012) Transcriptomic analyses of intestinal gene expression of juvenile Atlantic cod (*Gadus morhua*) fed diets with camelina oil as replacement for fish oil. Comp Biochem Physiol B - Biochem Mol Biol 161:283–293
Moser BR (2012) Biodiesel from alternative oilseed feedstocks: camelina and field pennycress. Biofuels 3:193–209
Mouzo D, Bernal J, López-Pedrouso M, Franco D, Zapata C (2018) Advances in the biology of seed and vegetative storage proteins based on two-dimensional electrophoresis coupled to mass spectrometry. Molecules 23:2462
Mudalkar S, Golla R, Ghatty S, Reddy AR (2014) *De novo* transcriptome analysis of an imminent biofuel crop, *Camelina sativa* L. using Illumina GAIIX sequencing platform and identification of SSR markers. Plant Mol Biol 84:159–171
Muntz K, Belozersky MA, Dunaevsky YE, Schlereth A, Tiedemann J (2001) Stored proteinases and the initiation of storage protein mobilization in seeds during germination and seedling growth. J Exp Bot 52:1741–1752
Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75:4646–4658
Nguyen HT, Silva JE, Podicheti R, Macrander J, Yang W, Nazarenus TJ, Nam JW, Jaworski JG, Lu C, Scheffler BE, Mockaitis K, Cahoon EB (2013) Camelina seed transcriptome: a tool for meal and oil improvement and translational research. Plant Biotechnol J 11:759–769
Nielsen HK, Hurrell RF (1985) Tryptophan determination of food proteins by HPLC after alkaline hydrolysis. J Sci Food Agric 36:893–907
Pekel AY, Kim JI, Chapple C, Adeola O (2015) Nutritional characteristics of camelina meal for 3-week-old broiler chickens. Poult Sci 94:371–378
Schmidt MA, Barbazuk WB, Sandford M, May G, Song Z, Zhou W, Nikolau BJ, Herman EM (2011) Silencing of soybean seed storage proteins results in a rebalanced protein composition preserving seed protein content without major collateral changes in the metabolome and transcriptome. Plant Physiol 156:330–545
Sharma G, Kumar VD, Haque A, Bhat SR, Prakash S, Chopra VL (2002) Brassica coenospecies: a rich reservoir for genetic

resistance to leaf spot caused by *Alternaria brassicae*. Euphytica 125:411–417
Shewry PR, Napier JA, Tatham AS (1995) Seed storage proteins: Structures and biosynthesis. Plant Cell 7:945–956
Shimada K, Matsushita S (1980) Relationship between thermocoagulation of proteins and amino acid compositions. J Agric Food Chem 28:413–417
Singh R, Bollina V, Higgins EE, Clarke WE, Eynck C, Sidebottom C, Gugel R, Snowdon R, Parkin IA (2015) Single-nucleotide polymorphism identification and genotyping in *Camelina sativa*. Mol Breed 35:35
So KKY, Duncan RW (2021) Breeding canola (Brassica napus L.) for protein in feed and food. Plants 10:2220
Soroka J, Olivier C, Grenkow L, Seguin-Swartz G (2015) Interactions between *Camelina sativa* (Brassicaceae) and insect pests of canola. Can Entomol 147:193–214
Stringam GR (1971) Genetics of four hypocotyl mutants in *Brassica campestris* L. J Hered 62:248–250
Suzuki Y, Kawazu T, Koyama H (2004) RNA isolation from siliques, dry seeds, and other tissues of *Arabidopsis thaliana*. Biotechniques 37:542–544
Szumacher-Strabel M, Cieślak A, Zmora P, Pers-Kamczyc E, Bielińska S, Stanisz M, Wójtowski J (2011) *Camelina sativa* cake improved unsaturated fatty acids in ewe's milk. J Sci Food Agric 91:2031–2037
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729
Tandang-Silvas MRG, Fukada T, Fukada C, Prak K, Cabanos C, Kimura A, Itoh T, Mikami B, Utsumi S, Maruyama N (2010) Conservation and divergence on plant seed 11S globulins based on crystal structures. Biochim Biophys Acta 1804:1432–1442
Tan-Wilson AL, Wilson KA (2011) Mobilization of seed protein reserves. Physiol Plant 145:140–153
Troeng S (1955) Oil determination of oilseed. Gravimetric routine method. J Am Oil Chem Soc 32:124–126
Tuan Y-H, Phillips RD (1997) Optimized determination of cystine/cysteine and acid-stable amino acids from a single hydrolysate of casein- and sorghum-based diet and digesta samples. J Agric Food Chem 45:3535–3540
Tuziak SM, Rise ML, Volkoff H (2014) An investigation of appetite-related peptide transcript expression in Atlantic cod (*Gadus morhua*) brain following a *Camelina sativa* meal-supplemented feeding trial. Gene 550:253–263
Ufaz S, Galili G (2008) Improving the content of essential amino acids in crop plants: goals and opportunities. Plant Physiol 147:954–961
van Gunstern WF (1996) Biomolecular simulation: the GROMOS96 manual and user guide. Biomos, Zürich, pp 1044
Vollmann J, Damboeck A, Eckl A, Schrems H, Ruckenbauer P (1996) Improvement of *Camelina sativa*, an underexploited oilseed. In: Janik J (ed) Progress in new crops. ASHS Press, Alexandria, Virginia, pp 357–362
Wan L, Ross ARS, Yang J, Hegedus DD, Kermode AR (2007) Phosphorylation of 12S globulin cruciferin in *Arabidopsis thaliana* seeds of wild type and abi1-1 mutant. Biochem J 404:247–256
Wanasundara JPD, McIntosh TC, Perera SP, Withana-Gamage TS, Mitra P (2016) Canola/rapeseed protein-functionality and nutrition. Oilseed Fats Crops Lipids 23:D407
Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T (2018) SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 46:W296–W303
Wilson AC, Halver JE (1986) Protein and amino acid requirements of fishes. Annu Rev Nutr 6:225–244
Withana-Gamage TS, Hegedus DD, Qiu X, Wanasundara J (2011) In silico homology modeling to predict protein functional properties of cruciferin. J Agric Food Chem 59:12925–12938
Withana-Gamage TS, Hegedus DD, Qui X, McIntosh T, Wanasundara JPD (2013a) Structural and physico-chemical property relationships of cruciferin homohexamers. J Agric Food Chem 61:5848–5859
Withana-Gamage TS, Hegedus DD, Qui X, Yu P, May T, Lydiate D, Wanasundara JPD (2013b) Characterization of *Arabidopsis thaliana* lines with altered seed storage protein profiles using synchrotron powered FTIR. J AGric Food Chem 61:901–912
Withana-Gamage TS, Hegedus DD, Qui X, Wanasundara J (2015) Solubility, heat-induced gelation and pepsin susceptibility of cruciferin protein as affected by subunit composition. Food Biophys 10:103–115
Withana-Gamage TS, Hegedus DD, Qui X, Coutu C, McIntosh T, Wanasundara J (2020) Subunit composition affects formation and stabilization of o/w emulsions by 11S seed storage protein cruciferin. Food Res Int 137:109387
Xue X, Feng CY, Hixson SM, Johnstone K, Anderson DM, Parrish CC, Rise ML (2014) Characterization of the fatty acyl elongase (*elovl*) gene family, and hepatic elovl and delta-6 fatty acyl desaturase transcript expression and fatty acid responses to diets containing camelina oil in Atlantic cod (*Gadus morhua*). Comp Biochem Physiol Part B - Biochem Mol Biol 175:9–22
Xue X, Hixson SM, Hori TS, Booman M, Parrish CC, Anderson DM, Rise ML (2015) Atlantic salmon (*Salmo salar*) liver transcriptome response to diets containing *Camelina sativa* products. Comp Biochem Physiol Part D 14:1–15
Ye CL, Andersen DM, Lall SP (2016) The effects of camelina oil and solvent extracted camelina meal on the growth, carcass composition and hindgut histology of Atlantic salmon (*Salmo salar*) parr in freshwater. Aquaculture 450:397–404
Zubr J (2003) Qualitative variation of *Camelina sativa* seed from different locations. Ind Crops Products 17:161–169