Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/4/8/eaar8602/DC1

Supplementary Materials for

Genome mapping of seed-borne allergens and immunoresponsive proteins in wheat

Angéla Juhász, Tatiana Belova, Chris G. Florides, Csaba Maulis, Iris Fischer, Gyöngyvér Gell, Zsófia Birinyi, Jamie Ong, Gabriel Keeble-Gagnère, Amudha Maharajan, Wujun Ma, Peter Gibson, Jizeng Jia, Daniel Lang, Klaus F. X. Mayer, Manuel Spannagl; International Wheat Genome Sequencing Consortium, Jason A. Tye-Din, Rudi Appels*, Odd-Arne Olsen*

*Corresponding author. Email: rudi.appels@unimelb.edu.au (R.A.); odd-arne.olsen@nmbu.no (O.-A.O.)

Published 17 August 2018, *Sci. Adv.* **4**, eaar8602 (2018)
DOI: 10.1126/sciadv.aar8602

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/4/8/eaar8602/DC1)

Data file S4 (Microsoft Excel format). Expression of peptides with known IFNγ-ELISPOT responses.
Science article authors that are IWGSC members (Microsoft Word format)

Supplementary material 1

Section S1. Manual annotation and curation of the Prolamin superfamily genes in the reference genome

Using the high-quality reference genome combined with a comprehensive analysis workflow we have identified and precisely characterized the members of prolamin superfamily. Protein classes identified include the low-molecular weight glutenins (LMW glutenins), the α- and γ-gliadins, the minor prolamin classes purinins and ALPs, the prolamin-superfamily members of ATIs, LTPs, nsLTPs, proline-rich proteins, hydrophob-seed domain containing proteins, egg-cell secreted proteins, and cortical cell delineating proteins. Additionally, high molecular weight glutenins (HMW glutenins) and domainless protein types including ω-gliadins, 19kDa Globulins and small cysteine-rich proteins were also identified. Seven hundred and six out of the 828 genes were present in the IWGSC Refseq v.1.0 annotation, 508 of which lack introns. The B sub-genome was more enriched in genes belonging to the prolamin superfamily. Clusters of 2 – 25 genes of the hydrophobic-seed domain proteins, proline-rich proteins, and cortical cell delineating proteins were present on all chromosomes. Gliadin and glutenin genes - primarily affecting nitrogen and sulphur storage, grain quality, and health responses - show a clear monophyletic origin and are clustered on homeologous group 1 and 6 chromosomes *(1)*, Data file S2). In addition, we refined gene models for 37 ω-gliadins using the PacBio whole genome shotgun sequence information *(1)*. The gene models were verified by checking for the presence of a signal peptide region and a variable number and positions of cysteine residues aligning to published prolamin sequences as well as by transcriptome data *(38)*. Thirty-three per cent of the γ-gliadins, 39% of the α-gliadin and 57% of the ω-gliadins sequences contain in-frame stop codons, thus representing potential pseudo genes. The majority of the prolamin superfamily genes, like gliadins, glutenins, puroindolines, avenin-like proteins and ATIs are grain specific, in contrast to the lipid transfer protein and hydrophobic seed domain-containing proteins that are also expressed in vegetative tissues including roots.

Genes encoding the gliadin Pfam domain (PF13016)

A total of 75 genes encoding gliadin pfam domain containing proteins (α-gliadins, γ-gliadins, LMW glutenins, purinins, gliadin-like proteins, ALPs and a 19kDa globulin) were identified and annotated (Data file S2). The major prolamin loci encoding γ-gliadins, LMW glutenins and the domainless ω-gliadins locate at the short arm of homeologues chromosome group 1 (fig. S1A, B). α-gliadins are located at the short arms of chromosome group 6 (fig. S1C). Genes encoding minor protein classes, such as ALPs (chromosome 4AL, 7AS and 7DS), 19kDa seed globulins (chromosome 1AL, 1BL and 1DL), purinins (chromosome 1AS, 1BS and 1DS), and gliadin-like proteins (chromosome 3AL, 3BL, 3DL) were also mapped to their chromosomal loci (Data file S1).

Genes encoding proteins with Protease inhibitor/seed storage/ LTP family (PF00234) domains

Altogether 158 gene models representing alpha amylase/trypsin inhibitors (ATIs), α-gliadins, puroindolines, grain softness proteins and non-specific lipid transfer proteins were identified. They are distributed across all chromosomes. Among the α-gliadins, more than 60% possess the PF00234 domain and they are present on all three sub-genomes. We have identified 35 ATI genes encoding monomeric, dimeric, and tetrameric forms of this protein family (Data file S2).

Genes encoding Probable lipid transfer protein (PF14368) domains

Gene models representing 298 LTP2 domain containing proteins were identified in the reference genome and they mainly encode lipid transfer and non-specific lipid transfer proteins ((*1),* fig. S3). Among the annotated 298 gene models encoding LTP2 domains, LTPs involved in various responses during cell growth and cell elongation (LTP Extensin-like), senescence (LTP YLS3-like), signaling in systemic acquired resistance (LTP DIR1) were identified in bread wheat and related species (*(16),* fig. S4). Non-specific lipid transfer proteins related to cell growth and cell wall elongation (GPI-anchored nsLTP, nsLTP PR60, nsLTP PR61) were identified, some of them with tissue specificity in reproductive tissues, while others expressed in roots (nsLTP PR61).

Hydrophobic seed proteins (PF14547) and Prolamin-like domain containing proteins (PF05617)

Hydrophobic seed domain containing proteins are represented by 14kDa and 36.4kDa proline-rich proteins (PRPs), pEARLI1-like proteins (LTP pEARLI1-like), hydrophobic seed proteins (HSDCP) and cortical cell delineating proteins (WRPs). Seventy-six genes encoding proline rich proteins and 95 genes encoding cortical cell delineating proteins were annotated. A large gene cluster of proline rich proteins with 12 genes was identified on chromosome 7D and a gene cluster encoding twenty-four cortical cell delineating proteins was identified on chromosome 4B. Majority of these genes show root specific expression and are related to root hair growth and cell expansion in roots. A few genes are also related to systemic acquired assistance (LTP pEARLI1-like). We also identified 6 genes with Prolamin-like domain, all of them encoded at chromosome group 7. These represent egg-cell secreted proteins (EC1.1 and EC1.2) and are involved in double fertilization.

HMW glutenins and sequences encoding prolamins without Pfam domains

HMW glutenins are a well characterised gene family with significant effect on end use quality encoded at the long arms of chromosome group 1. Prolamin superfamily genes without known Pfam domains are difficult to detect using automated annotations and they were therefore identified manually. ω-gliadins represent a sulphur-poor group of cereal prolamins without a functional protein domain structure. They are encoded in the major prolamin loci in the short arm of homologous group 1. Altogether 37 gene models were identified in Chinese Spring, many of them incomplete sequences that were corrected using the Pacbio whole genome shotgun sequencing results *(21)* (Fig. 1).

Section S2. Organization of major prolamin gene clusters of short arms of chromosomes 1 and 6

Precise orientation of chromosome region at the short arm of chromosome 1D

When initially examined in the IWGSC RefSeq v1.0 assembly the regions containing γ- and ω-gliadin, LMW-glutenin coding genes as well as one or two purinin coding genes appeared to vary extensively from each other in a synteny-based comparison. Inspection of the 1D assembly in the region indicated that the orientation of scaffolds and their relative positioning was ambiguous and when the scaffolds were rearranged to maximise alignment with 1AS and 1BS (confirmed by alignments of the segments from another variety) it was evident that overall there exists a high level of collinearity between the terminal regions of 1AS, 1BS and 1DS) (*(1),* fig. S1 D). Interestingly, gene copies of γ-gliadins and LMW glutenins are harbored between two clusters of ω-gliadin genes on chromosome arms 1AS and 1BS, while no second ω-gliadin gene cluster was found on 1DS (fig. S1 B). Similarly, loci orders were consistent except 1DS, where LMW glutenins were distributed in a region covering 6.15 Mb. Copies of Berberine-bridge enzyme-like protein coding genes along with NLR genes and PR60 non-specific lipid transfer proteins were also identified in the first 30 Mbp regions (fig. S1 A).

Identification of the major α-gliadin gene cluster at chromosome 6D

α-gliadins are located at the short arms of chromosome group 6 ((fig. S1C). Although no α-gliadin genes were previously annotated on IWGSC RefSeq v1.0 chromosome 6D our annotation shows that the unassigned scaffold collection (Chromosome Un) was enriched in regions encoding various prolamin superfamily member genes, including 31 α-gliadin genes and 8 ω-gliadins.

Correct chromosomal location of these genes was determined using blast analysis against the Chinese Spring genome assembly of the Earlham Institute (2016, UK), closely related species with sequenced genomes (*T. urartu, T. monococcum, Ae. speltoides, Ae. sharonensis, Ae. tauschii* and *T. durum*) and the assembled genome of cultivar AK58. A major cluster of α-gliadin genes is located on a single scaffold representing a genomic region of 3.64 Mbp in length (Chromosome Un Scaffold 53990) (fig. S1C). Genes with various annotated functions such as signal-peptide peptidase, glutamate receptor function, peptidyl-prolyl-cis-trans-isomerase, glycosyl transferase and serine/threonine receptor kinase were also mapped to the same scaffold. The conserved position of homeologue genes on the short arm of chromosome 6A and 6B confirmed that Chromosome Un scaffold 53990 represented the missing α-gliadin locus of chromosome 6D. The correct position of scaffold 53990 at position 26Mb in IWGSC RefSeq v1.0 Chromosome 6D was confirmed using the assembled 6D chromosome of cultivar AK58 (fig. S1E).

Section S3. Phylogenetic analysis of Prolamin superfamily genes

The allocation of the prolamin super-family into evolutionary clades using their Pfam domain information provided three broad mono-phyletic clades within the Triticeae species as well as other grasses such as rice, brachypodium, maize and sorghum (fig. S3). Hydrophobe-seed domain containing proteins (subtree labelled in blue) represent a few major protein families primarily expressed in the roots and related to cell wall development (Proline-rich proteins), root hair elongation (cortical cell delineating proteins), and stress defense mechanisms (glycine-rich proteins) (fig. S3). Gliadin and Tryp-alpha amyl domain containing proteins (subtree colored in red) represent the major seed protein types primarily related to food hypersensitive reactions with a molecular function in nutrient storage and protease inhibitor activities. Tryp-alpha amyl containing proteins – specifically ATIs and Pins – correspond with defense in all Poaceae species. Lipid transfer proteins and non-specific lipid transfer proteins belong to a third large clade and represent various molecular functions related to cell growth, cell elongation, senescence, lipid transport, and plant defense. Interestingly, classic prolamin protein types such as gliadins and glutenins were strongly under-represented in the tree, which is due to their expansion in gene numbers in the Triticeae *(1)* and partially due to the lack of correct annotation of these gene families in the publicly available genome sequences (fig. S3).

Section S4. Phylogenetic analysis of highly immunogenic α-gliadin sequences

The presence of the α-gliadin specific 33-mer peptide was investigated along with its five overlapping composing epitopes (fig. S4 B). The region was positioned at 76-108 aa in the Chinese Spring α-gliadin sequences. Based on the available sequence set the 33-mer is present in 21 sequences of the 534 sequences analysed α-gliadin sequences. The component epitopes were generally frequent in the α-gliadins of bread wheat, *Ae. tauschii*, *T. urartu* and some secalins of rye. In wheat, they were mostly characteristic of α-gliadins from the D genome. While the DQ2.5-glia-α1a and DQ2.5-glia-α2 epitopes were both present in *T. urartu* and *Ae. tauschii* sequences, none of the *T. urartu* sequences contained DQ2.5-glia-α1b peptides. In some of the *Ae. tauschii* α-

gliadins and rye secalins all three peptides were present however the number of peptide copies was less than in the alpha 33-mer (Fig. 3).

Section S5. Impact of temperature stress on protein composition

Size exclusion and reversed phase HPLC measurements of grain protein quantity and composition in two Norwegian bread wheat cultivars Bjarne and Berserk and the reference cultivar Chinese Spring showed significant effects of temperature stress applied at flowering. Table S1 shows changes in grain protein content of the major wheat protein classes, gliadins, glutenins and albumins/globulins. In this study, grain protein content (%) was similar under normal growth conditions in all studied cultivars but was differently modified by high and low temperatures (table S1). All three cultivars responded with an overall 1-3% protein content increase with the increase of the temperature during flowering, while 3.5-4.45% loss in protein content was observed when low temperature stress was applied.

Low temperature resulted in reduced gliadin accumulation by 24.1% (from 10.8% to 8.2%), 25% (from 10.05% to 7.54%) and 27.4% (from 10.18% to 7.39%) in Berserk, Chinese Spring and Bjarne, respectively. The percentage of glutenin protein also decreased with the largest drop in Bjarne by 20% and Chinese Spring by 31% and less in Berserk (by 13.9%) (table S1).
All three cultivars showed increased protein levels in response to high-temperature stress. While the change in protein content was greater in Berserk (by 19.5%) and CS (by 18.3%) it was less notable in Bjarne (by 6.8%). Interestingly, this smaller impact of high temperature stress in Bjarne was explained by almost no change in the amount of glutenins (table S1). Results show that high temperature stress increased the gliadin amount by 2.18g, 0.85g and 1.91g in 100g of flour which corresponds to 20.2%, 8.3% and 19% increase in gliadin content in Berserk, Bjarne and Chinese Spring, respectively. The gliadin composition was also differently affected in the three cultivars. In Berserk the increment was primarily caused by an increase of the proportion of α-gliadin content (1.12%), followed by the γ-gliadins (0.80%) and ω-gliadins (0.25%). A more balanced change was observed in Chinese Spring, with difference of 0.9% in α-gliadin levels, 0.49% in γ-gliadins and 0.52% in ω-gliadins. Gliadin composition of Bjarne was less affected by the high temperature conditions and was mainly due to changes in the γ-gliadin content.

Supplementary figures and table

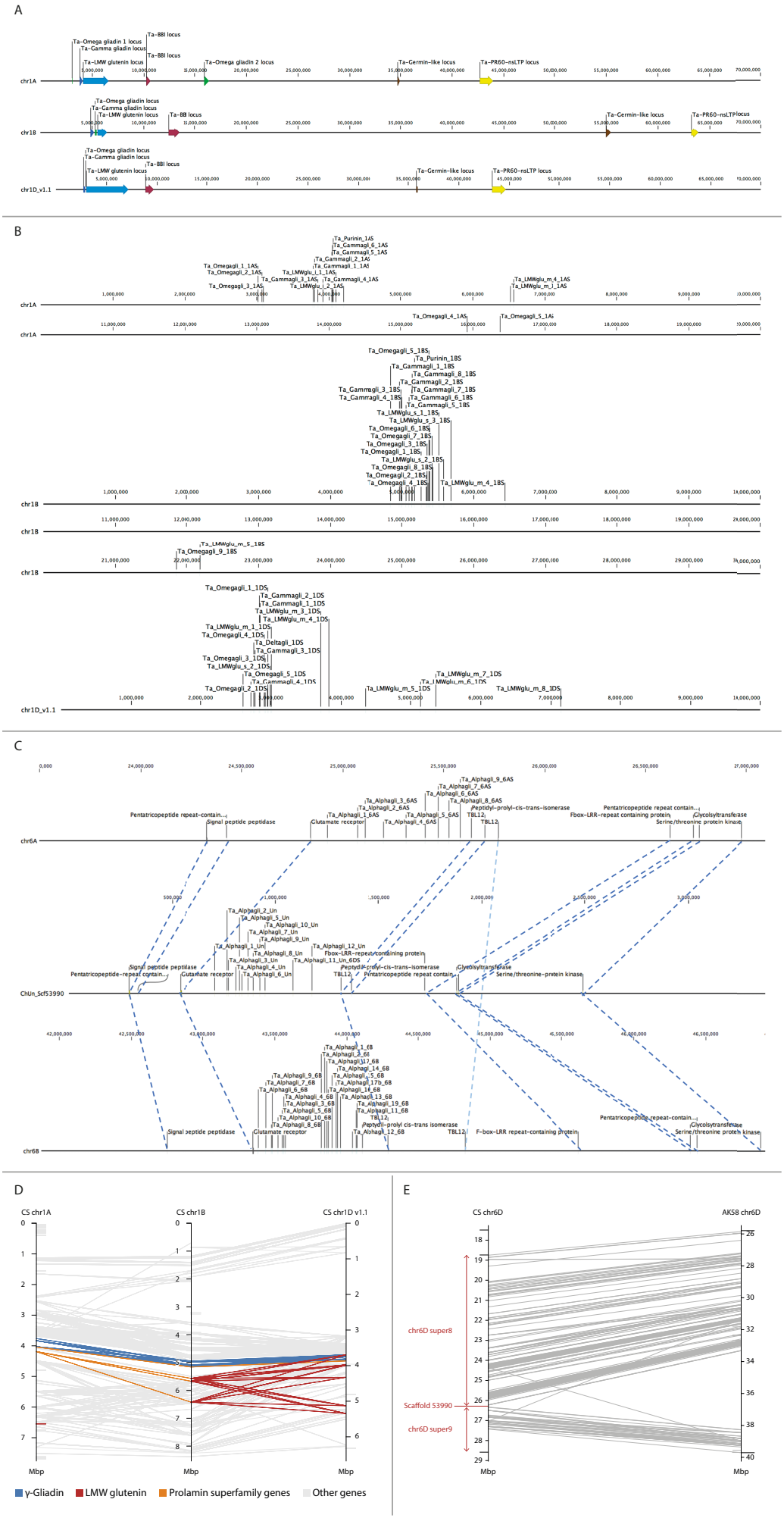

A
chr1A
chr1B
chr1D_v1.1
Ta-Omega gliadin 1 locus
Ta-Gamma gliadin locus
Ta-LMW glutenin locus
Ta-BBI locus
Ta-Omega gliadin 2 locus
Ta-Germin-like locus
Ta-PR60-nsLTP locus
B
chr1A
chr1B
chr1D_v1.1
C
chr6A
ChUn_Scf53990
chr6B
D
CS chr1A
CS chr1B
CS chr1D v1.1
Mbp
γ-Gliadin
LMW glutenin
Prolamin superfamily genes
Other genes
E
CS chr6D
AKS8 chr6D
chr6D super8
Scaffold 53990
chr6D super9
Mbp

Fig. S1. Chromosomal location of major food disease–related protein families on chromosome groups 1 and 6. (**A**) Conserved position of gene loci in chromosomes 1A, 1B and 1D; (**B**) Gliadin and LMW glutenin gene clusters present on short arm of group 1 chromosomes; (**C**) Identification of Chromosome 6D group α-gliadins on ChrUn scaffold 53990. Marker genes mapped within the same region of chromosome 6A and 6B are labelled in scaffold 53990; (**D**) Gene synteny of prolamins located in the short arms of chromosome group 1. Major prolamin types within the locus are labelled. Chromosome 1D v1.1 was produced based on syntheny comparisons to chromosomes 1A and 1B which showed the first two super-scaffolds needed to be re-ordered and one of them flipped, which then restored co-linearity. (**E**) Positional mapping of the gliadin-containing scaffold 53990 on chromosome 6D using the genome assembly of bread wheat cultivar AK58.

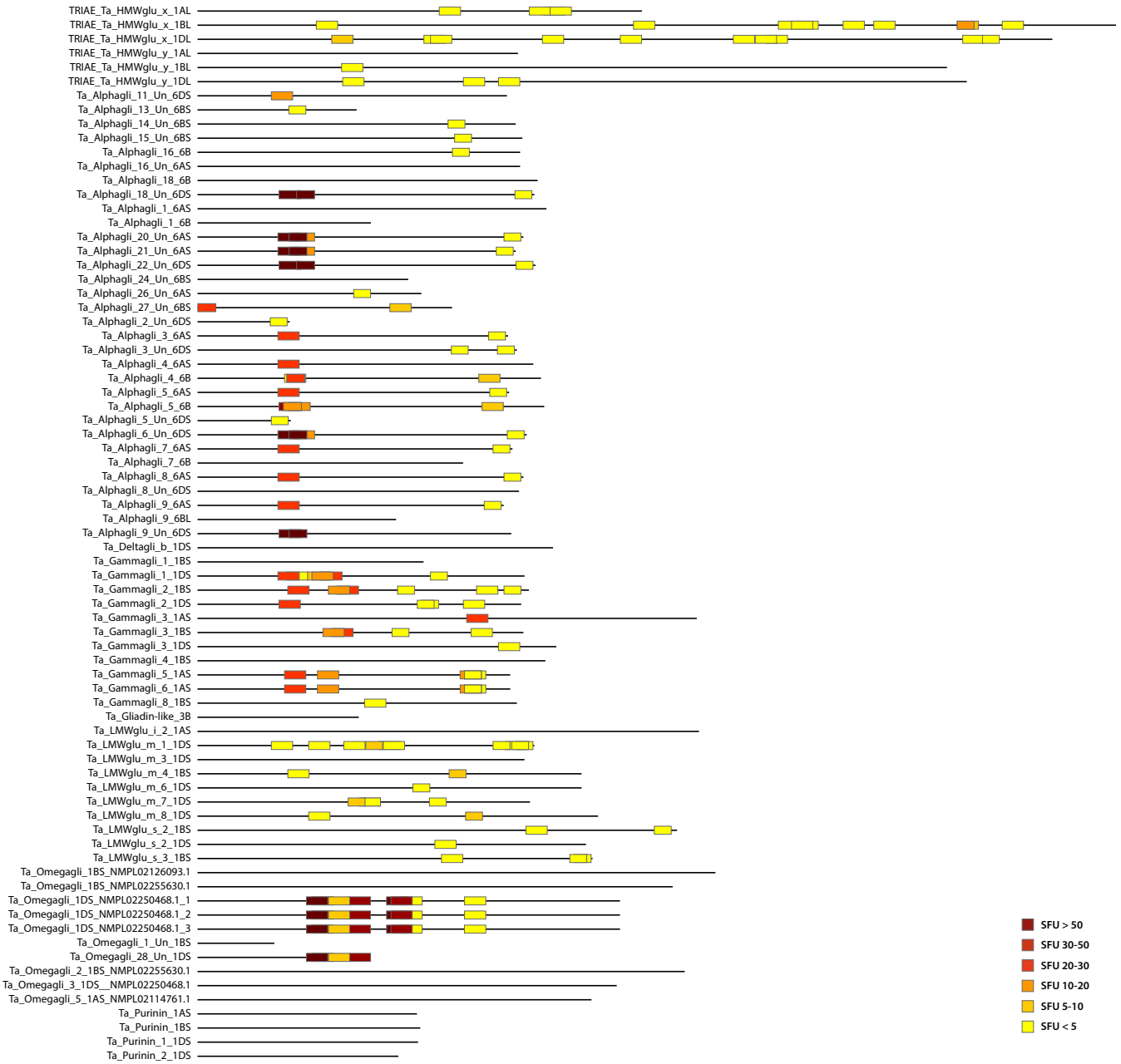


Fig. S2. Mapping of peptides with known IFNγ-ELISPOT response on the gliadin and glutenin sequences of Chinese Spring. The peptides clustered into six groups according their toxicity values and mapped to the protein sequences with 100% sequence identity. Overlapping peptides are represented in piled mode. Pseudo genes were excluded from the analysis.

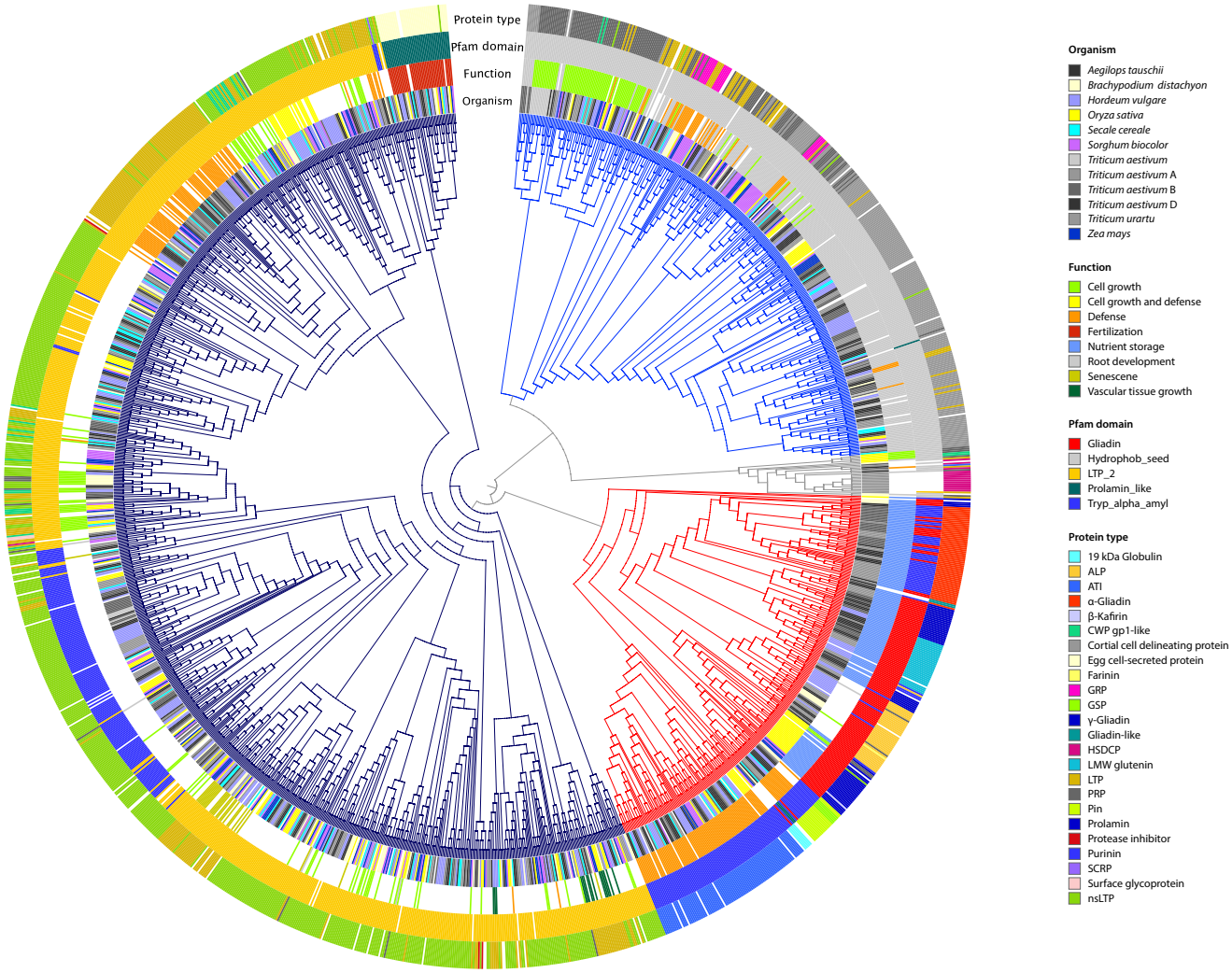
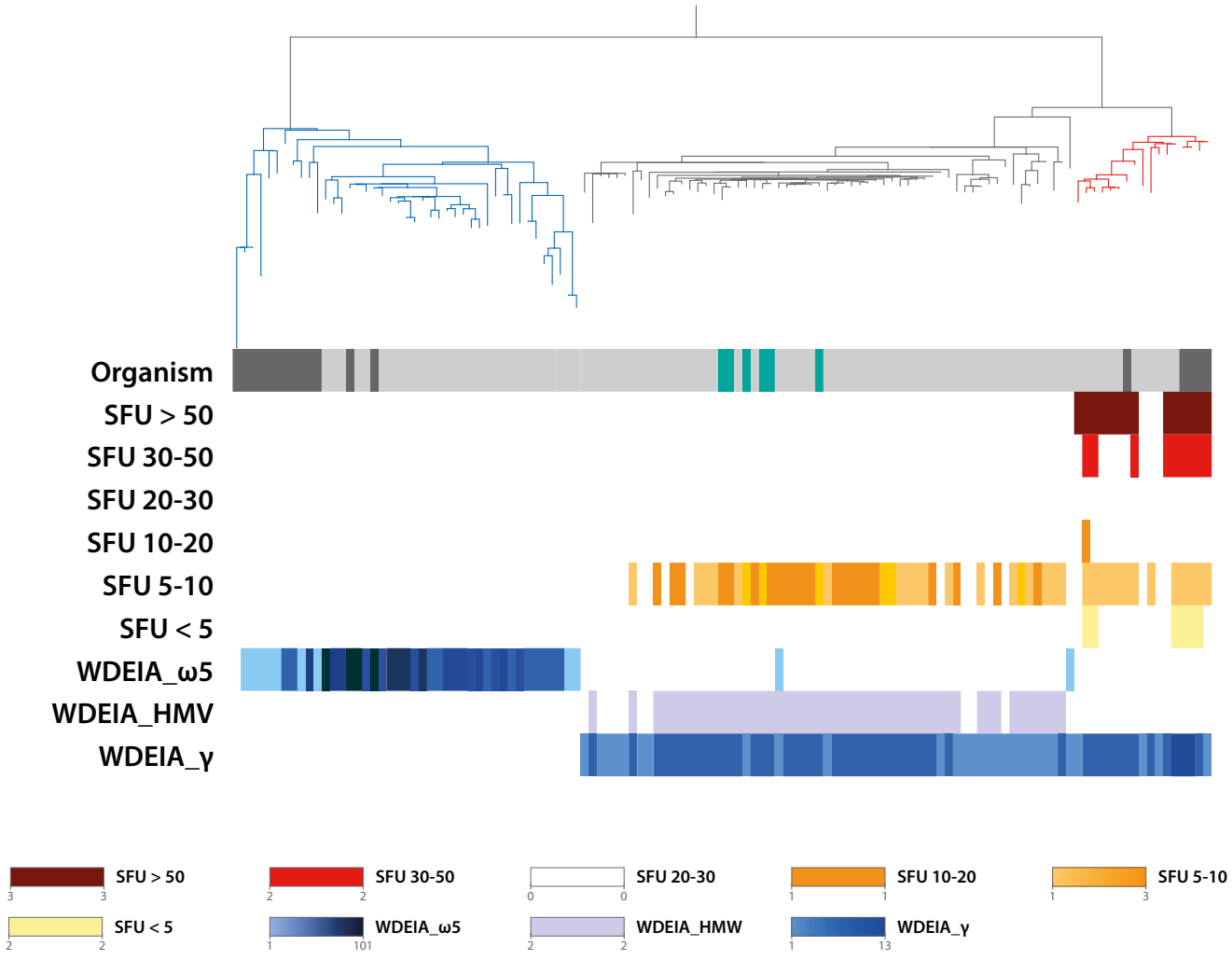


Fig. S3. Phylogenetic analysis of prolamin superfamily gene models. Proteins of Prolamin Clan (CL0482) were identified in wheat and its genome donors, barley, rye, rice, sorghum, maize and Brachypodium. Pfam domain sequence alignment was used to construct the tree. Protein type, pfam domain and organism are labelled. Known molecular functions are highlighted. The sub-tree labelled in light blue represent genes mainly expressed in roots. The sub-tree in red represents seed specific gliadin and tryp-alpha amyl domain containing proteins. The dark blue sub-tree is mainly composed of lipid transfer and non-specific lipid transfer proteins.

A


Organism
SFU > 50
SFU 30-50
SFU 20-30
SFU 10-20
SFU 5-10
SFU < 5
WDEIA_ω5
WDEIA_HMV
WDEIA_γ
SFU > 50
3 3
SFU 30-50
2 2
SFU 20-30
0 0
SFU 10-20
1 1
SFU 5-10
1 3
SFU < 5
2 2
WDEIA_ω5
1 101
WDEIA_HMW
2 2
WDEIA_γ
1 13

B

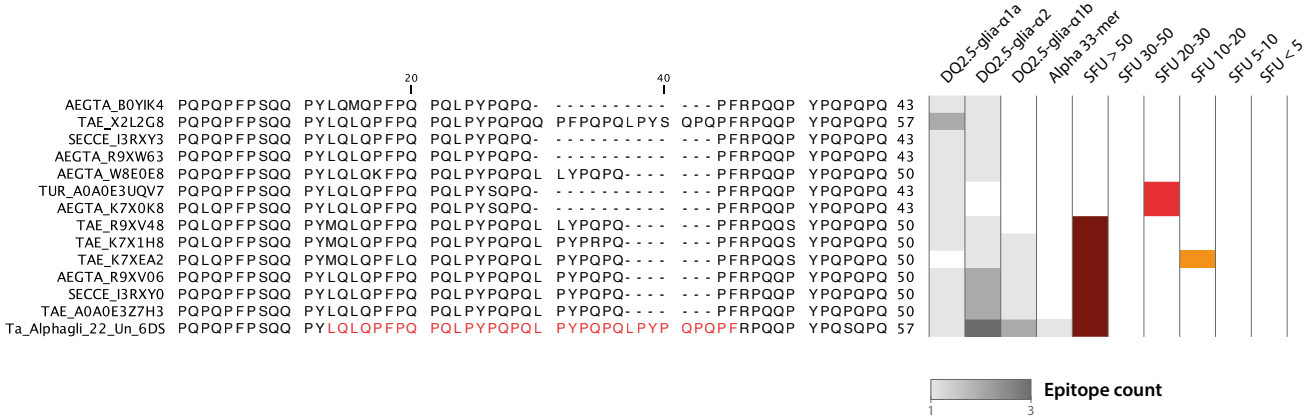

DQ2.5-glia-α1a
DQ2.5-glia-α2
DQ2.5-glia-α1b
Alpha 33-mer
SFU > 50
SFU 30-50
SFU 20-30
SFU 10-20
SFU 5-10
SFU < 5
20
40
AEGTA_B0YIK4 PQPQPFPSQQ PYLQMQPFPQ PQLPYPQPQ- ---------- ---PFRPQQP YPQPQPQ 43
TAE_X2L2G8 PQPQPFPSQQ PYLQLQPFPQ PQLPYPQPQQ PFPQPQLPYS QPQPFRPQQP YPQPQPQ 57
SECCE_I3RXY3 PQPQPFPSQQ PYLQLQPFPQ PQLPYPQPQ- ---------- ---PFRPQQP YPQPQPQ 43
AEGTA_R9XW63 PQPQPFPSQQ PYLQLQPFPQ PQLPYPQPQ- ---------- ---PFRPQQP YPQPQPQ 43
AEGTA_W8E0E8 PQPQPFPSQQ PYLQLQKFPQ PQLPYPQPQL LYPQPQ---- ---PFRPQQP YPQPQPQ 50
TUR_A0A0E3UQV7 PQPQPFPSQQ PYLQLQPFPQ PQLPYSQPQ- ---------- ---PFRPQQP YPQPQPQ 43
AEGTA_K7X0K8 PQLQPFPSQQ PYLQLQPFPQ PQLPYSQPQ- ---------- ---PFRPQQP YPQPQPQ 43
TAE_R9XV48 PQLQPFPSQQ PYMQLQPFPQ PQLPYPQPQL LYPQPQ---- ---PFRPQQS YPQPQPQ 50
TAE_K7X1H8 PQLQPFPSQQ PYMQLQPFPQ PQLPYPQPQL PYPRPQ---- ---PFRPQQS YPQPQPQ 50
TAE_K7XEA2 PQLQPFPSQQ PYMQLQPFLQ PQLPYPQPQL PYPQPQ---- ---PFRPQQS YPQPQPQ 50
AEGTA_R9XV06 PQPQPFPSQQ PYLQLQPFPQ PQLPYPQPQL PYPQPQ---- ---PFRPQQP YPQPQPQ 50
SECCE_I3RXY0 PQPQPFPSQQ PYLQLQPFPQ PQLPYPQPQL PYPQPQ---- ---PFRPQQP YPQPQPQ 50
TAE_A0A0E3Z7H3 PQPQPFPSQQ PYLQLQPFPQ PQLPYPQPQL PYPQPQ---- ---PFRPQQP YPQPQPQ 50
Ta_Alphagli_22_Un_6DS PQPQPFPSQQ PYLQLQPFPQ PQLPYPQPQL PYPQPQLPYP QPQPFRPQQP YPQSQPQ 57
Epitope count
1 3

Fig. S4. Phylogenetic analysis and epitope distribution of highly immunostimulatory gliadin and gliadin proteins in wheat and related species. (**A**) ω-gliadin phylogeny and epitope analysis. Phylogenetic tree was constructed of 92 ω-gliadin sequences of *T. aestivum*, *T. urartu*. *Ae. tauschii* and *S. cereale* origin retrieved from Uniprot. Omega-5 gliadins are highlighted as a blue sub-tree. Highly toxic omega-1,2 sequences are highlighted in red. Coloring of the Organism column represents sequences with D genome origin in dark grey, *S. cereale* sequences in cyan, A genome and B genome originated sequences or sequences without genome information are labelled in light grey. Color intensity in the SFU and WDEIA blocks represents differences in number of epitopes mapped to the individual proteins increasing from lighter to dark color. WDEIA epitopes originally identified from omega-5 gliadins, γ-gliadins and HMW glutenins are labelled. (**B**) Alignment of representative peptides spanning in the alpha 33-mer region. 33-mer peptide is underlined in red. Number of composing epitopes are labelled in grey squares next to alignment. Peptides containing sections with known immune response strength are labelled in the SFU block region.

Table S1. Effect of temperature stress on protein content and composition. Albumin-globulin, Gliadin and Glutenin contents are calculated from the area under the curve values of the SE-HPLC chromatograms and normalized for the protein content. Gliadin composition and amount of alpha 33-mer and toxic ω-gliadin proteins were calculated from the area under the curve values retrieved from the RP-HPLC chromatograms. Toxic ω-gliadin fraction was collected between retention time values 25.26-25.44 min; Alpha 33-mer fraction was quantified from the peak collected between retention time 38-38.8 min)

Cultivar	Treatment	Protein	Albumin-globulin	Glutenin	Gliadin	α-gliadin	γ-gliadin	ω-gliadin	Toxic ω-gliadin fraction	Alpha 33-mer fraction
		g/100g flour								
Berserk	Low	12.63±0.14	0.97±0.02	3.52±0.23	8.20±0.32	4.25±0.23	2.74±0.07	1.20±0.07	0.46±0.04	0.32±0.02
	Normal	16.15±0.12	1.24±0.02	4.09±0.16	10.81±0.16	5.44±0.13	3.88±0.02	1.50±0.03	0.60±0.03	0.45±0.02
	High	19.30±0.30	1.53±0.26	4.76±0.24	12.99±0.40	6.56±0.23	4.67±0.14	1.76±0.14	0.68±0.05	0.57±0.04
Bjarne	Low	11.9±0.13	1.11±0.07	3.82±0.15	7.39±1.05	3.46±0.27	2.66±0.18	1.28±0.11	0.64±0.10	0.36±0.05
	Normal	16.05±0.12	1.07±0.04	4.78±0.09	10.18±0.18	4.52±0.01	3.69±0.11	1.97±0.04	1.09±0.05	0.50±0.01
	High	17.15±0.12	1.35±0.06	4.76±0.12	11.03±0.15	4.82±0.09	4.19±0.12	2.02±0.03	1.13±0.05	0.58±0.02
Chinese Spring	Low	12.03±0.14	0.75±0.03	3.73±0.07	7.54±0.18	3.59±0.12	2.47±0.02	1.47±0.04	0.74±0.04	0.35±0.02
	Normal	16.48±0.19	1.03±0.05	5.39±0.07	10.05±0.20	4.93±0.08	3.26±0.04	1.86±0.08	0.90±0.05	0.50±0.01
	High	19.50±0.88	1.11±0.03	6.43±0.26	11.96±0.62	5.83±0.26	3.75±0.14	2.38±0.20	1.13±0.12	0.62±0.03

Supplementary data:

Data file S1. CDS sequences of prolamin superfamily gene models and identified reference allergens in fasta file format.

Data file S2. Annotation of prolamin superfamily genes and reference allergens used in this study. The annotation table includes gene identification, protein type, pfam clan and domain, genome and chromosomal assignment data for all the genes. Information on food-related disease proteins includes AllFam allergen group IDs, 'Tri a' assignment, caused diseases with references, route of exposure as well as reference homologs in other monocots.

Data file S3. Epitope annotation table of sequences used for the phylogenetic analyses. Sequences include protein sequences with Gliadin and Tryp_alpha_amyl domains of the reference genome and additional protein sequences obtained from the UniProt database. Annotation includes information of organism, Pfam domain, protein type. Epitope data obtained from the ProPepper database were mapped with 100% sequence identity. Table included epitope count data for each sequence. Peptide sequences with known immune-reactivity were retrieved from Tye-Din et al *(15)*. Peptides were grouped into six immunostrength groups as described in the materials and methods. Total peptide counts for each SFU group are highlighted.

Data file S4. Expression of peptides with known IFNγ-ELISPOT responses. Peptides with known immune-reactivity (in median SFU values per million cells) were mapped to the glutenin and gliadin sequences with 100% sequence identity. Expression values of each peptide was calculated as a sum of gene expression levels of genes that contain the peptide sequence in their translated sequence. Peptide expression table contains additional information on the SFU value of the peptide, its' SFU category, the number of sequences per protein type and sub-genome.