Analytical Biochemistry 665 (2023) 115048

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

Towards absolute quantification of protein genetic variants in *Pisum sativum* extracts

Gijs J.C. Vreeke [a,1], Maud G.J. Meijers [a,b,1], Jean-Paul Vincken [a], Peter A. Wierenga [a,*]

[a] Laboratory of Food Chemistry, Wageningen University and Research, P.O. Box 17, 6700 AA, Wageningen, the Netherlands
[b] TiFN, P.O. Box 557, 6700 AN, Wageningen, the Netherlands

ARTICLE INFO

Keywords:
Quantitative proteomics
Mass spectrometry
SDS-PAGE
Yellow pea
Food peptides

ABSTRACT

In recent years, several studies have used proteomics approaches to characterize genetic variant profiles of agricultural raw materials. In such studies, the challenge is the quantification of the individual protein variants. In this study a novel UPLC-PDA-MS method with absolute and label-free UV-based peptide quantification was applied to quantify the genetic variants of legumin, vicilin and albumins in pea extracts. The aim was to investigate the applicability of this method and to identify challenges in determining protein concentration from the measured peptide concentrations. Analysis of the protein mass balance showed significant losses of proteins in extraction (37%) and of peptides in further sample preparation (69%). The challenge in calculating the extractable individual protein concentrations was how to deal with these insoluble peptides. The quantification approach using average amino acid concentrations in each position of the sequence showed most reproducible results and allowed comparison of the genetic protein composition of 8 different cultivars. The extractable protein composition (μM/μM) was remarkably similar for all cultivar extracts and consisted of legumins A1 (12.8 ± 1.2%), A2 (1.1 ± 0.4%), B (9.9 ± 1.6%), J (7.5 ± 1.0%) and K (10.3 ± 2.1%), vicilin (15.2 ± 1.7%), provicilin (15.7 ± 2.5%), convicilin (9.8 ± 0.8%), albumin A1 (7.4 ± 2.0%), albumin 2 (10.0 ± 1.5%) and protease inhibitor (0.4 ± 0.4%).

1. Introduction

The protein composition of protein concentrates describes the different classes of protein e.g. globulins and albumins, or specific types of globulins e.g. legumin and vicilin present in the sample. In leguminosa seeds, the most abundant proteins are legumins and vicilins, e.g. in pea (*Pisum sativum*) isolates these proteins together represent approximately 53 ± 7% (w/w) of the total protein content [1,2]. The amount and relative ratio of the different types of proteins can be identified using electrophoresis techniques, such as SDS-PAGE [3]. Each type of protein, however, can be present in different genetic variants, which cannot be identified with this technique. For pea, the first reports about the existence of different legumin genetic variants are based on genomic data and date back 30 years [4,5]. Different genetic variants can be identified with proteomics techniques, which are already widely applied on agricultural raw materials such as milk and soy [6,7]. In that field there is quite some discussion on how the mass spectrometry data can be used to determine the absolute amount of individual proteins present (quantification) [8]. By combining mass spectrometric data with UV signals of individual peptides, we have shown that peptides can be accurately quantified in protein hydrolysates using UPLC equipped with a photodiode-array detector (PDA) and MS (UPLC-PDA-MS) [9–11]. The aim of this study is to illustrate the use of this method to obtain an absolute, label-free quantification of the protein composition of complex pea protein samples at genotype-level.

To identify which proteins are present in a sample, (plant) proteins are hydrolysed, the formed peptides are separated by (2D) gel-electrophoresis or liquid chromatography (LC) and analysed using mass-spectrometry (MS). This is similar in proteomics and in the approach described in this study. To quantify the proteins, first the peptides need to be quantified, which could be done via several strategies [12]. When the concentrations of the peptides are determined, they need to be converted to a concentration of a protein genetic variant. The quantification of peptides was initially done relatively to each other, using the MS peak intensities of the HPLC chromatogram, for genetic variants in bovine milk [13]. In this case, the intensity of a proteotypic peptide, which uniquely represented a protein and was (reproducibly)

* Corresponding author.
E-mail address: peter.wierenga@wur.nl (P.A. Wierenga).
[1] Joined first authorship.

https://doi.org/10.1016/j.ab.2023.115048
Received 4 October 2022; Received in revised form 14 December 2022; Accepted 9 January 2023
Available online 16 January 2023

Abbreviations

UPLC-PDA-MS	reverse phase ultra-high performance liquid chromatography-photodiode array-mass spectrometry
SDS-PAGE	sodium dodecyl sulphate polyacrylamide gel electrophoresis
BLP	*Bacillus* licheniformis protease
MQ	Milli-Q water
PPC	pea protein concentrate
PLF	pea legumin fraction
PVF	pea vicilin fraction
PAF	pea albumin fraction
LKE	Lente Krombek extract
VGKE	Vroegste Gele Krombek extract
VLE	Venlosche Lage extract
BelE	Belinda extract
Mir89E	Miranda 1989 extract
Mir20E	Miranda 2020 extract
PalE	Paloma extract
FlaE	Flavandra extract
Mon20E	Montana 2020 extract
Mon06E	Montana 2006 extract
YPE	yellow pea extract
OD	optical density
DTT	dithiothreitol
SPE	solid-phase extraction
TFA	trifluoroacetic acid
LOD	limit of detection
LOA	limit of annotation
PAS	periodic acid-Schiff
L:V	legumin to vicilin
Av	average

formed at high yield [14], was used to compare protein abundances. The downside of relative quantification is that individual MS intensities are highly affected by ion-suppression, matrix effects and day-to-day variation [15–17]. Therefore, later, absolute peptide concentrations were determined by comparing the MS peak intensity of the respective peptide with the intensity of an isotopically labelled reference peptide with similar ionisation properties [18]. The downside of using labelled reference peptides is that it can be costly and laborious. Due to the limited number of reference peptides, only one or a few peptides per protein are quantified. A benefit of this approach is that it has a lower relative standard deviation (RSD) on the calculated protein concentration (<10%) than the MS intensity based approaches (>10%) [19]. An alternative approach to quantification based on MS signals, is the quantification of peptides using the absolute UV absorbance [10,11]. For the absolute quantification of each peptide this method uses the UV peak area at 214 nm and the molar extinction coefficient predicted using the method of Kuipers et al. [9]. The benefit of this technique is that it has the accuracy of quantification techniques with labelling (~6% RSD for peptide concentrations in replicate injections [10]), but does not require chemical or isotopic labelling. A downside of applying UV-based quantification in complex digests, is that for coeluting peptides, the quantification is considerably less accurate [11].

Converting the peptide concentrations to a concentration of a protein genetic variant introduces three challenges.

- Enzymatic hydrolysis of a substrate does not always yield the same peptides at the same concentrations, when different hydrolysis conditions and incubation times are used [20,21]. Variations in digestion methods resulted in large relative standard deviations of 102% up to 1305% in quantification of individual peptides [21]. This could give problems when only one or a few (isotopically labelled reference) peptides are used to quantify a protein. Similarly, more than one peptide (sequence) could be released including the same amino acids of the original protein sequence. Therefore, one should sum peptide concentrations that cover the same amino acids in the protein sequence. Both issues could be solved when all peptides are quantified and used to determine the protein concentration, as for instance done with UV-based quantification or MS intensity based approaches as exponentially modified protein abundance index (emPAI) and intensity based absolute quantification (IBAQ) [11, 22–24].
- A second challenge is how to deal with peptides that are not unique to one genetic protein variant [19,25]. In proteomics analyses, these non-unique peptide sequences are typically excluded from the analysis. The loss of information by excluding these non-unique peptides leads to an underestimation of the protein concentration. The impact will depend on the proportion of non-unique peptide sequences, affected by the similarity between protein sequences.
- Underestimation could also result from peptide losses during sample preparation [26,27] or from intrinsic instability [20,28]. As a result, not all peptides that are formed during enzymatic hydrolysis are included in the analysis. Small molecules as free amino acids, di-peptides and (some) tri-peptides will also be excluded from the analysis since they are not detected in the typical RP-HPLC methods used [29]. The challenge is how to deal with this missing information in calculating protein concentrations.

To estimate the impact of these challenges on the protein quantification it is necessary to have knowledge on the mass balance *e.g.* how much of the initial protein(s) is included in the analysis. This is generally not considered in quantitative proteomics, but could easily be checked by analysis of the protein content before and after sample preparation and centrifugation, or by analysis of the amount of UV in the chromatograms compared with the expected amount based on protein content and composition. In this study the completeness of the analysis is evaluated in detail using the amino acid sequence coverage, UV recovery, matched UV, protein recovery and molar sequence coverage plot as described previously by Butre et al. and Vreeke et al. [10,11].

For pea, genetic variants were only quantified relatively in protein extracts using the volume percentage from 2-D electrophoresis [30]. Bourgeois et al. showed a 2-D map with 626 Coomassie-blue spots, of which 124 were analysed using MS-techniques (Maldi-TOF MS and LC-MS/MS). Altogether, 156 polypeptides were identified, belonging to 55 different proteins. This number can be an overestimation, since precursors, post-translationally cleaved proteins and proteins with modifications were reported as different proteins. Vicilin, convicilin, legumin, and albumin families represented 56% of the polypeptides identified [30]. Several genetic variants were identified, for example for pea legumin, three genetic variants were reported that were reviewed in Uniprot: legumin A2, J and K. In total, out of 11 genes that were described for pea legumin the primary structures of 5 different legumin genetic variants are reported in Uniprot: legumin A1, A2 (former LEG2), J, K and B (Uniprot) [31]. Legumin A1 and A2 are closely related with a sequence similarity of 97.5%. Legumin B is more similar to J and K (Table 1; Uniprot).

This study aims at quantifying the protein composition of pea extracts at genetic variant-level, using the recently developed UPLC-PDA-MS method by Vreeke et al. [11]. The method will be used to calculate protein concentrations using the concentrations of all peptides, based on UV absorbance. The applicability and challenges to quantify proteins will be investigated. The method will be tested on purified pea legumin, vicilin and albumin fractions and afterwards applied to characterize the

Table 1
Protein genetic variants in *Pisum sativum* reported on Uniprot used in peptide annotation screening.

Protein	Uniprot code	Molecular weight (Da)[a]	Sequence similarity (%)[b]	Identified?	Protein	Uniprot code	Molecular weight (Da)[a]	Sequence similarity (%)[b]	Identified?
Legumin A1	P02857	56604.56	–	Yes	Albumin-1 B	P62927	11273.69	96.9 -Alb-1 A	No
Legumin J	P05692	54587.24	40.1- Leg A1	Yes	Albumin-1 C	P62928	11233.73	90.8 -Alb-1 A	No
Legumin K	P05693	39799.74	66.7 - Leg J	Yes	Albumin-1 D	P62929	11237.69	96.9 -Alb-1 A	No
Legumin B	P14594	38989.64	47.2 -Leg K	Yes	Albumin-1 F	P62931	11234.74	96.9 -Alb-1 A	No
Legumin A2	P15838	56967.94	97.5 - Leg A1	Yes	Provicilin A	P02855	31539.98	47.9 - Vic	No
Vicilin	P13918	49340.84	–	Yes	Convicilin B	P13919	43275.25	50.9 -Con A	No
Provicilin B	P02854	44878.06	74.1 - Vic	Yes					
Convicilin A	P13915	63932.05	–	Yes					
Albumin 2	P08688	25620.88	–	Yes					
Albumin-1 A	P62926	11233.66	–	Yes					
IBBB	P56679	7863.97	–	Yes					
PIP20	Q41015	20861.83	–	Yes					

[a] Signal peptide and (post translational) modifications were not taken into account.
[b] The sequence similarity gives the amino acid sequence similarity of a protein and the most similar protein higher in the list for peptide annotation.

genetic protein composition of 8 pea cultivars.

2. Materials and methods

2.1. Materials

Yellow peas (*Pisum sativum* Leguminosae) were purchased from Alimex Europe B.V. (Sint-Kruis, Belgium). The following pea variants (*Pisum sativum* Leguminosae) were provided by the Centrum voor Genetische Bronnen Nederland (CGN, Wageningen, The Netherlands): Lente Krombek (CGN02949), Vroegste Gele Krombek (CGN02950), Venlosche Lage (CGN02962), Belinda (CGN10266), Miranda (CGN10292), Paloma (CGN10296), Flavandra (CGN13290), Montana (CGN24055). *Bacillus licheniformis* protease (BLP) was provided by Novozymes (Bagsvaerd, Denmark). BLP is a serine protease which is able to hydrolyse bonds at the C-terminal of aspartic acid (D) and glutamic acid (E) residues. Previous studies observed 1000x faster hydrolysis after glutamic acid (E) than aspartic acid (D) residues [10,32]. The BLP powder was further treated to remove insoluble material as described by Ref. [33]. In short, a suspension of BLP was made and centrifuged for 10 min (4000×*g*, 25 °C). The supernatant was dialyzed with a 12–14 kDa membrane against 150 mM NaCl and subsequently against demineralised water. Afterwards, the retentate was frozen and freeze-dried. The freeze-dried BLP had a protein content of 60% (w/w, as is) and an activity of 3.9 AU mg^{-1} min^{-2} according to analysis by Deng et al. [34]. SDS-PAGE marker, gels, sample buffer and running buffer were purchased from Bio-Rad Laboratories (Hercules, CA, USA). Coomassie blue stain was purchased form Expedeon (San Diego, CA, USA). A glycoprotein staining kit was purchased from Thermo Scientific (Waltham, MA, USA). Sep-Pak C18 6 cc Vac Cartridges (WAT043395) were purchased from Waters (Milford, MA, USA). All other chemicals were of analytical grade and purchased from either Merck (Darmstadt, Germany), Sigma-Aldrich (St. Louis, MO, USA) or Acros Organics (Geel, Belgium). All water was demineralised (conductivity of 2 μS cm^{-1}) or obtained from a Milli-Q system (Millipore, Billerica, MA, USA; conductivity of 0.5 μS cm^{-1}).

2.2. Methods

2.2.1. Protein isolation and fractionation from yellow pea

2.2.1.1. Preparation of pea protein concentrate. Pea protein concentrate (PPC) was prepared by alkaline extraction followed by iso-electric precipitation, as described by O'Kane [35], with minor alterations. Whole frozen peas (Alimex) were broken with a pin mill (LV 15 M Condux-Werk, Wolfgang bei Hanau, Germany) and subsequently milled (ZPS50 impact mill, Hosokawa-Alpine, Augsburg, Germany). The pea flour (10%, w/w) was suspended in Milli-Q water (MQ). The suspension was adjusted to pH 8.0, followed by centrifugation (17,000×*g*, 4 °C, 20 min). The supernatant was collected and adjusted to pH 4.5, followed by centrifugation (17,000×*g*, 4 °C, 20 min). The pellet was recovered and suspended in MQ at a final concentration of 10% (w/w, wet pellet) and adjusted to pH 8.0. The obtained solution was centrifuged (17,000×*g*, 4 °C, 20 min), and the resulting supernatant was frozen (PPC_{-20}), freeze-dried and named pea protein concentrate (PPC). Prior to all centrifugation steps, suspensions and solutions were kept at 4 °C and the set pH while being stirred for minimally 2 h.

2.2.1.2. Legumin and vicilin fractionation from PPC. The PPC_{-20} was further fractionated to obtain pea legumin fraction (PLF) and pea vicilin fraction (PVF) as described by O'Kane et al. with alterations [35]. The solution was adjusted to pH 8.0 with NaOH, and stirred for 1 h at 4 °C. The solution was subsequently diluted 1:1 with a McIlvaine buffer of pH 4.8, to a final concentration of 200 mM disodium phosphate and 100 mM citric acid containing 200 mM NaCl. The sample was stirred at 4 °C

Table 2
List of protein extracts and samples including their code and Centrum voor Genetische Bronnen Nederland (CGN) number, if applicable.

Code	Samples from CGN peas	CGN-number	Code	Samples non CGN peas
LKE	Lente Krombek extract	CGN02949	YPE	yellow pea extract
VGKE	Vroegste Gele Krombek extract	CGN02950	PPC	pea protein concentrate
VLE	Venlosche Lage extract	CGN02962	PLF	pea legumin fraction
BelE	Belinda extract	CGN10266	PVF	pea vicilin fraction
Mir89E	Miranda 1989 extract	CGN10292	PAF	pea albumin fraction
Mir20E	Miranda 2020 extract	CGN10292		
PalE	Paloma extract	CGN10296		
FlaE	Flavandra extract	CGN13290		
Mon20E	Montana 2020 extract	CGN24055		
Mon06E	Montana 2006 extract	CGN24055		

for minimally 2 h, followed by centrifugation (17,000×*g*, 4 °C, 20 min). The obtained supernatant containing the pea vicilin was filtered using an ultrafiltration system with a 5 kDa membrane (Hydrosart Ultrafilter, Sartorius AG, Frankfurt, Germany). The liquid removed during ultrafiltration was replenished by MQ. The PVF was frozen and freeze-dried. The legumin-rich pellet was resuspended in 20.0 mM Tris-HCl buffer, pH 8.0, (buffer A) at a final concentration of approximately 10 g L^{-1}. The solution was stirred for minimally 2 h, prior to centrifugation (17,000×*g*, 4 °C, 20 min). The obtained supernatant was filtered over a glass fiber pre-filter (13400-142-K, Sartorius) with a Whatman filter paper (black ribbon, 589/1, GE Healthcare, Uppsala, Sweden). The filtrate was applied onto a Source 15Q column (Fineline, Pfizer Manufacturing, Freiburg, Germany) coupled to an ÄKTA explorer system (GE Healthcare). Elution was similar to the method as described by O'Kane et al. and fractions were collected [35]. The fractions rich in legumin were pooled and filtered using an ultrafiltration system with a 5 kDa membrane (Hydrosart Ultrafilter, Sartorius AG). The liquid removed during ultrafiltration was replenished by MQ. The PLF was frozen and freeze-dried.

2.2.1.3. Preparation of pea albumin fraction. The pea albumin fraction (PAF) was isolated by grinding approximately 500 g of yellow peas (Alimex) using a centrifugal mill (Retsch ZM 200, Haan, Germany). The flour was suspended at (20%, w/w) in MQ. The pH of the suspension was adjusted to 8.0. Afterwards the sample was centrifuged (38,400×*g*, 15 min, 20 °C) and the obtained supernatant was adjusted to pH 4.5. The dispersion at pH 4.5 was centrifuged (38,400×*g*, 15 min, 20 °C) and the supernatant was dialysed using an ultra-filtration system with a 10 kDa cut-off, whilst stored on ice. The retentate was subsequently frozen, freeze-dried and labelled pea albumin fraction (PAF). Prior to centrifugation, the samples were stirred for 3 h at room temperature and the pH of the samples was checked regularly and adjusted to the desired pH if necessary.

2.2.2. Preparation of cultivar extracts

Protein was extracted from eight different pea varieties. From two varieties (Miranda and Montana) seeds were included from two different harvest years. Approximately 8–10 g of peas were ground using a centrifugal mill (Retsch ZM 200, Haan, Germany). The flour was suspended (20%, w/w) in MQ containing 2% SDS. The pH of the suspension was adjusted to 8.0, and the samples were stirred for 3 h at room temperature (RT). The pH of the samples was checked regularly and adjusted if necessary. Afterwards the samples were centrifuged (38,400×*g*, 15 min, 20 °C). The supernatants were dialysed against demineralised water using slide-a-lyzers (Thermo Scientific) with a 10 kDa cut-off and subsequently frozen and freeze-dried (Table 2).

2.2.3. Compositional analysis

2.2.3.1. Total nitrogen content. The total nitrogen content was determined in triplicate using the Dumas method (Flash EA 1112 N analyzer, Thermo Scientific), according to manufacturer's protocol. Methionine was used as standard for the nitrogen quantification. For the pea protein extracts, PPC, PLF and PVF a nitrogen conversion factor of 5.4 was used. This was calculated from the average nitrogen conversion factor of the following pea protein genetic variants: legumin A (P02857, Uniprot Database), legumin J (P05692, Uniprot Database), legumin A2 (P15838, Uniprot Database), legumin K (P05693, Uniprot Database), legumin B (P14594, Uniprot Database), and vicilin (P13918, Uniprot Database) [36]. For the PAF a nitrogen conversion factor of 6.22 was used, assuming only albumin 2 (P08688, Uniprot Database) to be present in the sample. This protein content of the samples was calculated assuming all nitrogen originated from protein. The signal peptide was not included in any of the sequences used. In addition it was assumed that there were no post-translational modifications to the proteins.

The protein recovery for extracts, concentrate and isolated fractions was calculated according to equation (1):

$$Protein\ recovery_{Dumas} = \frac{Protein\ in\ sample\ (g)}{Protein\ in\ flour\ (g)} \times 100\ \% \quad (1)$$

2.2.3.2. Protein composition using SDS-PAGE. The protein composition of the samples was determined using SDS-PAGE in the presence and absence of a reducing agent. The samples were diluted to 3 g L^{-1} and analysed according to the manufacturer's protocol. The samples were applied to gels (any kD™, Mini-protean TGX precast protein gels, Bio-Rad Laboratories), and separated on a Miniprotean II system (Bio-Rad Laboratories). The proteins were visualised by staining with Coomassie blue stain (InstantBlue, Expedeon). The gels were scanned and analysed using a densitometer (GS-900™, Bio-rad laboratories) and Image Lab software (Bio-Rad laboratories). Under reducing conditions the following bands were annotated: ~93 kDa lipoxygenase [37], ~70 kDa convicilin, ~50 kDa vicilin, ~38–40 kDa legumin acidic polypeptide, ~33 and 30 kDa vicilin αβ and βγ fragments [35], ~26 kDa albumin 2 [38], ~19–22 kDa legumin basic polypeptide, and ~19, 16 and 13.5 kDa vicilin α, β and γ fragments [35]. Legumin basic polypeptides and vicilin fragments were differentiated from one another, by comparing the gels under reducing and non-reducing conditions. Under non-reducing conditions legumin was present as a monomer consisting of an acidic and basic polypeptide chain, therefore bands of ~57–62 kDa were ascribed to legumin [35]. The intensity of all unidentified bands was summed and the total was referred to as "other proteins". The relative protein composition was determined from the optical density (OD), by dividing to OD of the protein of interest by the total OD in a lane. The relative composition under reducing and non-reducing conditions was averaged. SDS-PAGE analysis was also performed on the PPC, PLF, PVF, PAF, YPE after dithiothreitol (DTT) incubation and after the TFA addition on the supernatant (with and without SPE). The SDS-PAGE analysis was performed under reducing conditions as described above.

2.2.3.3. Detection of glycosylated protein using periodic acid – Schiff's reagent staining. The presence of glycosylated proteins was determined under non-reducing conditions. All samples were dissolved at approximately 2 g L^{-1} protein in MQ containing 2% SDS. Horseradish

peroxidase and soybean trypsin inhibitor were used as a positive and negative control, respectively. The controls were dissolved at 2 g L^{-1} protein in MQ. The proteins were separated using SDS-PAGE as described above. The gels were stained using a glycoprotein staining kit containing periodic acid – Schiff's reagent according to the manufacturer's protocol (Thermo Scientific).

2.2.4. Enzymatic protein hydrolysis

The freeze-dried protein extracts were dissolved at 1.0% (w/v) in 10 mL milli-pore water, adjusted to pH 8.0 and equilibrated for 30 min at 40 °C. The freeze-dried BLP was dissolved at 0.05 mg μL^{-1} of which 30 μL was added to start hydrolysis. The enzymatic hydrolysis was performed in duplicate for 2 h in a pH-stat device (Metrohm, Herisau, Switzerland). This device was used to keep the pH constant by titration of 0.2 M NaOH. Samples of 200 μL were taken before addition of the enzyme and after 10, 30 and 120 min of hydrolysis. The enzymatic hydrolysis was stopped by lowering the pH by addition of 20 μL mL^{-1} hydrolysate of 5 M HCl and changing the pH back after 10 min with 20 μL mL^{-1} hydrolysate of 5 M NaOH. Afterwards, the samples were stored frozen at −20 °C. The degree of hydrolysis was calculated according to equation (2).

$$DH_{stat}[\%] = V_b \times N_b \times \frac{1}{\alpha} \times \frac{1}{m_p} \times \frac{1}{h_{tot}} \times 100\ \% \quad (2)$$

where V_b [mL] is the volume of added NaOH; N_b [mol L^{-1}] is the normality of NaOH; α is the average degree of dissociation of the α-NH group ($1/\alpha = 1.257$ at 40 °C and pH 8 [33]); m_p [g] is the amount of protein in solution; h_{tot} [mmol g^{-1}] is the number of peptide bonds per gram of protein. h_{tot} [mmol g^{-1}] was calculated using the protein composition from SDS-PAGE to be 8.69 for PPC and the extracts, 8.74 for PLF, 8.68 for PVF and 8.77 for PAF.

2.2.5. RP-UPLC-MS analysis

2.2.5.1. Sample preparation for RP-UHPLC-MS. The hydrolysates were diluted (1:1) with a 100 mM Tris-HCl buffer at pH 8.0, containing 20 mM DTT and incubated for minimally 2 h at RT to reduce the disulphide bonds. Afterwards, part of the incubated sample was further processed with solid phase extraction (see section below) and part was used as is. The incubated PLF and PVF (one replicate) were mixed in ratios of 90:10, 75:25, 50:50, 25:75, 10:90 (v/v). The samples and mixtures were acidified by addition of 40 μL of 10% TFA per mL incubated hydrolysate, which lowered the pH to 1–2. The samples were centrifuged for 10 min at 14,000×*g* prior to injection. The PLF, PVF, PAF, PPC (hydrolysis in duplicate) and mixtures of PLF and PVF (originating from one hydrolysis) were injected twice. The hydrolysates of the different cultivars (hydrolysis in duplicate) were injected once.

2.2.5.2. Solid phase extraction (SPE). Solid phase extraction (SPE) was performed using Sep-Pak C18 columns according to the manufacturer's protocol (Waters). The Sep-Pak C18 columns were washed 3 times with 1 mL 50% acetonitrile in MQ and subsequently 3 times with 1 mL MQ, prior to loading the sample. Approximately 1 mL of the samples was loaded onto the columns. The impurities in the samples were removed by washing 3 times with 1 mL MQ and afterwards 3 times with 1 mL 3% acetonitrile in MQ. The peptides were removed from the column by washing with 1 mL 50% acetonitrile in MQ. The acetonitrile solution was evaporated under a N_2-flow. The dried samples were solubilized in 250 μL MQ using ultrasonication for 10 min.

2.2.5.3. Reverse phase ultra-high performance liquid chromatography (RP-UPLC). The hydrolysates were analysed on the Acquity Premier UPLC equipped with a PDA. A gradient was applied of two mobile phases: Eluent A, containing UPLC-grade water with 1% acetonitrile (ACN) + 0.1% TFA and eluent B, containing ACN with 1% water + 0.1% TFA. The gradient was 0–2 min isocratic on 3% B; 2–10 min linear gradient from 3 to 22% B; 10–16 min linear gradient 22–30% B; 16–21 min linear gradient 30–100% B; 21–26 min isocratic on 100% B; 26–28 min linear gradient 100-3% B and 28–32 min isocratic on 3% B. The peptides were separated on the Acquity Premier peptide column, BEH C18 2,1*150 300 A 1.7 μm, with a flowrate of 350 μL min^{-1}. The injected volume was 4 μL. The PDA was used to scan the UV absorbance at fixed wavelength of 214 nm at 1.2 nm resolution and 40 scans/second.

2.2.5.4. Electron spray ionisation time of flight mass spectrometry (ESI-Q-TOF-MS). The mass spectra (50–3000 *m/z*) were collected with the Select Series Cyclic IMS operating in time-of-flight and V-mode (Waters, Milford, MA, USA). The peptides were ionized in the electrospray ionisation source with a capillary voltage of 2.5 kV and a source temperature of 150 °C. The sample cone was operated at 40 V and nitrogen was used as desolvation gas (500 °C, 800 L h^{-1}) and cone gas (200 L h^{-1}). Online lock mass data were acquired by infusing 10 μL min^{-1} of 50 pg μL^{-1} Leucine-Enkephalin via the Waters LockSpray at a capillary voltage of 2.7 kV. The quadrupole was operated using the automatic quad profile. The collision energy applied in the trap was 6 V for the MS, and ramped up in the MSe method from 28 to 56 V for MS/MS. The cyclic ion mobility cell was not used in this experiment. The collision energy in the transfer was 4 V. Prior to analysis, the TOF-analyzer was calibrated up to 4000 *m/z* using sodium iodide.

2.2.6. Data processing

2.2.6.1. Peptide identification. Identification of the peptides was performed in UNIFI software version 1.8 according to the suggested settings in Vreeke et al. [11]. The amino acid sequences of the proteins, so without the signal peptide, in Table 1 were used in UNIFI. All different genetic variants were inserted as unique proteins. Post-translational modifications as oxidation, glycosylation and phosphorylation were not reported for these proteins on Uniprot and therefore not included in this analysis. First, a BLP specific analysis was performed on PPC, PLF, PVF, PAF, PPC and YPE with all potential proteins to evaluate their presence. Protein variants that showed no unique peptides were excluded (albumin variants B,C,D,F, provicilin A and convicilin B). This was done to reduce the number of non-unique sequences. Afterwards, a semi-specific analysis was performed on all samples. The semi-specific analysis included peptides of which either the peptide bond on the C- or on the N-terminal side that was hydrolysed did not match the specificity of BLP (assumed specificity for glutamic acid + aspartic acid). The semi-specific analysis method is essential for good coverage of proteins that naturally occur as linked poly-peptide fragments as vicilin. The protein variants were inserted in the same order as Table 1.

The processing parameters were set based on the guideline of Vreeke et al. [11]. In the peak detection, all *m/z* signals with an intensity above 1000 detector counts were processed, and all MS/MS signals with an intensity of more than 250 detector counts were processed. Peptides were annotated with a maximal acceptable mass error of 10 ppm. After processing, peptides were excluded that did not meet the criteria for MS/MS fragmentation as set by Vreeke et al. [11]. The average limit of detection was determined for the select series Cyclic IMS with a dilution series of a tryptic hydrolysate of alpha-lactalbumin with concentrations 2.5 mg L^{-1} to 5 g L^{-1}. The average MS intensity in the lowest dilution at which the parent ion *m/z* was detected was 1.6 * 10^4 Counts (LOD--lowest level of detection). Peptide annotations were excluded when the MS intensity was below the limit of detection, considering that at this intensity no clear MS/MS spectrum is acquired. In-source fragments, recognised in UNIFI or in PeptQuant, and adducts from water or ammonium were also removed (In this case all peptides annotated > LOD were included in analysis). As shown before a small part of these peptides may not be reproducibly identified in repeated analyses [11]. In the current study, all samples were injected in duplicate, minimizing

Table 3
Yield (%), total protein recovery and legumin:vicilin ratios of protein extracts, concentrate and fractions. Protein content (% w/w, on sample "as is") including standard deviations of pea flours and resulting extracts, concentrate and fractions.

Code	Yield (%)[a]	Total protein recovery (%)[b]	Protein content pea flour (%, w/w "as is")	Protein content extract/concentrate/fraction (%, w/w "as is")[c]	Legumin:vicilin ratio (w/w)[d]
PPC	11	46	17.3 ± 0.2	71.6 ± 0.5	28:72
PLF	2	10	17.3 ± 0.2	84.2 ± 0.9	94:6
PVF	5	21	17.3 ± 0.2	72.0 ± 1.4	6:94
PAF	1	5	17.3 ± 0.2	87.3 ± 0.5	38:62
BelE	22	67	19.3 ± 0.2	59.3 ± 0.7	31:69
FlaE	22	70	18.8 ± 0.1	59.5 ± 0.4	36:64
Mon20E	17	62	16.1 ± 0.5	58.4 ± 0.5	28:72
Mir20E	18	63	16.2 ≤ 0.0	58.6 ± 0.6	33:67
LKE	19	63	18.5 ± 0.2	60.3 ± 0.6	38:62
PalE	18	61	18.1 ± 0.3	60.2 ± 0.3	31:69
VLE	23	70	18.8 ± 0.2	58.5 ± 0.4	31:69
Mon06E	16	60	15.8 ± 0.4	58.9 ± 0.4	31:69
YPE	14	50	17.3 ± 0.2	57.3 ± 0.4	31:69
Mir89E	16	59	15.7 ± 0.1	57.4 ± 1.0	31:69
VGKE	23	70	19.6 ± 0.2	59.9 ± 0.9	41:59

[a] Powder (g)/flour (g) * 100.
[b] Protein in sample (g)/protein in flour (g) * 100.
[c] Determined by Dumas.
[d] Determined by SDS-PAGE densitometry.

possible errors in obtained quantification.

2.2.6.2. Peptide quantification. The absolute concentration of peptides was measured by analysis of the UV absorbance at 214 nm and the molar extinction coefficient of the particular peptide, predicted from Kuipers et al., 2007 [9]. The UV peak areas were integrated in Masslynx version 4.2, with the integration settings as described in Vreeke et al. [11]. The UV peak areas that were originating from the Tris and DTT were removed from the list. The list of peak areas was coupled to the peptide list, taking into account the time offset between PDA and MS (0.08 min) using PeptQuant, an in-house developed Matlab script. The concentration of the peptide was calculated with equation (3).

$$C_{peptide\ [\mu M]} = \frac{A_{214} \bullet Q}{\epsilon_{214} \bullet l \bullet V_{inj} \bullet k_{cell}} \quad (3)$$

where A_{214} [μAU·min] is the UV peak area at 214 nm, V_{inj} [μL] is the volume of sample injected, Q [μL min^{-1}] is the flow rate and l [cm] is the path length of the UV cell, which is 1 cm according to the manufacturer. The molar extinction coefficient ϵ_{214} [L Mol^{-1} cm^{-2}] for each peptide was calculated according to Kuipers et al. [9]. The cell constant, k_{cell} for the UV detector was 0.78 [11]. In case multiple peptides were assigned to the same UV peak, the corresponding area was divided based on the MS intensities and molar extinction coefficients of both peptides [11].

2.2.6.3. Protein quantification. The peptide concentrations were used to calculate the concentration of each protein (variant). To do this, first the concentration of each amino acid occupying a unique position of the protein sequence (unique amino acid) was calculated by summation of the peptide concentrations containing that unique amino acid. The concentration of each unique amino acid was plotted against the sequence of the protein (see results section for an example). If all peptides were completely included in the analysis, the unique amino acid concentrations would be identical for each amino acid in the protein sequence and equal to the initial protein concentration. Typically, variations were observed in the calculated concentrations of unique amino acids. Therefore, three different calculation methods (I-III) were used to calculate the concentration of a protein.

I Concentrations of all unique amino acids were averaged.
II All quantified concentrations of unique amino acids >0 μM were averaged.
III Averaging the concentrations of all unique amino acids with C > average of II.

2.2.6.4. Tools to analyse the completeness of peptide identification and quantification. The completeness of peptide identification was analysed by calculating the amino acid sequence coverage, also known as protein sequence coverage in proteomics [39]. This parameter describes how many of the unique amino acids in a certain protein are covered in at least one of the identified peptides (Equation (4)).

$$Amino\ acid\ sequence\ coverage\ [\%] = \frac{\#\ unique\ annotated\ amino\ acids}{\#\ amino\ acids\ in\ protein\ sequence} \cdot 100\,\% \quad (4)$$

The completeness of quantification was roughly estimated by the UV recovery, which was calculated by dividing the expected amount of UV by the total UV in the chromatogram. The expected amount of UV area was calculated with equation (3), with a correction of the molar extinction coefficient for broken peptide bonds during hydrolysis and the protein concentrations based on protein content and the estimated composition based on SDS-PAGE. To assess the completeness of quantification of each unique amino acid, plots were made of the sum of the absolute peptide concentration involving a certain unique amino acid residue.

3. Results and discussion

3.1. Pea flours and derived fractions: protein content, composition and losses during extraction

The protein contents of the prepared pea flours were: 17.6 ± 1.4% (w/w, on sample as is, Table 3). The protein extracts had protein contents of 58.9 ± 1.0% (w/w, on sample as is). The protein contents of the pea protein concentrate (PPC), pea legumin fraction (PLF), pea vicilin fraction (PVF) and pea albumin fraction (PAF) were higher than those of the crude protein extracts from the different cultivars: 71.6–87.3% (w/w, on sample as is, Table 3). For the different cultivars, the extracted protein represented 58–60% of the total amount of protein in the original sample (Table 3). Other authors report similar extractabilities using a Tris-HCl buffer at pH 8.0 (60 ± 7% [1]). The first challenge in analysis of the protein composition is that ~40% of the proteins in the pea flour are actually not extracted and therefore not analysed. PAS-staining SDS-PAGE gels did not indicate any glycosylated proteins in the extracts, concentrate and fractions (Supplementary Fig. S1). The protein

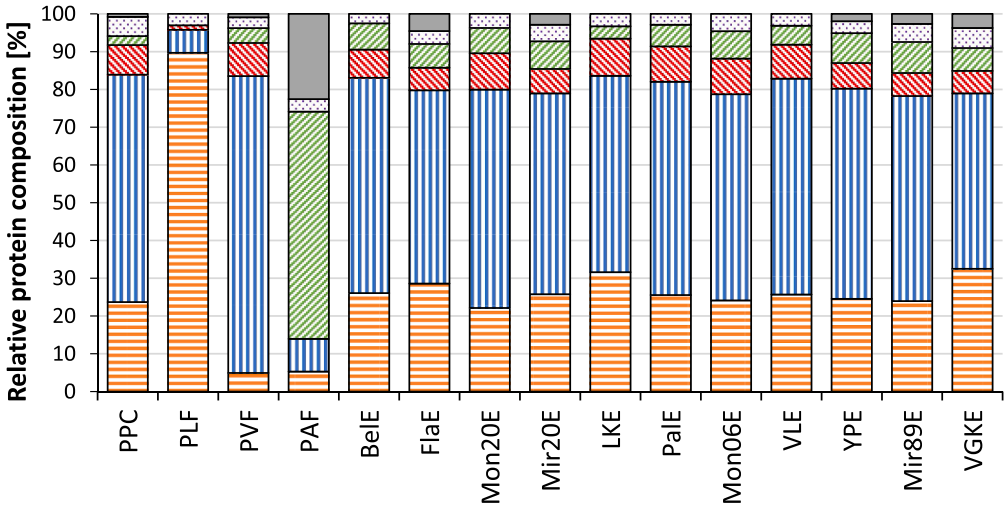

Fig. 1. Relative protein composition (w/w, %) based on densitometry of SDS-PAGE gels showing, legumin (▭), vicilin (▭), albumin (▭), convicilin (▭), lipoxygenase (▭) and other proteins (▭). Values are an average of the results under reducing and non-reducing conditions, with an average standard deviation 1% and a maximum standard deviation of 6%.

Table 4
Amount of integrated UV_{214} area in chromatograms, mean and standard deviations over replicates.

	Expected[a] UV area ($\cdot 10^5$ AU x min)	Total UV ($\cdot 10^5$ AU x min)	Annotated UV peak area ($\cdot 10^5$ AU x min)	Total UV/ Expected UV (%)	Relative Annotated UV area (%)	UV recovery: Annotated area/ expected area (%)
PPC	3.7	2.5 ± 0.0	1.8 ± 0.4	69 ± 0.6	71 ± 15	49 ± 11
PLF	4.9	3.7 ± 0.3	3.1 ± 0.3	76 ± 6.2	85 ± 7	64 ± 7
PVF	3.7	2.6 ± 0.0	2.2 ± 0.1	71 ± 0.4	83 ± 2	59 ± 1
PAF	4.6	1.8 ± 0.1	1.0 ± 0.1	38 ± 1.0	58 ± 5	22 ± 2
PLF:PVF_90:10	4.8	3.4 ± 0.0	2.9 ± 0.1	71 ± 0.4	86 ± 1	60 ± 1
PLF:PVF_75:25	4.6	3.1 ± 0.1	2.4 ± 0.1	68 ± 1.9	77 ± 2	52 ± 3
PLF:PVF_50:50	4.3	2.9 ± 0.0	2.2 ± 0.0	67 ± 0.2	78 ± 1	52 ± 1
PLF:PVF_25:75	4.0	2.5 ± 0.0	1.9 ± 0.1	64 ± 0.2	77 ± 2	49 ± 1
PLF:PVF_10:90	3.8	2.3 ± 0.0	1.7 ± 0.0	60 ± 0.4	76 ± 2	46 ± 1
BelE	3.2	0.9 ± 0.0	0.7 ± 0.0	27 ± 0.1	75 ± 0	20 ± 0
FlaE	3.1	1.3 ± 0.0	1.0 ± 0.0	42 ± 0.1	76 ± 0	32 ± 0
Mon20E	3.1	1.0 ± 0.1	0.8 ± 0.1	33 ± 2.9	75 ± 1	24 ± 2
Mir20E	3.0	0.9 ± 0.0	0.7 ± 0.0	30 ± 0.5	73 ± 1	22 ± 1
LKE	3.2	0.9 ± 0.1	0.8 ± 0.0	29 ± 1.5	82 ± 1	24 ± 1
PalE	3.2	0.9 ± 0.0	0.7 ± 0.0	26 ± 0.7	76 ± 2	20 ± 0
VLE	3.1	1.0 ± 0.1	0.8 ± 0.0	33 ± 3.0	79 ± 1	26 ± 3
Mon06E	3.1	1.0 ± 0.1	0.8 ± 0.0	33 ± 2.9	77 ± 4	25 ± 1
YPE	3.0	0.7 ± 0.0	0.6 ± 0.0	24 ± 0.6	78 ± 1	19 ± 1
Mir89E	3.0	0.9 ± 0.1	0.6 ± 0.1	29 ± 4.6	74 ± 0	21 ± 3
VGKE	3.1	1.0 ± 0.0	0.8 ± 0.1	32 ± 1.2	81 ± 2	26 ± 2

[a] Expected amount was calculated with protein content, estimated molar extinction coefficient based on (SDS-PAGE) protein composition and correction for degree of hydrolysis.

composition of the extracts as analysed by SDS-PAGE stained with Coomassie blue stain showed small differences in the legumin:vicilin ratio (L:V) 28:72–41:59 (w/w, Table 3, Supplementary Fig. S3), but no other differences in presence or absence of specific proteins (Fig. 1). The L:V ratio in PPC, PLF and PVF were 28:72, 94:6 and 6:94 (w/w), respectively. Comparable L:V ratios were obtained using size-exclusion chromatography (A_{214}): PPC 44:56, PLF 100:0, PVF 16:84 (results not shown).

3.2. BLP hydrolysis

The BLP hydrolysis of PPC, PLF, PVF reached a degree of hydrolysis of respectively 6.7 ± 0.2%, 6.7 ± 0.2% and 7.6 ± 0.3%. The hydrolysis of PAF yielded a lower degree of hydrolysis of 4.1 ± 0.2%. For the extracts the final DH was 6.9 ± 0.5%, which indicated that all extracts were hydrolysed to the same extent. The obtained degrees of hydrolysis were 64–85% of the expected value based on the percentage glutamic acid residues in the protein sequences (10.5% for legumin A1, 10.2% vicilin and 4.8% albumin 2). This was in line with the hydrolysis efficiencies observed by Butré et al. for BLP with dairy proteins [10]. SDS-PAGE of the protein hydrolysates showed the presence of intact protein after digestion (Supplementary Fig. S4).

3.3. Quantification of the protein genetic variants and challenges concerned

3.3.1. Peptide losses during sample preparation

The second challenge was the loss of peptides during sample preparation. The total UV peak areas in the chromatograms of the PPC, PLF, PVF and PAF were 68.8 ± 1%, 75.5 ± 6%, 70.8 ± 1%, and 38.0 ± 1.0%

of the expected amounts of UV based on protein content and composition, respectively (Table 4). The chromatograms of the extracts of different cultivars represented 31 ± 5% of the expected total UV. This means that only part of the extracted protein was included in the analysis. In our previous study on tryptic digests of milk protein isolates, the observed amount of UV absorbance was equal to the expected amount [11]. Typically, in (quantitative) proteomics studies, the recovery of injected protein material is not described. An exception, Wang et al. reported also low recoveries of 18–60% for plant protein extracts (barley leaves), dependent on the sample preparation procedure for proteomics analysis [40]. The low UV recoveries observed in the current study were attributed to insoluble aggregates formed when changing the pH to eluent conditions, which were visible as turbidity and then removed by centrifugation. To try and avoid this problem, samples were also prepared by applying solid phase extraction at the pH 8 (after reduction of the disulfide bonds), but the SPE treatment did not improve the UV recovery. The observed UV peak areas ranged between 14 and 53% of the expected UV absorbances. The UPLC-MS data of the same samples with and without SPE treatment did not show changes in *m/z* peaks and ion intensities. Therefore, further analyses were all performed on the dataset without SPE treatment.

3.3.2. Peptide identification in the PLF, PVF, PAF, PPC and YPE

Of the UV peaks that were present in the chromatograms, on average 77 ± 8% was attributed to peptides (Table 4). The highest matched UV was observed for PLF (91%) and the minimum was observed for PAF (54%). The matched UV for the extracts varied between 73 and 82%. UV areas that were not matched with peptide sequences were mostly from remaining intact proteins and phenolic compounds. The number of identified peptides was 301 ± 8 in the PLF, 293 ± 7 in the PVF, 98 ± 9 in the PAF, 264 ± 37 in the PPC and 186 ± 9 in the YPE. For 78 ± 3% of these peptides the MS intensity was above the average limit of annotation, which was previously used to indicate annotations with high repeatability [11]. To be as complete as possible for this study we also included the peptide annotation < limit of annotation (LOA, but these were all still > limit of detection (LOD) and confirmed with sufficient MS/MS fragments). Between replicate injections of PLF, PVF and PAF, 86 ± 2% of all the peptides were annotated similarly between replicates. Between duplicate hydrolyses of the same fractions, 85 ± 3% of the peptides were annotated similarly. This means that the variation in peptide composition between hydrolysates of replicate digestions, did not exceed the variation between replicate measurements of the same hydrolysate. The repeatability for duplicate injections in this study was higher (86%) than repeatability at peptide level observed in proteomics studies, where typically 35–79% were similarly annotated between duplicate injections [41–44].

Table 5
Amino acid sequence coverages (%) + standard deviation of pea proteins identified in PPC, PLF, PVF, PAF and YPE.

Protein	PPC	PLF	PVF	PAF	YPE
Legumin A1	53 ± 21	91 ± 4	47 ± 10	3 ± 1	34 ± 3
Legumin A2	14 ± 2	28 ± 2	9 ± 4	3 ± 0	2 ± 2
Legumin B	41 ± 8	66 ± 3	39 ± 11	2 ± 1	29 ± 1
Legumin J	30 ± 4	38 ± 10	13 ± 3	2 ± 1	24 ± 3
Legumin K	41 ± 8	69 ± 8	34 ± 7	2 ± 1	25 ± 4
Vicilin	71 ± 9	56 ± 5	80 ± 3	1 ± 0	55 ± 2
Provicilin	48 ± 24	48 ± 9	62 ± 4	14 ± 16	27 ± 0
Convicilin	50 ± 2	44 ± 4	53 ± 3	3 ± 2	35 ± 2
Albumin A1	46 ± 0	12 ± 2	71 ± 0	67 ± 6	46 ± 0
Albumin 2	47 ± 12	16 ± 15	36 ± 6	97 ± 1	31 ± 3
IBBB	0 ± 0	0 ± 0	5 ± 0	3 ± 0	3 ± 0
PIP20	5 ± 0	10 ± 8	20 ± 12	13 ± 16	3 ± 0

3.3.3. Completeness of peptide identification

The peptides identified for each protein genotype were visualised against the sequence of the protein, as illustrated with legumin A1 in PLF (Fig. 2). In some cases, part of the protein sequence was annotated in multiple peptides. For example, amino acid Leucine on position 1 for legumin A1 was present in peptides 1–3 and 1–9. In other cases, part of the sequence was not covered by any of the annotated peptides, e.g.

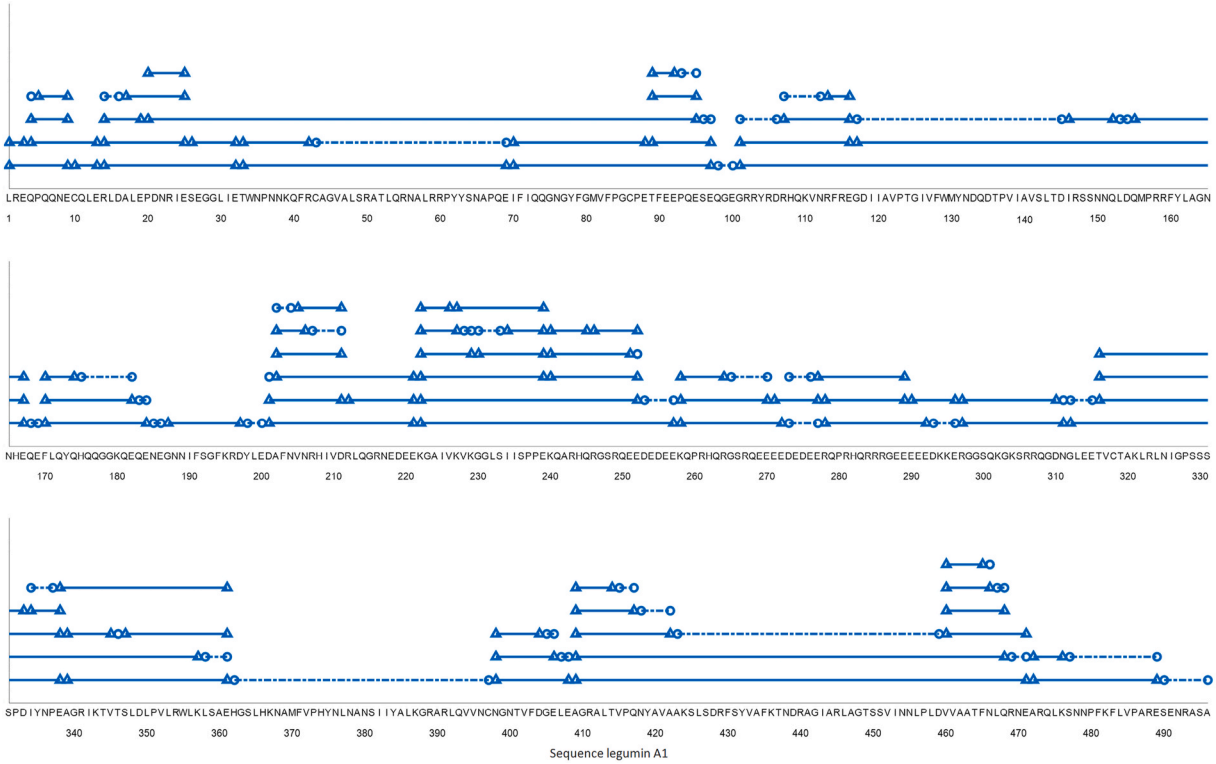

Fig. 2. Peptides of legumin A1 identified in PLF, visualised against the protein sequence of legumin A1. Dotted lines indicate missing peptides.

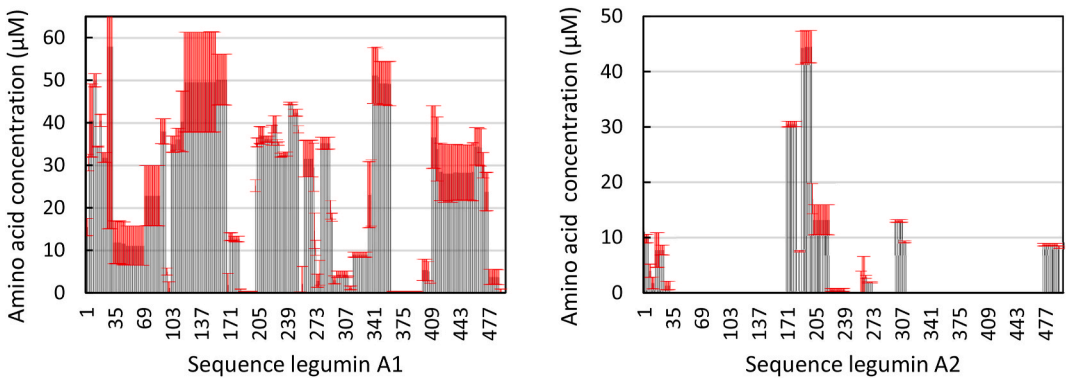

Fig. 3. The unique amino acid concentration (μM) + standard deviation (μM, in red) for legumin A1 (left) and A2 (right) in PLF, calculated with the peptide concentrations including that respective amino acid.

legumin A1: 362–397. The protein sequences of the most abundant proteins in the fractions were covered with high amino acid sequence coverages, respectively 91 ± 4% for legumin A1 in the PLF, 80 ± 3% for vicilin in the PVF and 97 ± 1% for albumin 2 in the PAF (Table 5). Substantial sequence coverages for the legumin variants A2, B, J and K (28–69%) confirmed that legumin was present in different genetic variants. The peptides identified in PVF yielded amino acid sequence coverages of 80 ± 3% for vicilin, 62 ± 4% for provicilin and 53 ± 3% for convicilin. For the two protease inhibitors that were included in the analysis the coverages were relatively low in all purified fractions (≤20%). The amino acid sequence coverages observed in this study were lower than coverages reported in previous studies with hydrolysates of 1–3 milk proteins (amino acid sequence coverages of 99–100%) [10,11], analysed with the same procedure. The amino acid sequence coverages were logically affected in the pea hydrolysates by the observed peptide losses in sample preparation. Furthermore, sample complexity could be relevant (higher number of peptides in similar injection volume), leading to lower concentrations of individual peptides. For instance, the coverage of legumin A1 in YPE (34 ± 3%), is ~1/3 of the coverage of legumin A1 in the purified legumin fraction from the same pea cultivar (PLF, 91 ± 4%).

3.3.4. *Peptide and protein quantification in the purified fractions*

All peptides in the PLF, PVF and PAF were quantified based on their UV absorbance and predicted molar extinction coefficient, which yielded a wide variety of individual peptide concentrations. For instance,

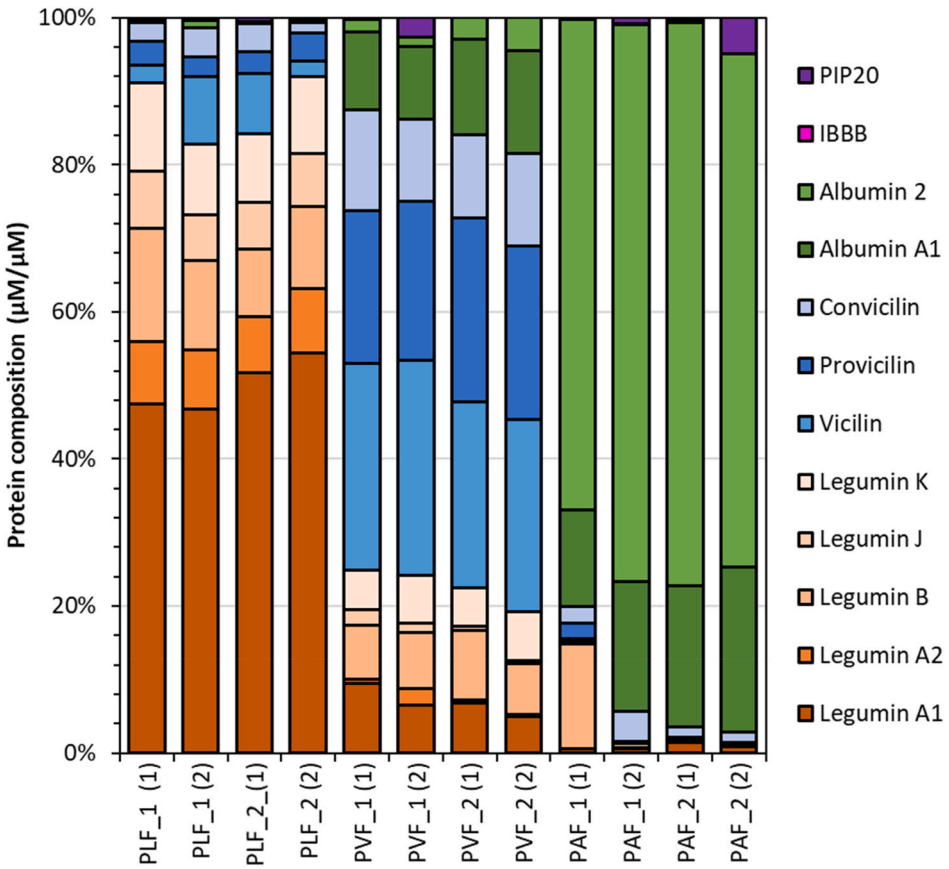

Fig. 4. Composition μM/μM (%/%) of the PLF, PVF and PAF determined with calculation I. Reproducibility was tested in duplicate hydrolyses (first number) and duplicate injections (second number).

the different peptides originating from legumin A1 in PLF had concentrations ranging between 0.4 nM (for peptide 258–264) to 55 μM (for peptide 26–32). This implies that using one of these quantified peptides as reference for the protein concentration, as done often with isotopically labelling, would yield very different protein concentrations dependent on the reference peptide chosen. Therefore, to tackle this challenge, we choose to sum all peptide concentrations to determine the protein concentration. To visualise this way of protein quantification, the amino acid concentrations for each position of the sequence were plotted against the amino acid sequence of each protein (Fig. 3). For legumin A1 in PLF, the standard deviation of the observed amino acid concentration in four replicates was 4%. The average amino acid concentration was 25 ± 3 μM over the four replicates. In absence of losses during sample preparation, the retrieved amino acid concentration would be similar to the (molar) concentration of protein before hydrolysis. The composition of the PLF, PVF and PAF were determined using the average amino acid concentration as indication for the protein concentrations (calculation I) (Fig. 4). With this calculation, minor differences were observed between replicate analyses and major differences in the relative protein composition for the different fractions were observed. The PLF consisted for 87 ± 5% of the legumins, the PVF had 62 ± 1% of vicilin, provicilin and convicilin and the PAF had 72 ± 5% of albumin 2.

3.4. Correcting protein concentrations for missing peptide information

3.4.1. Generic peptide sequences

The third challenge was that some peptide sequences can occur in several protein variants. This could lead to overestimation of one variant, and underestimation of the other variant in protein quantification. For the 134 peptides annotated to legumin A1 or A2 in PLF, only 1 peptide could also originate from the sequence of legumin B, J or K. Out of these 133 peptides unique for legumin A, 29 were unique for genetic variant A1, 20 were unique for genetic variant A2 and 75 peptides could originate from either legumin A1 or A2. Since these 75 peptides were -in the software-now attributed to legumin A1, the amino acid sequence coverage for A2 had large parts of the sequence where no peptides were attributed to (Fig. 3). Therefore, the concentrations of proteins should ideally be calculated with peptides that were unique for that genetic variant.

3.4.2. How to tackle missing peptide information in calculated protein concentrations

Two additional calculations were evaluated to transform the peptide concentrations into an absolute protein concentration, taking into account the unique sequences per variant (calculation II) and possibly low recovery of part of the sequence (calculation III). Based on protein content and the SDS-PAGE composition, 67 μM of legumins were expected to be present in the PLF and 57 μM of vicilins would be present in the PVF (Table 6). These values are higher than what was calculated from averaging the amino acid concentrations of quantified peptides; legumins in PLF (43 ± 3 μM) and vicilins in PVF (20 ± 1 μM). A correction for non-unique sequences, which was done by calculating the concentrations with only peptides that were unique for a certain variant, overestimated the concentration of the minor legumin variants (A2,B,J, K) in the PVF. When the protein concentrations were calculated with all amino acids that were above the average, which was done to exclude the sequences of the protein that were quantified clearly lower than the maximum in the plot, the observed concentration of legumins was almost 2x the (maximum) expected concentration. Similarly, an unrealistic amount of legumins was observed in the PVF. This error seems to originate from parts of the protein sequence of legumins that were quantified at concentrations far above the (observed) average amino acid concentrations. This was for instance the case for sequences of legumin B: 43–50, 83–90 and 103–107 and legumin A2: 170–182, 187–200. For these peptides the identification might be incorrect. For the LegA2 peptides 170–182 and 187–200 the identifications were confirmed with 19 and 21 MS/MS fragments, respectively, which excludes the possibility of having alternative annotations. For legumin B 103–107, with sequence KEEED, an isobaric alternative assignment would be provicilin DKEEE of 307–311. Leg B peptides 43–50 and 83–90 did not have alternative assignments either. These flaws in the analysis of the peptides, in combination with the losses during sample preparation will affect the absolute concentrations calculated. However, it is expected that the current flaws will be similar for all samples and have thereby a minor effect on the conclusions on relative differences in genetic composition.

3.4.3. Genetic composition of mixtures of PLF and PVF

The fractions PLF and PVF were mixed in different ratios to evaluate the robustness in the genetic protein composition. For the majority of the calculated amino acid concentrations, the height changed gradually with the amount of the protein in the mixture and the overall the pattern of the plot remained similar (Fig. 5). This means that the majority of the peptides were identified and quantified consistently. Exceptions were for example the legumin B sequences, 83–90 and 103–107 mentioned in the previous section: For these, the maximum concentrations were not observed in the PLF. Regardless of the calculation used, the analysed composition of the 50:50 mixture was a good representative of the purified PLF and PVF fractions. Since the composition of the purified fractions was most reproducible when determined with calculation I, so without correction for the missing peptide information, this calculation was also used to determine the composition of the different pea cultivars.

3.5. Variation in genetic composition of different pea cultivars

For the pea cultivar extracts, approximately 20 ± 5% of the total protein in the peas was analysed by RP-UHPLC-MS, based on the measured UV absorbance, amount of injected sample, and protein contents of the extracts and flours. In total, 15 ± 4% of the total protein in the different pea cultivars was annotated. The protein composition of

Table 6
Illustrating the effect of different approaches to convert peptide to protein concentrations (μM) in PLF and PVF based on UPLC-PDA-MS (calculation I-III) and protein composition based on SDS-PAGE.

	SDS-PAGE		Calculation I (Av)[a,d]		Calculation II (Av>0)[b,d]		Calculation III (Av > Av)[c,d]	
	PLF	PVF	PLF	PVF	PLF	PVF	PLF	PVF
Legumin A1	–	–	24.7 ± 3.2	2.8 ± 0.7	27.1 ± 1.3	3.1 ± 1.3	41.3 ± 3.5	16.3 ± 14.2
Legumins (A1, A2, B, J, K)	66.7	3.2	43.0 ± 2.8	9.2 ± 0.9	67.0 ± 4.9	28.3 ± 6.4	116.0 ± 7.9	27.2 ± 2.2
Vicilin	–	–	2.8 ± 2.0	11.1 ± 0.6	4.8 ± 3.3	13.8 ± 0.6	12.9 ± 3.2	23.1 ± 14.1
Vicilins (Vicilin + provicilin)	5.2	57.4	4.4 ± 1.9	20.4 ± 0.9	8.1 ± 3.5	28.9 ± 2.7	67.6 ± 10.4	51.8 ± 2.3

[a] Concentrations of all unique amino acids were averaged.
[b] All quantified concentrations of unique amino acids >0 μM were averaged.
[c] Averaging the concentrations of all unique amino acids with C > average of II.
[d] For calculations, see materials and methods section 2.2.6.3.

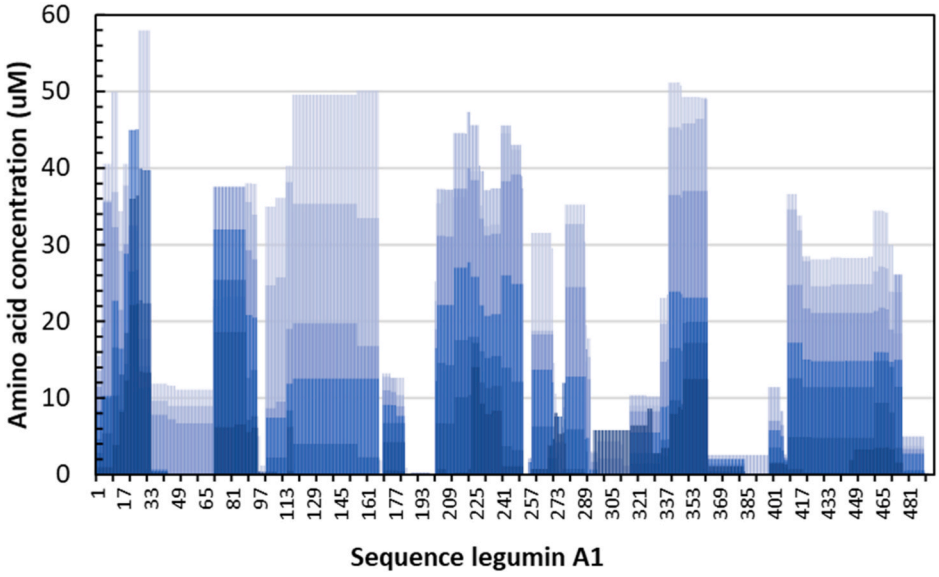

Fig. 5. Amino acid concentration (μM) observed for legumin A1 for different PLF:PVF ratios: 100:0 ( ), 90:10 ( ), 75:25 ( ) 50:50 (25:75 ( ), 10:90 ( ) and 0:100 ( ).

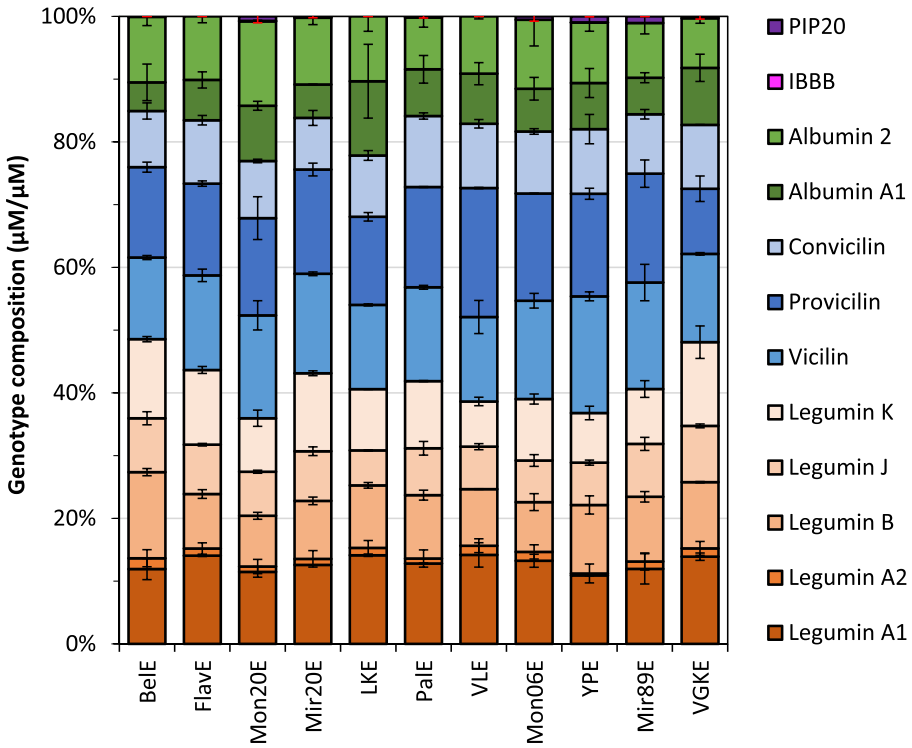

Fig. 6. Absolute protein composition μM/μM (%/%) on genetic-variant level for extracts of different pea cultivars determined with calculation I.

the analysed part of the different pea cultivar extracts showed no significant differences (Fig. 6). Based on the calculation using all amino acids the cultivars were composed of 12.8 ± 1.2% legumin A1, 1.1 ± 0.4% legumin A2, 9.9 ± 1.6% legumin B, 7.5 ± 1.0% legumin J, 10.3 ± 2.1% legumin K, 15.2 ± 1.7% vicilin, 15.7 ± 2.5% provicilin, 9.8 ± 0.8% convicilin, 7.4 ± 2.0% albumin A1, 10.0 ± 1.5% albumin 2 and 0.4 ± 0.4% PIP20. The L:V ratio found in these samples was 57:43% (μM μM^{-1}), whereas approximately 70% of the samples described in literature and measured with SDS-PAGE has a ratio ranging between 17:83–38:62 (w/w %) ([1,2,45–47]. The ratio determined with UPLC-PDA-MS reflected the differences in recovery of legumins (65%) compared with vicilin (36%) in the purified fractions. Legumin A1, A2,

B, J and K were present in all samples, and legumin A (A1 + A2) was most abundant, 14.1 ± 1.4%. The other legumin genetic variants occurred in relatively similar quantities in each pea cultivar and the composition (n/n %) between the genetic variants was comparable. Besides IBBB, all the protein genetic variants considered in this study were found in all pea cultivars in similar quantities. A recent study by Burstin et al. provided insights into the pea genome [48]. Our study shows that the genes responsible for the production of the proteins considered in this study are expressed in all cultivars in similar quantities.

4. Conclusion

In this study, a new way of protein quantification was illustrated, in which all peptides in the analysis were used to quantify protein genetic variants. Using UV quantification, we were not limited to a small number of reference peptides, enabling the quantification of all peptides, normally only achieved with MS intensity-based quantification. Analysis of the protein mass balance showed losses during sample preparation and protein extraction. This allowed to describe how much of the original pea protein was described by the analysis. With the approach taken, the high impact of wrong annotations on calculated protein concentrations was identified. Without correcting for the peptide losses, differences in composition were still reproducibly determined for fractions of legumins, vicilins and albumins, as well as their mixtures. For the pea protein extracts from different cultivars this method showed that all considered protein genetic variants were present in similar amounts.

CRediT authorship contribution statement

Gijs J.C. Vreeke: Conceptualization, Methodology, Formal analysis, Investigation, Visualization, Writing – original draft. **Maud G.J. Meijers:** Conceptualization, Methodology, Formal analysis, Investigation, Visualization, Writing – original draft. **Jean-Paul Vincken:** Supervision, Writing – review & editing. **Peter A. Wierenga:** Conceptualization, Software, Supervision, Writing – review & editing.

Declaration of competing interest

The project is partially organized by and executed under the auspices of TiFN, a public - private partnership on precompetitive research in food and nutrition. The authors have declared that no competing interests exist in the writing of this publication. Funding for this research was partially obtained from Bel S.A., Nutricia Research B.V., Pepsico Inc., Unilever Nederland holdings B.V., the Netherlands Organisation for Scientific Research and the Top-sector Agri&Food.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ab.2023.115048.

References

[1] M.B. Barać, S. Čabrilo, M.B. Pešić, S.P. Stanojević, S. Žilić, O. Maćej, et al., Profile and functional properties of seed proteins from six pea (*Pisum sativum*) genotypes, Int. J. Mol. Sci. 11 (12) (2010) 4973–4990, https://doi.org/10.3390/ijms11124973.

[2] E.N. Tzitzikas, J.-P. Vincken, J. de Groot, H. Gruppen, R.G.F. Visser, Genetic variation in pea seed globulin composition, J. Agric. Food Chem. 54 (2) (2006) 425–433, https://doi.org/10.1021/jf0519008.

[3] V. García Arteaga, S. Kraus, M. Schott, I. Muranyi, U. Schweiggert-Weisz, P. Eisner, Screening of twelve pea (*Pisum sativum L.*) cultivars and their isolates focusing on the protein characterization, functionality, and sensory profiles, Foods 10 (4) (2021), https://doi.org/10.3390/foods10040758.

[4] P.R. Shewry, R. Casey, Seed Proteins. Seed Proteins, Springer, 1999, pp. 1–10.

[5] A.J. Thompson, D. Bown, S. Yaish, J.A. Gatehouse, Differential expression of seed storage protein genes in the pea legJ subfamily, Sequence of Gene legK. Biochemie und Physiologie der Pflanzen 187 (1) (1991) 1–12, https://doi.org/10.1016/S0015-3796(11)80177-0.

[6] R. Agregán, N. Echegaray, M. López-Pedrouso, R. Kharabsheh, D. Franco, J. M. Lorenzo, Proteomic advances in milk and dairy products, Molecules 26 (13) (2021), https://doi.org/10.3390/molecules26133832.

[7] N.L. Houston, D.-G. Lee, S.E. Stevenson, G.S. Ladics, G.A. Bannon, S. McClain, et al., Quantitation of soybean allergens using tandem mass spectrometry, J. Proteome Res. 10 (2) (2011) 763–773, https://doi.org/10.1021/pr100913w.

[8] N. Colaert, J. Vandekerckhove, L. Martens, K. Gevaert, A Case Study on the Comparison of Different Software Tools for Automated Quantification of Peptides. Gel-free Proteomics, Springer, 2011, pp. 373–398, https://doi.org/10.1007/978-1-61779-148-2_25.

[9] B.J.H. Kuipers, H. Gruppen, Prediction of molar extinction coefficients of proteins and peptides using UV absorption of the constituent amino acids at 214 nm to enable quantitative reverse phase high-performance liquid chromatography-mass spectrometry analysis, J. Agric. Food Chem. 55 (14) (2007) 5445–5451, https://doi.org/10.1021/jf070337l.

[10] C.I. Butré, S. Sforza, H. Gruppen, P.A. Wierenga, Introducing enzyme selectivity: a quantitative parameter to describe enzymatic protein hydrolysis, Anal. Bioanal. Chem. 406 (24) (2014) 5827–5841, https://doi.org/10.1007/s00216-014-8006-2.

[11] G.J.C. Vreeke, W. Lubbers, J.-P. Vincken, P.A. Wierenga, A method to identify and quantify the complete peptide composition in protein hydrolysates, Anal. Chim. Acta 1201 (2022), 339616, https://doi.org/10.1016/j.aca.2022.339616.

[12] K.W. Lau, A.R. Jones, N. Swainston, J.A. Siepen, S.J. Hubbard, Capture and analysis of quantitative proteomic data, Proteomics 7 (16) (2007) 2787–2799, https://doi.org/10.1002/pmic.200700127.

[13] H. Jensen, N. Poulsen, K. Andersen, M. Hammershøj, H. Poulsen, L. Larsen, Distinct composition of bovine milk from Jersey and Holstein-Friesian cows with good, poor, or noncoagulation properties as reflected in protein genetic variants and isoforms, J. Dairy Sci. 95 (12) (2012) 6905–6917, https://doi.org/10.3168/jds.2012-5675.

[14] S. Keerthikumar, S. Mathivanan, Proteotypic peptides and their applications, in: S. Keerthikumar, S. Mathivanan (Eds.), Proteome Bioinformatics, Springer New York, New York, NY, 2017, pp. 101–107, https://doi.org/10.1007/978-1-4939-6740-7_8.

[15] T.M. Annesley, Ion suppression in mass spectrometry, Clin. Chem. 49 (7) (2003) 1041–1044, https://doi.org/10.1373/49.7.1041.

[16] H. Trufelli, P. Palma, G. Famiglini, A. Cappiello, An overview of matrix effects in liquid chromatography–mass spectrometry, Mass Spectrom. Rev. 30 (3) (2011) 491–509, https://doi.org/10.1002/mas.20298.

[17] P.D. Piehowski, V.A. Petyuk, D.J. Orton, F. Xie, R.J. Moore, M. Ramirez-Restrepo, et al., Sources of technical variability in quantitative LC–MS proteomics: human brain tissue sample analysis, J. Proteome Res. 12 (5) (2013) 2128–2137, https://doi.org/10.1021/pr301146m.

[18] C. Bär, D. Mathis, P. Neuhaus, D. Dürr, W. Bisig, L. Egger, et al., Protein profile of dairy products: simultaneous quantification of twenty bovine milk proteins, Int. Dairy J. 97 (2019) 167–175, https://doi.org/10.1016/j.idairyj.2019.01.001.

[19] W.X. Schulze, B. Usadel, Quantitation in mass-spectrometry-based proteomics, Annu. Rev. Plant Biol. 61 (2010) 491–516, https://doi.org/10.1146/annurev-arplant-042809-112132.

[20] C.I. Butré, S. Buhler, S. Sforza, H. Gruppen, P.A. Wierenga, Spontaneous, non-enzymatic breakdown of peptides during enzymatic protein hydrolysis, Biochim. Biophys. Acta Protein Proteonomics 1854 (8) (2015) 987–994, https://doi.org/10.1016/j.bbapap.2015.03.004.

[21] M.S. Lowenthal, Y. Liang, K.W. Phinney, S.E. Stein, Quantitative bottom-up proteomics depends on digestion conditions, Anal. Chem. 86 (1) (2013) 551–558, https://doi.org/10.1021/ac4027274.

[22] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, et al., Global quantification of mammalian gene expression control, Nature 473 (7347) (2011) 337–342, https://doi.org/10.1038/nature10098.

[23] I. Battisti, L.B. Ebinezer, G. Lomolino, A. Masi, G. Arrigoni, Protein profile of commercial soybean milks analyzed by label-free quantitative proteomics, Food Chem. 352 (2021), 129299, https://doi.org/10.1016/j.foodchem.2021.129299.

[24] Y. Ishihama, Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber, et al., Exponentially Modified Protein Abundance Index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein*S, Mol. Cell. Proteomics 4 (9) (2005) 1265–1272, https://doi.org/10.1074/mcp.M500061-MCP200.

[25] M. Blein-Nicolas, H. Xu, D. Vienne, C. Giraud, S. Huet, M. Zivy, Including shared peptides for estimating protein abundances: a significant improvement for quantitative proteomics, Proteomics 12 (2012) 2797–2801, https://doi.org/10.1002/pmic.201100660.

[26] A. Kraut, M. Marcellin, A. Adrait, L. Kuhn, M. Louwagie, S. Kieffer-Jaquinod, et al., Peptide storage: are you getting the best return on your investment? Defining optimal storage conditions for proteomics samples, J. Proteome Res. 8 (7) (2009) 3778–3785, https://doi.org/10.1021/pr900095u.

[27] K.L. Zapadka, F.J. Becher, A.L. Gomes Dos Santos, S.E. Jackson, Factors affecting the physical stability (aggregation) of peptide therapeutics, Interface Focus 7 (6) (2017), 20170030, https://doi.org/10.1098/rsfs.2017.0030.

[28] M. Planyavsky, M.L. Huber, N.A. Staller, A.C. Müller, K.L. Bennett, A longitudinal proteomic assessment of peptide degradation and loss under acidic storage

conditions, Anal. Biochem. 473 (2015) 11–13, https://doi.org/10.1016/j.ab.2014.11.020.
[29] C.I. Butré, Introducing Enzyme Selectivity as a Quantitative Parameter to Describe the Effects of Substrate Concentration on Protein Hydrolysis [PhD Thesis Wageningen University for the Degree of Doctor in the Year 2014, Wageningen University, Wageningen, 2014.
[30] M. Bourgeois, F. Jacquin, V. Savois, N. Sommerer, V. Labas, C. Henry, et al., Dissecting the proteome of pea mature seeds reveals the phenotypic plasticity of seed protein composition, Proteomics 9 (2) (2009) 254–271, https://doi.org/10.1002/pmic.200700903.
[31] R. Casey, C. Domoney, Pea globulins, in: P.R. Shewry, R. Casey (Eds.), Seed Proteins, Springer Netherlands, Dordrecht, 1999, pp. 171–208.
[32] K. Breddam, M. Meldal, Substrate preferences of glutamic acid-specific endopeptidases assessed by synthetic peptide substrates based on intramolecular fluorescence quenching, Eur. J. Biochem. 206 (1) (1992) 103–107, https://doi.org/10.1111/j.1432-1033.1992.tb16906.x.
[33] C.I. Butré, P.A. Wierenga, H. Gruppen, Influence of water availability on the enzymatic hydrolysis of proteins, Process Biochem. 49 (2014) 1903–1912, https://doi.org/10.1016/j.procbio.2014.08.009.
[34] Y. Deng, C.I. Butré, P.A. Wierenga, Influence of substrate concentration on the extent of protein enzymatic hydrolysis, Int. Dairy J. 86 (2018) 39–48, https://doi.org/10.1016/j.idairyj.2018.06.018.
[35] F.E. O'Kane, R.P. Happe, J.M. Vereijken, H. Gruppen, M.A.J.S. van Boekel, Characterization of pea vicilin. 1. Denoting convicilin as the α-subunit of the *Pisum* vicilin family, J. Agric. Food Chem. 52 (10) (2004) 3141–3148, https://doi.org/10.1021/jf035104i.
[36] UniProtKB. http://www.uniprot.org/. 04/06/2020.
[37] U. Szymanowska, A. Jakubczyk, B. Baraniak, A. Kur, Characterisation of lipoxygenase from pea seeds (*Pisum sativum* var. Telephone L.), Food Chem. 116 (4) (2009) 906–910, https://doi.org/10.1016/j.foodchem.2009.03.045.
[38] T.J.V. Higgins, L.R. Beach, D. Spencer, P.M. Chandler, P.J. Randall, R.J. Blagrove, et al., cDNA and protein sequence of a major pea seed albumin (PA 2 : Mr≈26 000), Plant Mol. Biol. 8 (1) (1987) 37–45, https://doi.org/10.1007/BF00016432.
[39] B. Meyer, D.G. Papasotiriou, M. Karas, 100% protein sequence coverage: a modern form of surrealism in proteomics, Amino Acids 41 (2) (2011) 291–310, https://doi.org/10.1007/s00726-010-0680-6.
[40] W.-Q. Wang, O.N. Jensen, I.M. Møller, K.H. Hebelstrup, A. Rogowska-Wrzesinska, Evaluation of sample preparation methods for mass spectrometry-based proteomic analysis of barley leaves, Plant Methods 14 (1) (2018) 1–13, https://doi.org/10.1186/s13007-018-0341-4.
[41] D.L. Tabb, L. Vega-Montoto, P.A. Rudnick, A.M. Variyath, A.-J.L. Ham, D.M. Bunk, et al., Repeatability and reproducibility in proteomic identifications by liquid chromatography− tandem mass spectrometry, J. Proteome Res. 9 (2) (2010) 761–776, https://doi.org/10.1021/pr9006365.
[42] M. Berg, A. Parbel, H. Pettersen, D. Fenyö, L. Björkesten, Reproducibility of LC-MS-based protein identification, J. Exp. Bot. 57 (7) (2006) 1509–1514, https://doi.org/10.1093/jxb/erj139.
[43] N. Delmotte, M. Lasaosa, A. Tholey, E. Heinzle, A. van Dorsselaer, C.G. Huber, Repeatability of peptide identifications in shotgun proteome analysis employing off-line two-dimensional chromatographic separations and ion-trap MS, J. Separ. Sci. 32 (8) (2009) 1156–1164, https://doi.org/10.1002/jssc.200800615.
[44] C.C. Tsou, C.F. Tsai, G.C. Teo, Y.J. Chen, A.I. Nesvizhskii, Untargeted, spectral library-free analysis of data-independent acquisition proteomics data generated using Orbitrap mass spectrometers, Proteomics 16 (15–16) (2016) 2257–2271, https://doi.org/10.1002/pmic.201500526.
[45] R. Casey, J.E. Sharman, D.J. Wright, J.R. Bacon, P. Guldager, Quantitative variability in *Pisum* seed globulins: its assessment and significance, Plant Foods Hum. Nutr. 31 (4) (1982) 333–346, https://doi.org/10.1007/BF01094045.
[46] J. Gueguen, J. Barbot, Quantitative and qualitative variability of pea (*Pisum sativum* L.) protein composition, J. Sci. Food Agric. 42 (3) (1988) 209–224, https://doi.org/10.1002/jsfa.2740420304.
[47] H.E. Schroeder, Quantitative studies on the cotyledonary proteins in the genus *Pisum*, J. Sci. Food Agric. 33 (7) (1982) 623–633, https://doi.org/10.1002/jsfa.2740330707.
[48] J. Burstin, J. Kreplak, J. Macas, J. Lichtenzveig, *Pisum sativum* (pea), Trends Genet. 36 (4) (2020) 312–313, https://doi.org/10.1016/j.tig.2019.12.009.