GigaScience, 2023, **12**, 1–25
DOI: 10.1093/gigascience/giad084
Research

A community resource to mass explore the wheat grain proteome and its application to the late-maturity alpha-amylase (LMA) problem

Delphine Vincent[1,*], AnhDuyen Bui[1], Vilnis Ezernieks[1], Saleh Shahinfar[1], Timothy Luke[1], Doris Ram[1], Nicholas Rigas[2], Joe Panozzo[2,3], Simone Rochfort[1,4], Hans Daetwyler[1,4] and Matthew Hayden[1,4]

[1]Agriculture Victoria Research, AgriBio, Center Centre for AgriBioscience, Bundoora, VIC 3083, Australia
[2]Agriculture Victoria Research, Grains Innovation Park, Horsham, VIC 3400, Australia
[3]Centre for Agricultural Innovation, University of Melbourne, Parkville, VIC 3010, Australia
[4]School of Applied Systems Biology, La Trobe University, Bundoora, VIC 3083, Australia
*Correspondence address. Agriculture Victoria, AgriBio, 5 Ring road, Bundoora, VIC 3083, Australia. E-mail: delphine.vincent@agriculture.vic.gov.au

Abstract

Background: Late-maturity alpha-amylase (LMA) is a wheat genetic defect causing the synthesis of high isoelectric point alpha-amylase following a temperature shock during mid-grain development or prolonged cold throughout grain development, both leading to starch degradation. While the physiology is well understood, the biochemical mechanisms involved in grain LMA response remain unclear. We have applied high-throughput proteomics to 4,061 wheat flours displaying a range of LMA activities. Using an array of statistical analyses to select LMA-responsive biomarkers, we have mined them using a suite of tools applicable to wheat proteins.

Results: We observed that LMA-affected grains activated their primary metabolisms such as glycolysis and gluconeogenesis; TCA cycle, along with DNA- and RNA- binding mechanisms; and protein translation. This logically transitioned to protein folding activities driven by chaperones and protein disulfide isomerase, as well as protein assembly via dimerisation and complexing. The secondary metabolism was also mobilized with the upregulation of phytohormones and chemical and defence responses. LMA further invoked cellular structures, including ribosomes, microtubules, and chromatin. Finally, and unsurprisingly, LMA expression greatly impacted grain storage proteins, as well as starch and other carbohydrates, with the upregulation of alpha-gliadins and starch metabolism, whereas LMW glutenin, stachyose, sucrose, UDP-galactose, and UDP-glucose were downregulated.

Conclusions: To our knowledge, this is not only the first proteomics study tackling the wheat LMA issue but also the largest plant-based proteomics study published to date. Logistics, technicalities, requirements, and bottlenecks of such an ambitious large-scale high-throughput proteomics experiment along with the challenges associated with big data analyses are discussed.

Keywords: *Triticum aestivum*, large-scale high-throughput workflow, bottom-up shotgun proteomics, LC-MS/MS, late-maturity alpha-amylase, LMA, big data, statistics, data mining, circos plot

Key points

- Largest plant proteomics dataset
- First LMA proteomics study
- Molecular toolkit to assist wheat breeders to select for or against quantitative traits such as LMA

Introduction

Common bread wheat (*Triticum aestivum L.*) is the dominant crop in temperate regions, currently covering more than 220 million hectares worldwide, exceeding 749 million tons in production annually [1] and predicted to reach 835 million tons by 2030 [2]. Millennia of domestication have accrued an enormous genetic diversity in this species, with potentially more than 50,000 *T. aestivum* cultivars [3]. Wheat owes its success to adaptability to temperate, Mediterranean, and subtropical climates; high yields; storability; but above all to the unique properties of doughs, which can be processed into a vast range of foods [4, 5]. Wheat grains are not only a major source of carbohydrate in the form of starch, which can reach levels of up to 75% in white flour, but also a substantial source of protein, representing up to 15% of grain dry weight [5]. Wheat proteins can, however, trigger adverse reactions such as dietary intolerance or food and respiratory allergies [5]. Current breeding programs mainly aim at sustaining wheat production and quality with reduced agrochemical inputs, as well as developing new disease-resistant and stress-tolerant varieties with enhanced quality for specific end uses [6]. Wheat research and breeding must accelerate genetic gain to keep augmenting crop yield while maintaining or improving grain quality traits if the demands of the growing human population are to be met [7].

A critical element in the equation was the sequencing and functional annotation of the genome. Sequencing the hexaploid bread wheat genome was a gigantic achievement proportionate to its large size, abundance of repetitive DNA, and the immense difficulty of discerning homoeologs from subgenomes A, B, and D. Whilst this required the commitment of 20 countries

Received: April 24, 2023. **Revised:** August 2, 2023. **Accepted:** September 19, 2023

collaborating as a consortium (International Wheat Genome Sequencing Consortium [IWGSC]) and a lot of strategizing from 2005 onward, including sequencing diploid and tetraploid ancestors, it was the advent of next-generation sequencing technologies producing long but error-prone or accurate yet short reads that made this massive endeavour successful [8]. A 13-year effort ensued, drafting in 2014 the *T. aestivum* genome [9] and culminating in 2018 with the release of the long-awaited fully assembled and annotated 14.5-Gb reference genome, cataloguing 107,891 high-confidence genes along 21 chromosome-like sequence assemblies (IWGSC RefSeq v1.0) [7]. A refined version of the reference genome using optical mapping and long sequence reads was recently released (IWGSC RefSeq v2.1) [10]. With such worthwhile genomic resources in store, wheat can now be instated as a model for plant genetic research and employed to tackle complex biological questions on evolution, domestication, polyploidization, and genetic and epigenetic interaction between homoeologous genes and genomes [8]. Genome annotations pave the way to investigate pathways and biochemical attributes behind bread wheat quality using transcriptomics [11] or proteomics [2] approaches.

The industry will equally benefit from these latest scientific developments since processing companies, markets, and food industries demand not only high-yielding and resistant varieties but also those with specific end-use qualities [1, 4]. Market requirements have influenced wheat breeding as not to neglect essential protein content and quality. Because wheat is generally traded according to grain protein content and hardness, standards must be abided to by producers and distributors. Intact starch polymers provide the gelatinization and retrogradation needed for an acceptable product. Failure to meet receival standards for milling grades due to starch degradation measured in the wheat industry using the Hagberg–Perten falling number (FN) method [12] leads to grain discount and downgrading to animal feed, which incurs a loss of profit [13]. The low FN values manifest as a loss of viscosity upon mixing starch-degraded flour with water can alter appearance and texture of end products [14], but it might not deteriorate baking functionality [15] and could be used instead in alternate preparations [16]. There are multiple causes of low FN symptomatic of starch degradation, including preharvest sprouting, late-maturity alpha-amylase (LMA), and variation in kernel starch and protein [17]. LMA is a wheat genetic defect causing the synthesis of high isoelectric point (pI) alpha-amylase in the aleurone because of a temperature shock during mid-grain development or prolonged cold throughout grain development, leading to an unacceptable low FN at harvest or during storage [18–20]. High pI alpha-amylase is normally not synthesized until after maturity in seeds when they may sprout in response to rain or germinate following sowing the next season's crop [21].

Four alpha-amylase isoforms have been identified to date in wheat. Several alpha-amylase 1 (TaAMY1) loci have been localized on the long arm of group 6 chromosomes [22]. In LMA-prone wheat genotypes and under given temperatures, Amy-1 genes are transcribed in isolated cells or cell islands distributed throughout the aleurone system of grains with a 50–60% moisture content before they have reached physiological maturity [21]. Appearance of high pI alpha-amylase protein is preceded by a short-lived transient period of mRNA synthesis leading to a stable enzyme and retained through to seed maturity [18, 23]. Multiple alpha-amylase 2 (TaAMY2) loci are positioned on the long arm of the group 7 chromosomes and produce a low pI alpha-amylase in the pericarp of the developing grain [24]. A single locus encodes alpha-amylase 3 (TaAMY3) on group 5 chromosomes and is transcribed throughout the grain development, suggesting a role in grain development and maturation [25]. Like TaAMY2, TaAMY3 enzyme mainly appears during grain development in the pericarp and would be the predominant alpha-amylase enzyme throughout grain development [26]. Despite its shorter length and elevated pI, TaAMY3 displays equal numbers of calcium-binding and active sites relative to the other 3 isoforms; however, the distance between key AA residues and the last 2 active site residues is shortened [27]. Overexpressing TaAMY3 in the endosperm of developing grain to levels of up to 100-fold higher than the wild-type results in low FN similar to those seen in LMA-affected grains, yet has no detrimental effect on starch structure, flour composition, and baking quality of bread [28] or on noodle colour or firmness [29]. A fourth isoform, alpha-amylase 4 (TaAMY4), is also encoded by a single locus on group 5 chromosomes and is coexpressed with TaAMY1 in LMA-affected grains [27]. Comparison of the 4 isoforms revealed that they contain 385 to 439 AAs, with a molecular mass between 45.4 and 48.3 kD, and a pI ranging from 5.5 to 8.6. All isoforms differ slightly in their 3-dimensional (3D) protein structure, including the presence of additional sugar binding domains hinting to various enzymatic properties [27, 30].

Although LMA expression correlates with measurable changes in both hormone content and transcript profiles during grain maturation, there is no obvious visual effect on grain appearance, development, or morphology [20], hence the need to perform assays to test for its activity [12]. Enzyme-linked immunosorbent assay (ELISA) [31] and quantitative reverse transcription polymerase chain reaction (RT-qPCR) [32] assays were developed to specifically target TaAMY1, the main enzyme involved in LMA. One limitation to the RT-qPCR method relates to the apparent short life of the high pI alpha-amylase mRNA [18]. Commonly employed is the colorimetric Ceralpha assay [33], whereby the alpha-amylase activity is expressed in terms of Ceralpha units per gram of flour (U/g). A single unit corresponds to the amount of enzyme required to release 1 μM p-nitrophenyl in the presence of excess quantities of alpha-glucosidase in 1 minute at 40°C [34]. Such measurements have revealed that LMA is more prevalent than originally thought, with reports arising from North America, Australia, Japan, Canada, South Africa, China, Mexico, Germany, and the United Kingdom [35]. The presence of LMA in breeding populations could be attributed to unexplained positive effects on grain production/quality or alternately simply manifest the lack of significant selection pressure against this trait [20]. Both a cool temperature shock near physiological maturity or continuous cool maximum temperatures during grain development can induce LMA synthesis in wheat [19]. The prediction of LMA occurrence during LMA dedicated field trial is impeded by the stochastic nature of LMA expression resulting from specific genetics, climatic conditions, and developmental stages.

LMA has a genetic (G) component (alpha-amylase gene required), yet it is only expressed and enzymatically active under particular environmental (E) conditions (temperature shock) at a given developmental stage, making it the product of a G × E interaction, which lends itself to postgenomic quantitative studies to shed some lights into the biological mechanisms underpinning LMA expression. Yet, to date, only 1 LMA-related transcriptomics study has been published and no proteomics work has been attempted despite the potential this technology offers to help improve bread wheat quality [2]. Using microarray technology, Barrero and colleagues [18] reported that LMA resulted from a very narrow and transitory peak of expression of genes encoding high-pI alpha-amylase during grain development. Furthermore, the LMA phenotype triggered elevated levels of gibberellins such as GA19 and much lower levels of auxin in the de-embryonated fraction of grains sampled shortly after the initiation of LMA synthesis. A recent report questions this hormonal response since alpha-amylase synthesis by wheat aleurone during grain development appears to be independent of gibberellin [36]. Even though, on one hand, genomics can catalogue genes present in a sample and, on the other hand, transcriptomics can validate expression levels, only proteomics can measure the actual protein abundance, record posttranslational modification (PTM), and identify interacting proteins [2]. We have developed a high-throughput proteomics method to rapidly profile *T. aestivum* grains and datamine their proteome [37]. In the present study, we have applied our optimised procedure to a collection of 4,061 wheat flours whose LMA content ranged from 0 to 8 U/g of flour. We have applied multiple statistical analyses to our big data to select LMA-responsive biomarkers that we have mined using a suite of tools applicable to wheat proteins, yet not necessarily embraced by grain scientists. To our knowledge, this is not only the first proteomics study tackling the wheat LMA issue but also the largest plant-based proteomics study published to date. Logistics, technicalities, requirements, and bottlenecks of such an ambitious large-scale high-throughput proteomics experiment along with the challenges associated with big data analyses are discussed.

Results and Discussion

Resources for scientific studies on wheat

Wheat resources

A total 858 wheat genotypes, sourced from all over the world, grown over 8 years since 2012 and stored in optimal conditions amounting to 4,061 grain samples, were analysed in this work (Supplementary Table S1). Because LMA measurements occurred simultaneously to the proteomics analyses in 2019, we did not consider storage time for the statistics. We also did not statistically test for varietal differences, which was outside the focus of this study.

High-throughput proteomics workflow to efficiently process and analyse thousands of samples

We have developed a high-throughput proteomics liquid chromatography–mass spectrometry (LC-MS) method [37] that was applied to 4,061 wheat grain samples following the workflow described in Fig. 1. The technical aspects pertaining to sample preparation/tracking and data acquisition steps that ensured a high-throughput workflow are available in Supplementary File SF1. Overall, the LC-MS continuous run lasted for 143 days (20.4 weeks or 4.5 months) and included regular system maintenance (mass calibration, source cleaning, high-performance liquid chromatography [HPLC] column swapping). A total of 4,370 RAW files were acquired. A Gantt chart illustrates the timeline of the workflow steps along with data accumulation (Fig. 2).

The wet experiment bottlenecks were resolved where possible as explained in [37]. Most time was spent grinding, transferring, weighing, and extracting the samples as there was no option to greatly upscale those steps (Fig. 2). The workflow became much faster when 96-well plates were introduced (from digestion step onward), allowing for high-throughput multipipetting and multidispensing activities, as well as minimising the footprint of sample freezer storage. Although steps were sequential, they could overlap with 2 experimenters operating in a staggered fashion from one lab workstation to the next.

LC-MS1 acquisition started when enough plates were ready to ensure continuous instrument run while samples processing was still happening. Data acquisition was completed 18 days after the last wheat sample was fully processed, demonstrating minimum time loss (Fig. 2). The Genedata Refiner workflow used to process LC-MS1 data was previously optimised [37] (Supplementary Fig. S1); its first step was applied to batches of ∼200 LC-MS1 files during MS run. The time-limiting factor was the server computing ability.

Overall, all 4,061 wheat samples were processed and analysed (from receiving the samples to processing the LC-MS1 data) in 334 days (∼11 months). Purchasing all required consumables ahead, keeping track of the samples, well-organised logistics by setting up working stations for each wet lab step, and overlapping activities across experimenters guaranteed efficient time management. Stowing samples in the freezer in between steps allowed to safely interrupt the sample preparation procedure to accommodate equipment/experimenter downtime without compromising the quality of the samples processed so far.

The subsequent steps had to follow one another. LC-MS2 acquisition necessitated LC-MS1 data processing to be finished to produce parent mass lists and consequently had to be performed post hoc. Whilst LC-MS2 acquisition was rapid (2 weeks), its processing took longer (3 months) because it required another Genedata Refiner workflow (Supplementary Fig. S2), a more recent nonredundant database with decoy sequences, testing several Mascot parameters (data not shown), and linking LC-MS2 clusters to LC-MS1 clusters (data not shown).

The final bottleneck in the workflow pertained to statistical analyses and data mining (8 months), which necessitated trying different statistical methods with multiple trial and error stages working out optimal parameters, testing and using different data mining tools, which required training and a lot of strategizing on how best to present big data. Running such large datasets proved computationally taxing, necessitating extensive dwell times; it often ran out of memory and triggered server crashes.

One way to increase the throughput and therefore shrink the timeline would be to use an automated sample preparation station. A robot (Bravo Automated Liquid Handling Platform from Agilent) was used to automate peptide clean-up and phosphopeptide enrichment from wheat and maize vegetative samples [38]. We could not find any other high-throughput method in wheat or cereals.

LC-MS1 quantitative data processing, normalisation, correction, and standardisation to remove technical biases

The Genedata Refiner workflow was applied to 4,147 LC-MS1 files (4,061 wheat + 86 quality controls [QCs]; Supplementary Fig. S1). Step 1 covered noise subtraction nodes that could be run on individual data file. It was performed throughout LC-MS1 acquisition activity on weekly batches (∼230 files) to optimise server dwell time. Step 1 helped assess data reproducibility and nonreproducible files (71 samples) were omitted from the remainder of the processing, leaving 3,990 wheat and 86 QC data files. Step 2 encapsulated all alignment, peak detection, and quantitation, as well as isotope clustering and singleton filtering activities. This step had to be performed on all 4,076 reproducible data files simultaneously and therefore could only be undertaken when the LC-MS1 run was finalised. The experiment metadata captured in Excel were associated to the quantitative data and exported to Genedata Analyst for data normalisation purpose.

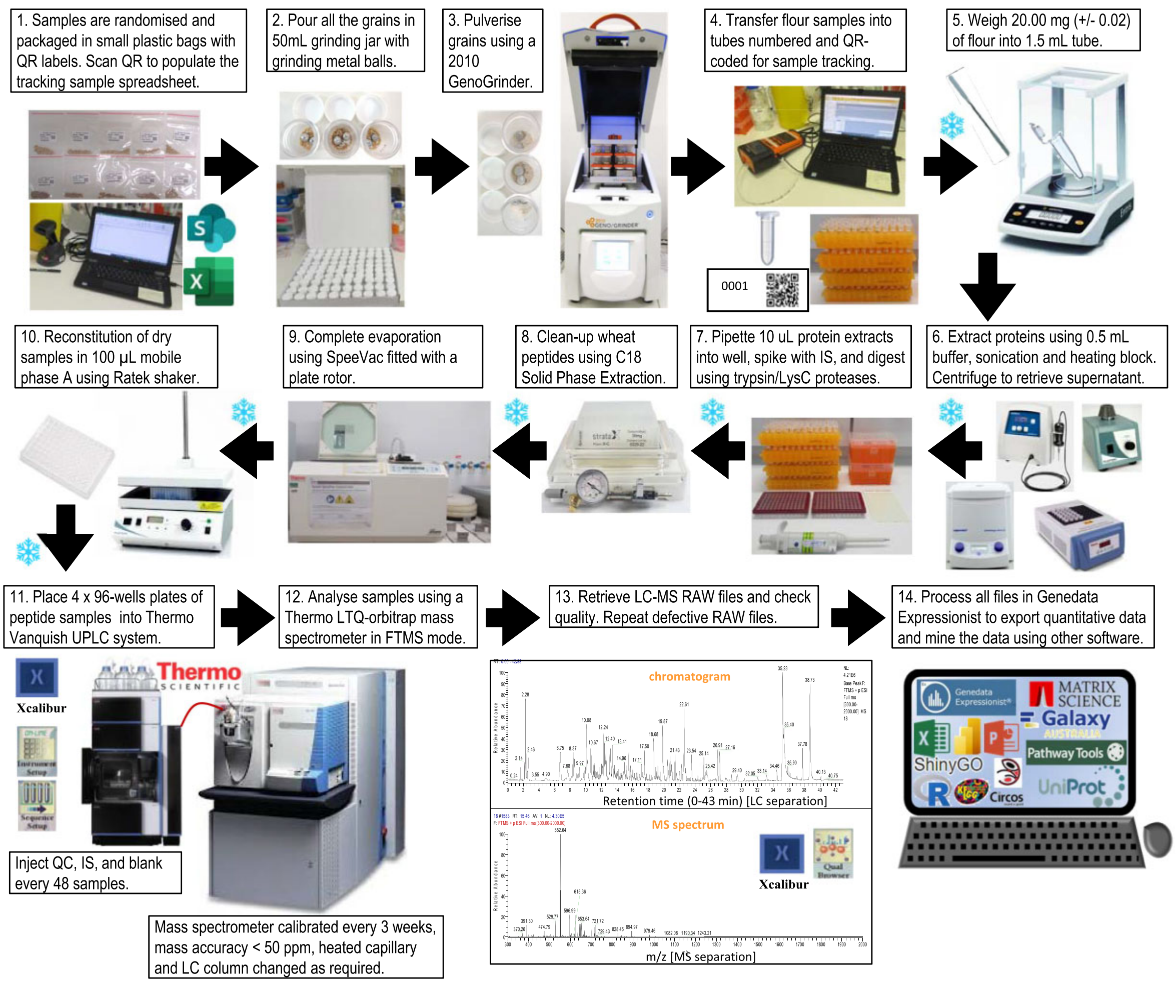

1. Samples are randomised and packaged in small plastic bags with QR labels. Scan QR to populate the tracking sample spreadsheet.
2. Pour all the grains in 50mL grinding jar with grinding metal balls.
3. Pulverise grains using a 2010 GenoGrinder.
4. Transfer flour samples into tubes numbered and QR-coded for sample tracking.
0001
5. Weigh 20.00 mg (+/- 0.02) of flour into 1.5 mL tube.
6. Extract proteins using 0.5 mL buffer, sonication and heating block. Centrifuge to retrieve supernatant.
7. Pipette 10 uL protein extracts into well, spike with IS, and digest using trypsin/LysC proteases.
8. Clean-up wheat peptides using C18 Solid Phase Extraction.
9. Complete evaporation using SpeeVac fitted with a plate rotor.
10. Reconstitution of dry samples in 100 µL mobile phase A using Ratek shaker.
11. Place 4 x 96-wells plates of peptide samples into Thermo Vanquish UPLC system.
12. Analyse samples using a Thermo LTQ-orbitrap mass spectrometer in FTMS mode.
13. Retrieve LC-MS RAW files and check quality. Repeat defective RAW files.
14. Process all files in Genedata Expressionist to export quantitative data and mine the data using other software.
Xcalibur
Thermo SCIENTIFIC
Inject QC, IS, and blank every 48 samples.
Mass spectrometer calibrated every 3 weeks, mass accuracy < 50 ppm, heated capillary and LC column changed as required.
chromatogram
Relative Abundance
Retention time (0-43 min) [LC separation]
MS spectrum
Xcalibur
Qual Browser
m/z [MS separation]
Genedata Expressionist
MATRIX SCIENCE
Galaxy
Pathway Tools
ShinyGO
R
Circos
UniProt

Figure 1: High-throughput workflow used on the 4,061 wheat samples. The snowflakes indicate storage in −80°C freezers.

The data were normalised as described in [37] following 3 steps: using flour weights, internal standard (IS) cluster, and QC replicates along with LC-MS injection order (Fig. 3).

Raw data displayed a clear sample grouping based on injection order during the LC-MS1 run (Fig. 3A) and mirrored the instrument maintenance events (mass calibration, etc.). Two large groups appeared that could not be explained by any experimental steps. Normalising using flour weight accuracy of 1% helped create tighter wheat sample groups with 4 outliers and isolated QCs (Fig. 3B). The 2 larger groups of samples were less distinct. This first normalisation step did not significantly impact the peptide distribution, as can be seen on the principal componcent analysis (PCA) loading plots (Supplementary Fig. S3G, H). Normalising against the IS shifted the sample groups around but did not combine or homogenise them (Fig. 3C). The two larger sample groups observed in panels A and B became indistinguishable in panel C. This normalisation step also affected peptide distribution assuming a more oval shape on the loading plot (Supplementary Fig. S3I). The final normalisation step further scattered the samples more widely across the PCA plot and accentuated the technical variation gradually expanding over time during the instrument run (Fig. 3D). Yet at the peptide level, this last normalisation activity further shrunk the grouping, assuming a more circular distribution with less outliers (Supplementary Fig. S3J). The benefits of normalisation were discussed before [37] with respect to precise sample weights mandated by metabolomics [39], spiking IS postdigestion to alleviate for sample-to-sample variations [40, 41], and QCs to account for batch differences over time and minimise cross-run effects [41–43]. In their ground-breaking study to assess and ameliorate the reproducibility of large-scale proteomics experiments, Poulos and colleagues [44] have highlighted the decrease over time in mass analyser sensitivity in between cleaning events and how technical replicates, such as QCs, help remove unwanted variation. Despite all the normalisation steps applied to our data, not all technical biases could be removed, thus necessitating further data correction.

The fully normalised dataset of 3,990 wheat samples and 32,336 reproducible peptides was exported as a CSV file and imported into R to run a linear model fitting the technical factors that bore the greatest variance and were associated with LC-MS maintenance. The experimental variation was successfully eradicated as illustrated by PCA (Fig. 3E, K). The results showed that while instrument mass calibration had a much bigger effect, all 3 technical factors had a significant effect ($P < 0.05$ based on permutation testing with 100 iterations) on the spectral data (data not shown). This correction method was initially developed in a metabolomics study to account for uncontrollable environmental effects [45]. Quantitative geneticists routinely exploit linear

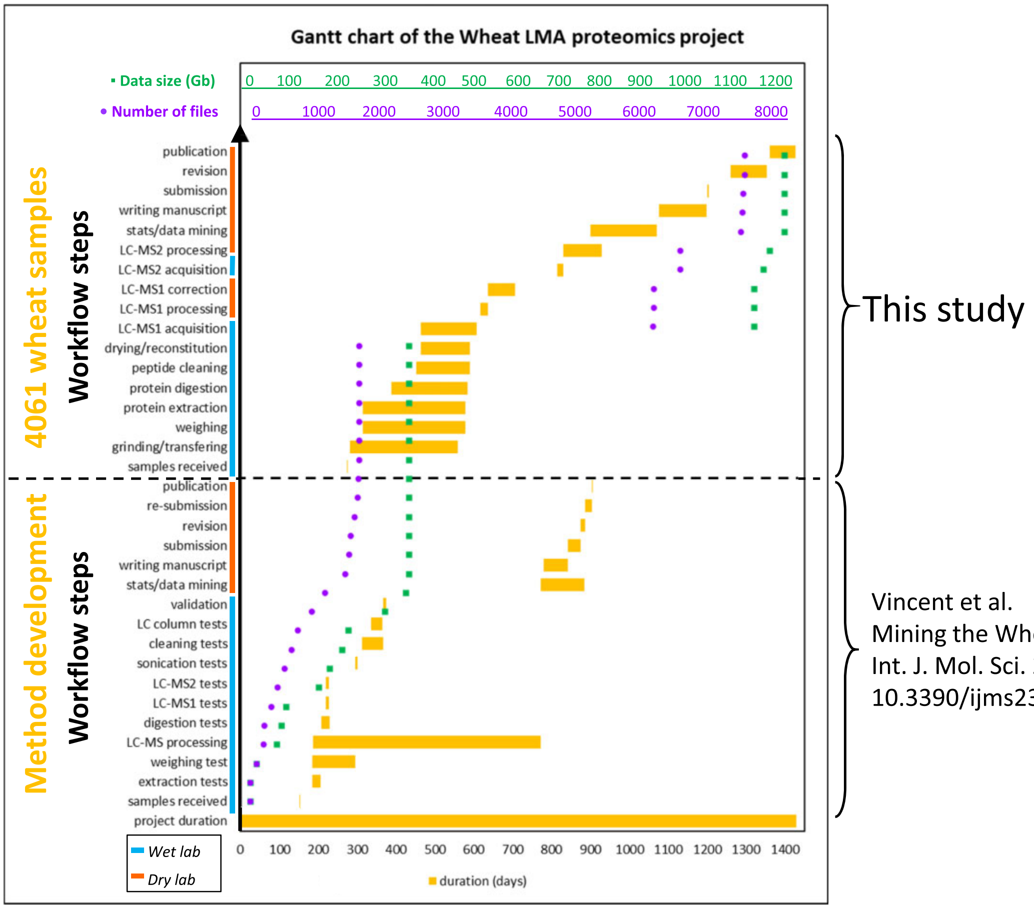


Figure 2: Gantt chart capturing the timeline for each step of the proteomics workflow and data accumulation during both the method development and large-scale analysis of the 4,061 wheat samples.

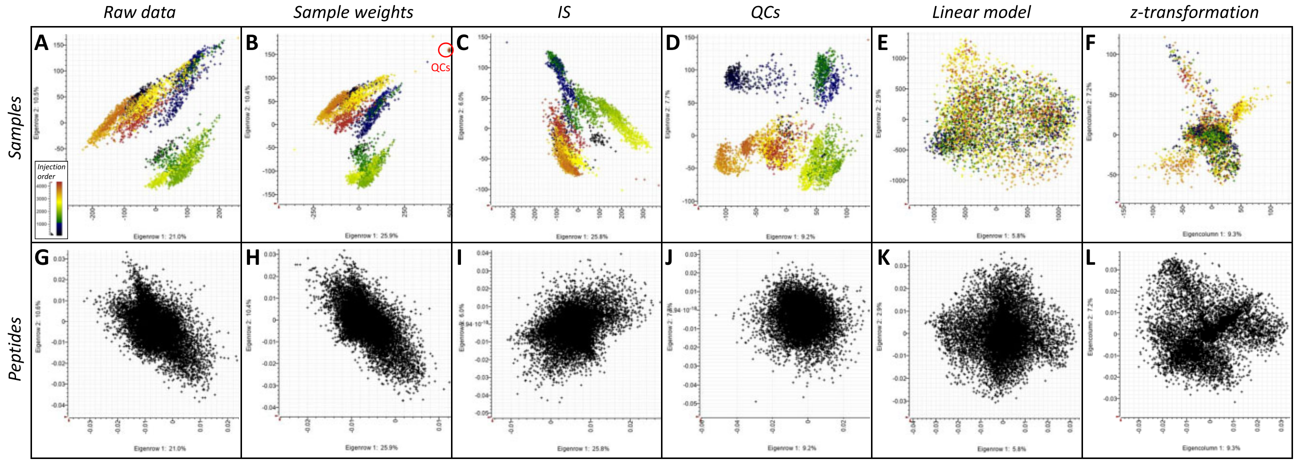


Figure 3: Normalisation, correction, and standardisation of the raw data visualised using PCA projection plots of the samples (A–F) and loading plots of the peptides (F–K). Samples are coloured accordingly to LC-MS injection order from blue-green to yellow-orange-red. (A, G) PC1 vs. PC2 plot based on unnormalised LC-MS1 quantitative data. (B, H) PC1 vs. PC2 plot based on data from panels A, G normalised using the sample weights; QCs are all condensed in a tight group. (C, I) PC1 vs. PC2 plot based on data from panels B, H normalised using the IS cluster. (D, J) PC1 vs. PC2 plot based using data from panels C, I normalised using the injection order and the "intensity drift" algorithm. (E, K) PC1 vs. PC2 plot using normalised data from panels D, J corrected using a linear model and keeping the residuals. (F, L) PC1 vs. PC2 plot using corrected data from panels E, L and z-transformed per row (peptides).

models to measure the influence of systematic environmental effects (fixed effects), which impact phenotypic variation and unscramble genetic from nongenetic factors [46]. To our knowledge, this is the first time such correction method was applied to proteomics data.

The final data transformation step involved a z-transformation (scaling and centring) to level out extreme quantities and facilitate the comparison and clustering of peptide profiles during statistical analyses. Finding linear combinations of predictors based on how much variation they explain is achieved by centring to a mean of 0 and scaling to a standard deviation of 1 [47]. Such mathematical transformation is common practice in postgenomics expression studies, and MS-proteomics is no exception [48, 49]. In our study, z-transformation radically modified the data from an homogeneous plot to defined groups stretching in 4 main directions (Fig. 3F, L), which could not be attributed to any of our metadata. Peptide quantities that originally ranged from 0 to 1 × 10^7 ultimately spanned a mere −22 to 63 scale.

A nonredundant wheat database to annotate LC-MS2 results

A *T. aestivum* database was created by combining all the protein sequences publicly available from UniProt and IWGSC EnsemblPlants repositories. The database was reversed to create a decoy database, which was then concatenated to the latter. This way, not only a single file has to be interrogated in the Mascot system, but also false positives are only recorded when a match from the decoy sequences exceeds any match from the target sequences [50, 51]. All LC-MS2 files were searched using the Mascot algorithm with an error-tolerant search to maximise PTM discovery.

Our strategy to quickly identify as many peptides as possible was to multiply the number of data-dependent LC-MS2 methods rather than multiplying the number of samples analysed. We thus pooled 10% of the wheat samples randomly chosen into 1 tube and subjected this pooled sample to 11 methods (passes) with replicates, varied ITMS parameters, and 10 unique parent lists of 2,000 ions each (Supplementary Files SF1, SF2). Each method had a drastic impact on the selection of the precursor ion, with some areas being thoroughly sampled whilst others were ignored (Supplementary Fig. S3).

A total of 63 LC-MS2 files were thus obtained. The LC-MS2 methods fluctuated in their efficiencies, identifying as few as 104 peptides (pass 7) up to 11,662 peptides (pass 8), irrespective of the number of MS2 events (Supplementary Fig. S4).

Passes 8–10 yielded by far the largest identity counts across all 10 parent lists, even though they did not feature the highest MS2 event counts (Supplementary Fig. S4). We concluded that key MS parameters to maximise peptide identifications were the inclusion of the parent lists into the data-dependent settings (passes 8–11), albeit not the at the global level (pass 7), as well as allowing for wider mass tolerance window during precursor selection. The widest tolerance (2 *m/z*) achieved the greatest counts (pass 8, Supplementary Fig. S4). Overall, a total of 315,934 peptides were identified, comprising only 6,550 unique peptides, which matched 10,437 unique wheat proteins, 277 decoy accessions, and 3 contaminant proteins. The huge peptide redundancy was explained by the fact that a single pooled sample (from 400 individual samples) was repeatedly analysed using various LC-MS2 methods. Pooling digests erased sample-to-sample variation. More protein identities could have been realised with a diverse sample set subject to all the methods developed here, but that would have extended the data acquisition, analysis, and mining by many more months. An array of strategies can be employed to increase the proteome coverage of plant seeds, including depletion and prefractionation strategies as well as exploring different organs, developmental stages, and cell cultures [52, 53]. However, these additional experimental steps are time-consuming, labour-intensive, and costly and thus unsuitable for large-scale high-throughput experiments like ours. Our strategy was first to quantify peptides rapidly and reproducibly from thousands of wheat samples using a label-free LC-MS approach and apply robust statistical analyses to detect potential trait-related biomarkers and, second, to quickly identify as many peptides as possible using LC-MS2. Large-scale proteomics studies have been applied to humans [54]; to our knowledge, this is the largest plant proteomics study carried out to date.

Posttranslational modifications (PTMs)

In this study, we opted for an error-tolerant search, which accrued a plethora of modifications (Supplementary Table S2). A total of 21,486 carbamidomethylations of Cys residues were identified as fixed modifications. This was expected to occur during our denaturing protein extraction procedure. The most prevalent dynamic modifications were nonspecific cleavages (5,480), followed by N-terminal ammonia losses (907) and conversion from N-terminal Gln to pyroGlu (815). During the digestion process involving trypsin, proteomics studies have often reported the formation of semi-tryptic and nonspecific peptides besides cleavages after Arg or Lys residues [55]. Therefore, some of our nonspecific peptides could have resulted from the digestion step, but we cannot rule out that nontryptic peptides were naturally present on our stored grains, resulting from residual enzymatic activities.

Ammonia losses are neutral losses commonly triggered by CID upon creating b and y ions and can be detected by high-resolution mass analysers such as FTMS instruments [56]. C-terminal Arg or Lys of tryptic peptides often leads to abundant y ions with ammonia loss [57] as well as b ions specific enough to detect the presence of Gln, Asn, His, Lys, and Arg residues [56]. PyroGlu formation is a common cyclization side reaction of Glu and/or Gln residues in peptides and proteins that occurs when those residues are located at the N-terminus and under slightly acidic conditions [58], such as our experimental conditions; therefore, this PTM could also be a process artefact. Other frequent PTMs in our study were N-terminal ethylation (265 occurrences); deamidation (147 occurrences); guanidylation (141 occurrences), the latter of which could have been triggered during protein resuspension in Guanidine-HCl solution as discussed in [37]; and oxidation of Met (100 occurrences) (Supplementary Table S2).

Numerous PTMs have been identified in plants [52] and cereals in particular [59], including barley [60] and wheat [2, 61, 62]. Deamidations of glutamine residues in glutenins have been reported [5], along with C-terminal loss of tyrosine, potentially facilitating protein sorting during seed maturation [2]. Starch content and storage proteins are prominent in wheat grain; PTMs involved in starch quality have been reviewed [63]. Our study lists numerous potential PTMs; this warrants more experiments to validate them and decipher their role in LMA response. Future proteomics experiments should endeavour to explore the relationship between structure and functionality of gluten proteoforms arising from key PTMs in response to the LMA phenotype.

Linking LC-MS1 and LC-MS2 data to annotate quantities with identities

LC-MS1 files resolved 32,336 reproducible clusters, which had to be matched to 29,908 clusters from LC-MS2 data files. Using tol-

Table 1: Experiment summary.

Items quantified	Occurrences
Number of wheat genotypes	858
Number of wheat samples	4,061
Sampling years	8 (2012–2019)
Trait (LMA)	1
Digestion types	1
Number of reproducible LC-MS1 files	3,990
Number of LC-MS1 peaks	137,669
Number of reproducible LC-MS1 clusters	32,336
Cluster size range	2–10
Cluster charge range	2–7
Cluster *m/z* range	300.13–1,921.55
Cluster mass range	598.26–6,527.06
Base peak range	120–520,083
Number of clusters with peptide identity	5,414
Number of identified accessions	8,044
Range of peptides/accession	1–64
Range of accessions/peptide	1–212

erances of 20 ppm for *m/z* and mass and 1 minute for retention times, 16,874 (52%) peptide clusters were matched across both datasets, of which 5,414 bore peptide identification results. These identified peptides matched 8,044 *T. aestivum* protein accessions. Our experimental results are summarised in Table 1; number of identified peptide numbers aside, they compared well with our previous findings during method optimisation [37].

Our strategy was to consider all 8,044 protein hits identified from the 5,414 sequenced peptides irrespective of their homology. We thus turned the wide table of 5,414 peptides × 212 protein accessions into a long table containing 32,347 rows of peptides assigned to unique protein entries and replicated the quantitative data accordingly for statistical analysis purposes. The list of all identities is captured in Supplementary Table S3. Up to 64 unique peptides matched a particular protein with an average of 4 peptides per hit (Supplementary Fig. S5A, B).

A given peptide matched to up to 212 protein accessions with an average of 6 hits per peptide (Cluster_29,452, VLQQLNPCK, Supplementary Fig. S5C, D). This mirrored the high frequency of homoeologous proteins in the hexaploid wheat samples expressed from 3 similar subgenomes, A, B, and D [64]. Another compounding factor was that wheat protein accessions were created from genomic sequences, resulting in multiple protein entries bearing identical sequences but arising from different gene accessions [2]. This created, on one hand, protein identities labelled as "fragments" despite having a complete coding region and, on the other hand, other entries lacking this tag despite having an incomplete coding region (Supplementary Table S3). Finally, the vast number of PTMs identified here also contributed to boosting hits against a particular peptide AA sequence. The most dominant wheat grain proteins are storage proteins such as gliadins and glutenins, which featured prominently in our proteome (Supplementary Fig. S5E, Supplementary Table S3), despite the fact that their low Lys/Arg content makes them less prone to trypsin digestion [2]. Other major proteins comprised histones, beta-D-glucosidases, and ubiquitin. This list of identified proteins compared well with our previous methodological work [37]. Other recent studies on mature wheat seed proteome using gel-based or gel-free technologies also published a comparable list of protein identities [65–67].

Application to a wheat industry problem: late maturity alpha-amylase (LMA)

By unravelling the genetic, biochemical, and physiological mechanisms that lead to LMA expression, scientists strive to understand and eliminate LMA from wheat breeding programs [35]. Surprisingly, postgenomics is not one of the strategies adopted by researchers to close the biological knowledge gap, with only 1 transcriptomics study registered so far [18]. Our study constitutes the first proteomics experiment performed to decipher the mechanisms involved. Machine learning was performed on the complete dataset to distinguish LMA-susceptible from nonsusceptible wheat genotypes without success (data not shown). Results from statistics and data mining are described and discussed below.

Getting the quantitative data ready for statistical analyses

Assessing the normality of LC-MS1 datasets

To assess whether our LC-MS1 datasets following the correction and z-transformation steps were normally distributed, we plotted the data as histogram and boxplot. We further performed the nonparametric 1-sample Kolmogorov–Smirnov (K-S) test [68] well suited to analysing big data [69]. Both histogram and boxplot of the corrected data were asymmetrical, with most values being on the low range (Supplementary Fig. S6A, B), which revealed that this dataset was not normally distributed. This was confirmed by the high K-S statistics (D) of 0.41 and a very low *P* value (< 2.2 e^{16}).

Using the z-transformed data, the histogram and boxplot were more symmetrical (Supplementary Fig. S6C, D). Whilst the K-S statistics (D) was reduced to 0.27, it was still too high to conclude to normality. Even though we did not achieve a Gaussian distribution by standardising the data, we managed to make it more even, which improved statistical analyses for biomarker discovery.

Partial least squares of unbiased samples to select a meaningful set of LMA-responsive peptides

Analysing such a large dataset (3,990 columns × 32,337 rows) was computationally taxing, necessitating extensive dwell times to finalise statistical analyses and often triggering Genedata sever crashes due to out-of-memory failures despite recent upgrades. Consequently, we devised a strategy to select a subset of relevant peptides via the supervised cluster method partial least squares (PLS). Using the 934 unbiased samples and all 32,337 peptides (including Cluster_AAA), we executed a PLS analysis with LMA trait as a response. The score plot of the first 2 components showed that the PLS successfully pulled out the grain samples exhibiting high LMA activities (Supplementary Fig. S7A).

The corresponding loading plot allowed us to categorise peptides according to their contribution to the PLS model via their Variable Importance in Projection (VIP) scores. The most-contributing peptides (i.e., exhibiting the highest VIP score) were in the plot area equivalent to that of high LMA samples (Supplementary Fig. S7B).

VIP scores indicated the importance of each variable (peptide) in the projection used in the PLS model. Peptide VIP scores were calculated as weighted sums of the squared correlations between the PLS components and the original peptides; weights were inferred from the percentage variation explained by the PLS component in the model [70]. VIP scores greater than 0.5, 1.0, and 1.5 segregated 14,440 (45%), 7,252 (22%), and 2,996 (9%) peptides, respectively. By setting up 3 VIP score thresholds of increasing stringency, we thus created 3 subsets of peptides of decreasing sizes that could be used in more computationally demanding processes.

Wheat subsampling to create an unbiased dataset and transforming LMA trait profile to achieve normal distribution

In the 3,990 reproducible wheat samples, 3,773 featured LMA measurements that ranged from 0.04 to 7.95 U/g (Supplementary Table S1), albeit mostly on the low scale with 88% of the values recording less than 0.2 U/g (Fig. 4A), which corresponds to the 300-second receival threshold of FN [14, 19].

Our range far exceeded those reported earlier, spanning either 0.08 to 0.67 U/g across 33 spring wheat cultivars grown across 18 field sites [71], 0.023 to 1.417 U/g over 39 varieties grown under controlled and triggering LMA conditions [19], or 0.002 to 1.977 U/g among 196 genotypes from 3 experimental locations [15]. We chose a threshold of 0.17 as a tipping point to delineate between grain samples displaying either low (3,306 samples) or high (467 samples) alpha-amylase activity. The LMA profiles below and above this arbitrary value showed a slow gradual increase of enzyme activity up to 3.2 units where datapoints became more scattered (Fig. 4B, C). Because the LMA distribution was significantly skewed towards low values and to restore balance to the trait profile, we retained all the wheat samples with an LMA above 0.17 (467 samples) and randomly selected 467 samples (out of 3,306) for which LMA fell below this threshold. The LMA profile of this unbiased subset of 934 samples (Fig. 4D) was very similar to the complete distribution (Fig. 4A).

When LMA measurements were plotted as a histogram, it confirmed the skewness towards low activities and highlighted that most values fell between 0.068 and 0.203 U/g (Fig. 4E). A natural logarithm transformation did not make the data Gaussian (Fig. 4F), nor did other logarithmic bases (data not shown). A binary logarithm function was used to transform LMA data to ascertain the significant negative correlation with FNs [15, 19]. FNs inferior to 300 seconds, which is the commercial trade cutoff manifesting significant alpha-amylase activity, corresponded to a $\log_2$ LMA value of -3 [19]. In our work, an inverse function normally distributed LMA values, albeit as a slightly asymmetrical bell curve (Fig. 4G). This INV(LMA) data were further standardised (centred around zero and scaled down to comparable variance) when it was incorporated at the peptide level, which did not compromise its Gaussian distribution (Fig. 4H).

Predicting LMA missing values

Out of the 3,990 reproducibly processed grain samples, 217 were not measured for LMA. We employed a univariate PLS regression strategy to impute them. Using our 2,996 peptide set with the highest VIP scores, we tested various PLS regression models (data not shown) on a random selection of 179 samples out of the 934 unbiased sample set, which ranged from 0.5 to 4.9. This testing set was analysed against the remainder of the unbiased set (755 samples). The best regression model utilised 20% of the valid values and 20 latent factors; it predicted the 179 tested values with 93% accuracy (Supplementary Fig. S8A).

This model was not accurate for small LMA values with an R^2 of 6%, even imputing negative values (Supplementary Fig. S8B). Yet, it was 98% accurate for LMA measurements greater than 0.17 U/g (Supplementary Fig. S8C). It was more critical to faithfully estimate high LMA values given that it was the criterion for grain soundness; our PLS regression (PLSR) model fulfilled this. We applied the model's parameters to predict the 217 LMA missing values against the unbiased set of 934 samples; the imputations ranged from -0.29 to 0.63 U/g (Supplementary Fig. S8D). The negative values were converted to zeros. LMA predictions are reported in Supplementary Table S1.

The simplest method for imputing missing data relied on single-value imputation, such as the mean [72], whilst more complex methods were based on regression [73] or K-nearest neighbours (KNN), which estimates a missing data point using distances calculated from its most similar neighbours [74]. Invented in 1966 [75], PLS regression has become very popular notably in the fields of bioinformatics [76] and spectroscopy [77]. Nengsih and colleagues [78] demonstrated that while computation times increased with the proportion of missing data, up to 30% missing values could be imputed using PLSR. In our study, LMA was the single trait provided to analyse LC-MS1 data. Not imputing missing LMA measurements meant that 5.4% (217/3,990 samples) of our dataset would have been useless; therefore, it was a worthwhile effort. Along with PLSR, we have also tested multivariate linear regression (MLR), univariate polynomial regression, and KNN imputation by varying several parameters, including valid value percentage, number of latent factors, number of parameters (for MLR), and distance computation and number of K (for KNN), albeit without success (data not shown).

Incorporating LMA trait at the peptide level for biomarker discovery

Because we only had a single trait to make biological sense of our big data, we introduced all 3,990 LMA values (including the predicted values), which characterised wheat samples at the peptide level by transposing it and renaming "Cluster_AAA." This added 1 extra row to our dataset of 32,336 peptides to make a final matrix of 3,990 columns (wheat samples) and 32,337 rows. This way, we could apply statistical analyses that would group peptides that behaved similarly or conversely to our LMA trait, thereby facilitating biomarker discovery. To permit the comparison between LMA and grain peptides, we first needed to normalise and standardise LMA values prior to their transposition.

Having LMA incorporated with wheat grain peptides (as Cluster_AAA) further helped us assess the relevance of the statistical tests carried out by validating anticipated results. For instance, when performing a correlation analysis with LMA, as expected, Cluster_AAA achieved a positive correlation of 1. In another instance, when executing a 1-factor linear model with LMA as a covariate, Cluster_AAA was confirmed to yield a q value of 0. Finally, when performing multivariate clustering analyses (hierarchical clustering analysis [HCA], self-organising map [SOM], k-means), this strategy assisted us in finding peptides with profiles similar to that of Cluster_AAA.

Statistical analyses to discover LMA-responsive biomarkers

Big data produced by gene expression studies are too large to analyse by mere sorting in spreadsheets or plotting on few charts. Multivariate data analyses such as clustering and correlating methods are required to make sense of the data [79, 80]. Yet, as helpful as these multivariate analyses are, they are not as statistically robust as uni- or bivariate analyses [79] to test the relationship between peptides and LMA. We thus performed a few uni-, bi-, and multivariate analyses to explore our large dataset against our single LMA trait.

Unsupervised multivariate clustering analyses (SOM, k-means, HCA) for pattern recognition and peptide profiling of LMA phenotype

As multivariate analyses handle integral datasets and iteratively impute many statistics, they incur heavy computational costs. Suffering multiple Genedata server crashes, we could only apply such methods to a subset of our data. Using the unbiased set

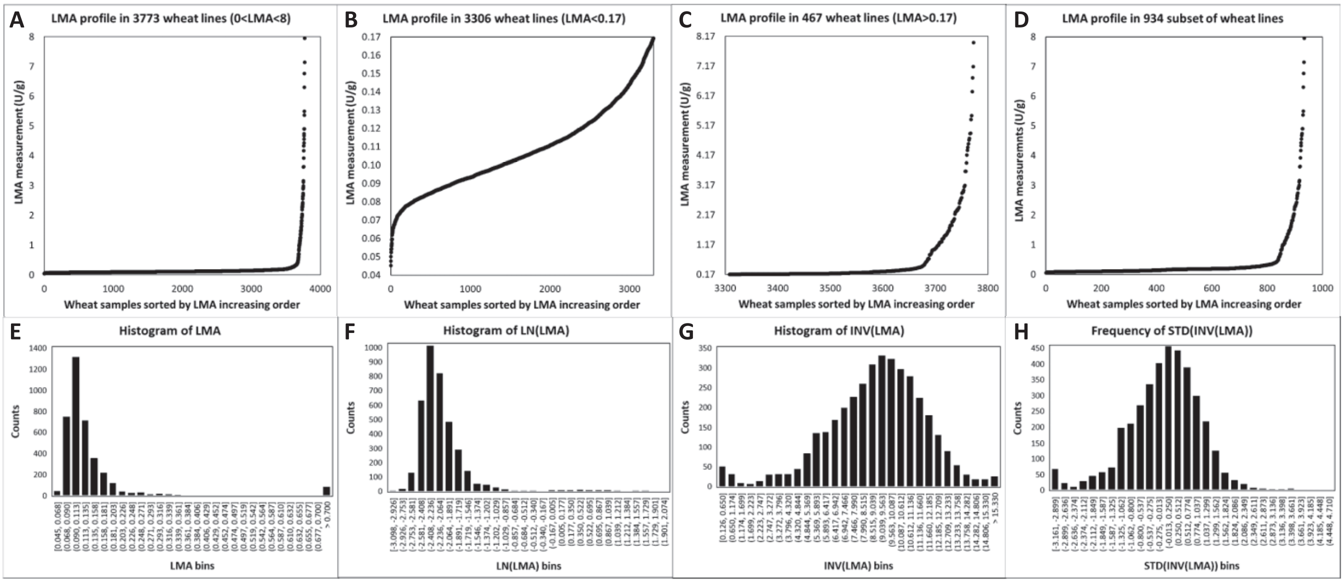


Figure 4: Profiles of LMA measurements for each wheat sample sorted by increasing values illustrated as scatterplots (A–D) and histograms (E–H). (A) Scatterplot of LMA values assayed in 3,773 wheat samples. (B) Scatterplot of LMA values less than 0.17 U/g in 3,306 wheat samples. (C) Scatterplot of LMA values equal to or greater than 0.17 U/g in 467 wheat samples. (D) Scatterplot of LMA values in unbiased set containing 934 samples (see Section 2.8.2 for explanation). (E) Histogram of LMA values assayed in 3,773 wheat samples along 30 bins. (F) Histogram of LMA values assayed in 3,773 wheat samples and transformed using a natural logarithm (LN) function along 30 bins. (G) Histogram of LMA values assayed in 3,773 wheat samples and transformed using an inverse function (1/LMA = INV(LMA)) along 30 bins. (H) Histogram of LMA values assayed in 3,773 wheat samples and transformed standardising the inversion function (STD(INV(LMA))) from panel G along 30 bins.

of 934 wheat samples and the list of 7,254 peptides with LMA-responsive VIP scores above 1, we have performed 3 unsupervised clustering analyses, SOM, k-means, and divisive HCA. Because we had incorporated the LMA trait at the peptide level as Cluster_AAA, we could look for groups resulting from these analyses, which assembled peptides behaving similarly to Cluster_AAA. Clustering or cluster analysis corresponds to a set of learning methods grouping observations that share similar characteristics. Within a set of related values of the variables analysed, these methods find feature patterns that generate clusters that group similar observations [81]. Unsupervised clustering analyses are commonly employed in gene expression studies [80].

In our experiment, the SOM model yielded 48 groups comprising 8 to 555 peptides with mean distances from 0.09 to 0.80. The group including Cluster_AAA (4,3) contained 26 biomarker peptides; its distance from the group centre ranged from 0.00 to 0.83 with a mean of 0.38 and an SD of 0.31 (Supplementary Table S4). Cluster_AAA stood 0.70 from the group centre. While SOM has been widely used in exploratory data analyses in diverse fields [82], it has only been applied to proteomics in the context of animal cell culture [83], GPI anchor prediction [84], transmembrane helix predictor [85], protein conformation [86], or protein–protein interaction [87] but never in plant grains.

We tested different number of neighbours (k) and observed that the larger k, the greater the variance explained by the k-means model (data not shown). Applying the biggest k possible (20 neighbour groups) produced a model that overall explained 71.1% of the variance. Neighbour group 14 with a variance of 35% contained 93 biomarker peptides spanning a distance of 0.12 to 0.94, including Cluster_AAA, whose distance was 0.79 (Supplementary Table S4). K-means clustering was well adopted by the proteomics community to group gene products of similar profiles, notably in plants such as bamboo [88], nightshade [89], or grape [90] but, to our knowledge, not in wheat. In developing corn grains, coordinated protein expression associated with different functional categories was revealed by a k-means clustering analysis [91].

We successfully applied an agglomerative 2D HCA to cluster both samples and peptides (data not shown) but failed to select individual cluster groups to retain the one hosting Cluster_AAA. Instead, we performed a divisive HCA, which ordered the peptides into clusters that could then be chosen individually. Cluster_AAA belonged to a group of 33 biomarker peptides (order 1915–1947, Supplementary Table S4). We could not find in the literature any proteomics study that resorted to divisive HCA; conversely, classic (agglomerative) HCA created in 1998 [92] and its extension 2D HCA [93] are widely used by the community, including wheat scientists [94–98]. Using agglomerative HCA on two-dimensional electrophoresis–resolved proteins, Tasleem-Tahir distinguished 9 expression profiles throughout wheat grain growth, from anthesis to maturity [98]. In their gel-free iTRAQ analysis of early developing wheat endosperms (from 7–28 days postanthesis [DPA]), Ma and colleagues employed HCA to delineate starch processes [96]. Similarly, 5 major protein expression patterns across developmental stages 4–12 DPA were outlined using HCA [99]. HCA was also employed to explore the change in expression of embryo and endorsperm proteomes during wheat seed germination [100]. In their comprehensive proteomics and proteogenomics study of key developmental stages of 24 wheat organs and tissues, Duncan and colleagues showed that HCA faithfully assigned samples to 3 main clusters corresponding to photosynthetic tissues (leaves, bracts, and other green organs); non-photosynthetic, developmental, and reproductive organs (pollen, stem, anther, coleoptiles, roots, immature spike); and grain (developmental series, embryo, pericarp, endosperm) [94]. More recently, Cao and colleagues discriminated differentially expressed proteins in 2 wheat lines using HCA [65]. All these reports demonstrate that genotype, sample, and tissue specificity of protein profiles can be highlighted using unsupervised clustering tools.

Bivariate analyses (correlation and linear regression) to consider each individual peptide against LMA

As bivariate analyses handle only 2 variables at a time, they are not computationally taxing. We were thus able to apply such methods on our complete dataset comprising 3,990 samples and 32,337 peptides (including Cluster_AAA). Due to the quantitative nature of the LMA trait, we could not perform an analysis of variance (ANOVA). We have thus carried out 2 bivariate analyses: a correlation and a linear model. Because we had incorporated the LMA trait at the peptide level as Cluster_AAA, we could assess the validity of our analyses based on the outputs produced by the latter.

In our experiment, correlation coefficients ranged from −0.07 to 0.3, except for Cluster_AAA, which as expected attained absolute positive correlation with an R^2 of 1 (Supplementary Table S4). Our coefficients do not show a strong relationship between peptide profiles and LMA. We arbitrarily chose an absolute value of 0.15 to retain any LMA-associated peptide, which excluded all negatively correlated features but included 28 positively correlated biomarkers. Correlation analyses are frequently employed in proteomics to unravel proteins underpinning particular sample types, conditions, or traits [101], and wheat is no exception [102–110]. Concordance of transcript and protein profiles in wheat grain was assessed via correlation coefficients, which increased with seed maturity [103, 109]. Grain yield and grain protein content were observed to be negatively correlated, yet both also positively correlated to nitrogen availability in a wheat genotype-specific manner [111].

The *q* value for the linear regression slope indicates whether changes in the explanatory variable are significantly linked with changes in the outcome. In our work, we looked for significant relationships between the 32,337 peptides (including Cluster_AAA) and the inverse function of LMA, which assumed normality as a covariate factor. The *q* values ranged from 6×10^{-8} to 1, except for Cluster_AAA, which exhibited a *q* -value of 0, as expected (Supplementary Table S4). We arbitrarily applied a 5% *q* -value threshold to consider 494 biomarker peptides whose change in expression profiles was significantly linked to variation in LMA measurements. Linear mixed models are regularly employed by the proteomics community for biomarker discovery approaches [112–115] but, as far as we know, not on wheat grains.

Compiling all statistical analyses to generate a list of candidate peptides and binning LMA values for biomarker profiling and *t* test

In this study, LMA-responsive biomarkers were selected based on the statistical analyses presented above and had to fulfil at least one of the following criteria: belong to the SOM group (4,3), be included in k-means group 14, bear a divisive HCA order from 1915 to 194, exhibit a correlation R^2 greater than 15%, or display a *q* value inferior to 5%. This created a list of 531 biomarkers, most of which fulfilled several statistical criteria and all of them exhibiting a VIP score for the LMA-responsive PLS greater than 1 (Supplementary Table S4).

When attempting to chart the biomarker profiles, we were faced with the challenge of plotting 3,990 datapoints per gene product, which ruled out typical line graphs, scatterplots, histograms, or utilising oversized illegible heatmaps to represent all data points simultaneously (data not shown). We consequently adopted a data reduction strategy involving binning the samples into 8 or 2 arbitrary bins based on their LMA values.

The 8-bin profiling comprised all 3,990 samples sorted by increasing LMA measurements and partitioned into 8 groups of equal sample size (~499 samples/bin, Supplementary Table S1). Plotting the average of each bin as a line chart faithfully maintained the pattern of LMA measurement observed in Fig. 4A with a flat profile for the first 7 bins followed by a steep increase in the last bin (Supplementary Fig. S9A).

This profiling strategy was not used for statistical purpose but proved very useful during data mining of all identified 5,514 peptides upon using tools that offered quantitative charting such as Pathway Tools and Circos (see below).

The 2-bin profiling only featured the 934 unbiased samples separated according to an arbitrary 0.17-U/g threshold (Supplementary Table S1). Plotting the average of each bin as a histogram clearly displayed a marked quantitative increase from bin 1 to bin 2 (Supplementary Fig. S9B). This simple representation tool allowed us to categorise the 531 biomarkers as being either upregulated when bin 2 was taller than bin 1, denoting an accumulation in samples with LMA >0.17 U/g, or downregulated when bin 1 was taller than bin 2, denoting an accumulation in samples with LMA <0.17 U/g.

This oversimplified binning scheme allowed us to perform one last statistical analysis on the 532 biomarkers (including Cluster_AAA) using the unbiased set of 934 samples—namely, a Student *t* test with an effect size. We generated a volcano plot based on the *P* values and the directed effect size (i.e., fold change), which clearly delineated the biomarkers according to their accumulation in bin 1 or 2 (Fig. 5A).

More LMA-related biomarkers were upregulated (325) than downregulated (206) according to our 2-bin profiling. This was explained by the fact that all our statistical analyses, bar the PLS and linear model, favoured peptides behaving similarly to Cluster_AAA, a proxy to LMA actual measurements. Some exemplary patterns are displayed as histograms with error bars and compared to that of Cluster_AAA to expose the assortment of up- and downregulation profiles (Fig. 5B). Because the 2-bin representation was very reductive, we also present a heatmap of all the intensities of the 532 biomarkers (including Cluster_AAA) sorted by directed effect size (i.e., fold change) in each of 934 unbiased wheat samples organised by HCA cluster order (Fig. 5C). No strong differential expression trend appeared apart from a horizontal gradient of colours from left to right denoting the change from up- to downregulation of the biomarkers and a swap in colour vertically, suggesting that samples were efficiently classified by the HCA. Despite merely featuring a small subset (934 × 532) of our global dataset (3,990 × 32,337), the heatmap looked noisy and remained very hard to interpret due to an excessive number of data points (469,888 quantities) and the lack of a visually striking pattern. This further reinforced the need to devise simple representations tools such as a volcano plot when reporting results on big data.

To our knowledge, volcano plots have not been widely adopted by the proteomics community, let alone wheat grain scientists with only 1 report so far [67], unlike heatmaps, which are frequently reported in proteomics publications [116]. In our work, we sorted the 531 biomarker peptides according to their 2-bin fold changes and wheat sample based on their LC-MS molecular similarity (Fig. 5C). Zang and colleagues have adopted heatmaps to profile the proteins underpinning seed tissue organogenesis [117].

Mining biomarkers to make biological sense of the data

Among the 531 biomarkers that exhibited significance levels in response to LMA measurements, 390 were identified by LC-MS2 and

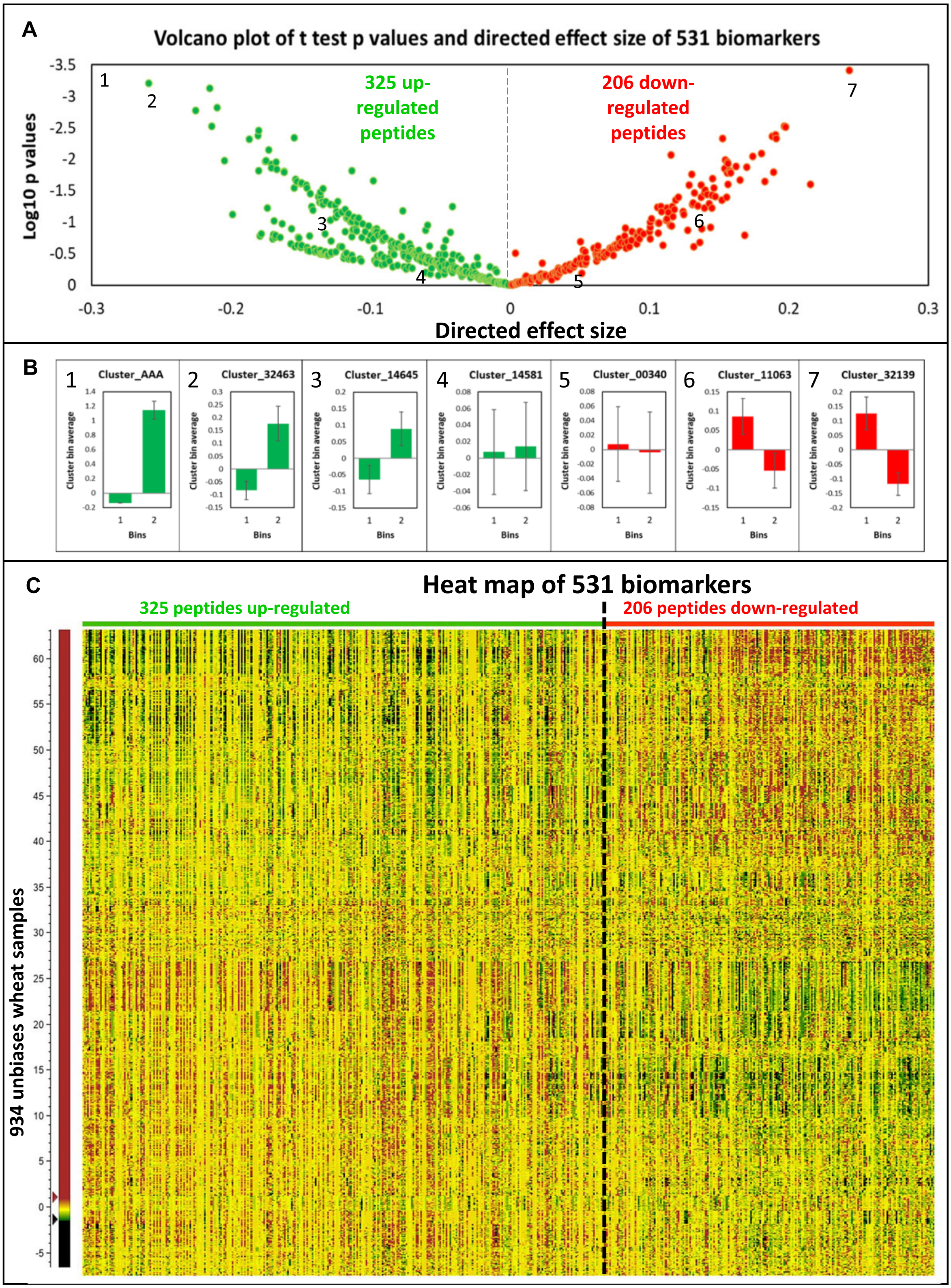


Figure 5: Volcano plot from *t* test and heatmap of up- and downregulated 531 biomarkers using the unbiased set of 934 wheat samples. (A) Volcano plot of the 325 upregulated and 206 downregulated biomarkers. Numbers position exemplary peptides plotted in panel B. Cluster_AAA with coordinates (−1.2, −23.5) is an outlier in the upper left corner and is not featured for display purpose. (B) Mean histograms along 2 bins of clusters illustrating up- and downregulation patterns and located with numbers on panel A. Standard errors are depicted with the vertical bars. Bin 1 corresponds to 467 samples with LMA <0.17 U/g and bin 2 corresponds to 467 samples with LMA >0.17 U/g. (C) Heatmap corresponding to the volcano plot in panel A with peptides sorted according to directed effect size and samples sorted based on HCA cluster order.

matched 3,798 protein accessions (Supplementary Table S5). This list included the most abundant and homoeologous proteins such as the prominent storage and starch-related proteins, gliadins, glutenins, avenins, and starch synthases as well as constitutive proteins such as histones, protein disulfide isomerases, and tubulin, or else stress-related proteins such as heat shock and 14–3–3 proteins. We did not identify any peptides belonging to LMA in this study, likely because we did not target high LMA samples. To visualise our peptides of interest in a biological context, we have undertaken a series of data-mining steps. We have also made use of our 8- or 2-bin profiling strategy when using quantitative mapping tools. The 2-bin profiling is hereafter referred to it as up- or downregulated gene products. The data-mining tools presented below suited wheat proteins. Many other *in silico* tools are freely available online, which we encourage the community to employ; however, we would not recommend using String or PlantReactome, which in our hands yielded very little results (data not shown).

Protein descriptions and GO terms from UniProtKB

Out of the 8,044 identities, 7,939 could be mapped in UniProtKB, which flagged 6,457 GOMF terms, 3,769 GOCC terms, 3,991 GOBP terms, and 1,385 unique protein names (Supplementary Table S3). Power BI proved very useful to mine identified peptides and simultaneously plot some of their features as histogram, scatterplot, pie chart, violin plot, tree map, and word cloud into a single dashboard (Supplementary Fig. S10A) and then drill down on some aspects, for instance, inhibitor (Supplementary Fig. S10B) or deamidation (Supplementary Fig. S10C).

The protein names were turned into word clouds and the most frequent GO terms for each category were presented as tree maps. Standing out from the cloud were the words "protein," "containing," "domain," "subunit," "glutenin," "LMW," "molecular," and "weight," confirming the preponderance of LMW glutenin subunits and domain-containing proteins such as AAI domain-containing protein homoeologous to alpha-amylase inhibitors (Supplementary Fig. S11B–D). Also predominant among identified proteins were the words "alpha" and "gliadin." Word cloud is a text-processing method that offers an efficient and compact visualization of the most frequent terms in a text [118], yet it seldom appears in the scientific literature. It has been cleverly used to categorise moonlighting proteins [119] or depict the history of GOMF terms [120], but not in the wheat proteome. Representing our 390 identified LMA-responsive biomarkers as word clouds revealed that upregulated peptides belonged predominantly to alpha-gliadins, whereas downregulated peptides mostly matched LMW glutenins (Fig. 6A, F).

Rather than adopting a pie chart or histogram to plot the GO terms of all identified proteins as commonly reported, we opted for tree maps, which were initially implemented for microarray data [121, 122] and later integrated into the web server REVIGO [123] used during our wheat method optimisation [37]. For all 8,044 identified proteins in the present study, we generated the tree maps for all 3 GO classes using Power BI as it afforded more display options than REVIGO. The most frequent biological processes (GOBP) were "polysaccharide catabolic process" (5,643), "starch biosynthetic process" (3,688), "nucleosome assembly" (3,626), "protein folding" (2,950), and "protein refolding" (2,499) (Supplementary Fig. S11E). "Cytoplasm" (11,888), "extracellular region" (9,964), and nucleus" (7,478) were the most common cellular components (GOCC); recording 3,687 entries, the amyloplast was listed in the sixth position (Supplementary Fig. S11F). With 37,308 occurrences, the "nutrient reservoir activity" was by far the most recurrent molecular function (GOMF), followed by "ATP binding" (7,012) and "serine-type endopeptidase inhibitor activity" (5,811) (Supplementary Fig. S11G). The list of dominant proteins and associated GO terms in this work pointed to a storage organ such as the wheat seed and confirmed what has previously been reported in wheat grain [37, 109, 117, 124–126]. All GO terms against the 390 identified LMA-related biomarkers are listed in Supplementary Table S5. The 207 upregulated biomarkers came mostly from cytoplasmic and chloroplastic proteins involved in protein translation and folding, with ATP binding activities (Fig. 6B). The 183 downregulated peptides predominantly belonged to cytoplasmic and cytosolic proteins acting in protein folding and TCA cycle and bearing ATP binding activity (Fig. 6G).

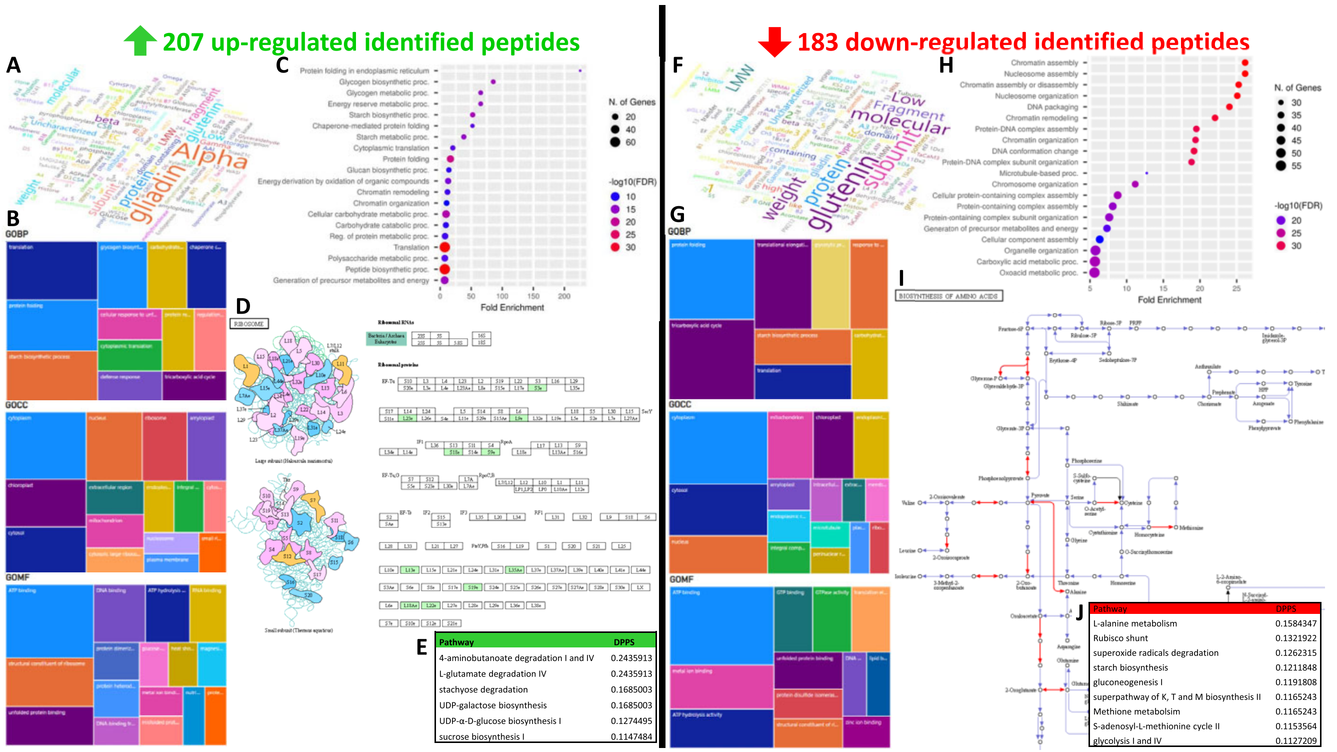
KEGG to retrieve Pathway, Brite, and Module names

From the 8,044 fasta sequences, 677 unique KEGG Orthologs (KOs) could be retrieved, which mapped to 327 KEGG pathways, 41 brites, and 117 modules and annotated 11,888 peptides (Supplementary Table S3). Identified proteins belonged to 179 (26%) KEGG metabolic pathways with 109 (16%) KOs involved in the biosynthesis of secondary metabolites (Supplementary Fig. S12A), including sugar-related enzymes such as amylases, sucrose synthases, hexokinases, fructokinases, and beta-glucosidases.

Half of KOs pointed to enzymes (336), then exosomes (71, 10%), ribosomes (62, 9%), and chromosome-associated proteins (60, 9%) (Supplementary Fig. S12B). Primary metabolisms such as glycolysis, TCA cycle, and gluconeogenesis were prominent KEGG modules (Supplementary Fig. S12C). Unexpectedly, 62 KOs (exclusively ribosomal proteins) were associated with the "coronavirus disease—COVID 19" pathway. Similarly, many proteins were linked with other human-related afflictions (e.g., sclerosis, neurodegeneration, and Parkinson, Huntington, Alzheimer, and prion diseases; Supplementary Fig. S12A). This demonstrated the limitations of using generalist databases like KEGG that are mostly relevant to human research to map plant proteins. While KEGG plant interface exists [127], plant-related datasets are dispersed throughout the whole KEGG server so that one cannot exclusively mine plant-specific entries. There is a need for future KEGG iterations to restrict searches to relevant taxa. Notwithstanding nonplant hits, pathways symptomatic of grains were accurately captured in this experiment such as the carbon metabolism (42, 6%), glycolysis/gluconeogenesis (25, 4%), and the starch and sucrose metabolism (18, 3%) (Supplementary Fig. S12D–F). Despite the constraint raised above, KEGG remains a database widely employed to explore plant proteomes, including wheat grain proteins [37, 128–130]. Mapping our 390 LMA-associated biomarkers (Supplementary Table S5) highlighted that many upregulated peptides came from ribosomal proteins (Fig. 6D) while several downregulated peptides belonged to enzymes acting in the biosynthesis of AAs (Fig. 6I).

ShinyGO to retrieve enriched functional categories and chromosomal positions

Multiple online tools exist to efficiently mine GO terms, but only a few cater for nonmodel species, let alone plants [131–133]. When looking for relevant mining tools during our method development stage, we resorted to the AgriGO online program, which specifically focused on agricultural species and offered valuable illustrations to display enrichment sets [37]. Unfortunately, AgriGO server is no longer available. We have found instead ShinyGO [134], recently developed, which not only surpassed AgriGO in terms of enrichment visualisations but also provided wheat protein chromosomal positions, desirable for Circos plots. A downside of ShinyGO



Pathway	DPPS
4-aminobutanoate degradation I and IV	0.2435913
L-glutamate degradation IV	0.2435913
stachyose degradation	0.1685003
UDP-galactose biosynthesis	0.1685003
UDP-α-D-glucose biosynthesis I	0.1274495
sucrose biosynthesis I	0.1147484

Pathway	DPPS
L-alanine metabolism	0.1584347
Rubisco shunt	0.1321922
superoxide radicals degradation	0.1262315
starch biosynthesis	0.1211848
gluconeogenesis I	0.1191808
superpathway of K, T and M biosynthesis II	0.1165243
Methione metabolsim	0.1165243
S-adenosyl-L-methionine cycle II	0.1153564
glycolysis I and IV	0.1127209

Figure 6: Data mining of up- and downregulated biomarkers. (A, F) Word cloud of protein names. (B, G) Tree maps of GO terms for BP, CC, and MF categories. (C, H) Dot plots from ShinyGO. (D, I) Most significant KEGG pathways, ribosomes for upregulated biomarkers and AA biosynthesis for downregulated biomarkers. (E, J) Differentially perturbed pathways (DPPS) from Pathway Tools.

was that it did not perform well with UniProt accession IDs, hence the prerequisite to retrieve TRAES IDs from UniProtKB. A total of 6,622 TRAES accessions corresponding to the 8,044 UniProt proteins were thus retrieved, of which 4,571 could be mapped by ShinyGO (Supplementary Table S6). An enrichment analysis ensued and could be visualised as a chart, tree, network, and chromosomal map; density plots and histograms were also produced (Supplementary Fig. S13).

The most enriched category was the TCA cycle with a fold enrichment in excess of 12.5 and the most significant GO classes were translation and peptide biosynthesis with an False Discovery Rate (FDR) inferior to e^{-160} (Supplementary Fig. S13A, E). Protein folding and ribonucleoprotein complex biogenesis stood out as well among the proteins identified in this study (Supplementary Fig. S13B). Identities covered the whole genome with lower density around centromeres (Supplementary Fig. S13F). ShinyGO and other online data-mining algorithms were employed to predict genetic components systems implicated in the plant model species *Arabidopsis* in response to highlight from transcriptomics datasets publicly available [135]. Our results exemplify the relevance of ShinyGO for nonmodel plant species; we could not find other cereal reports making use of it, probably due to its recent emergence [134]. A fold enrichment exceeding 200 was found among the 207 upregulated peptides from gene products involved in protein folding in endoplasmic reticulum (Fig. 6C), followed by glycogen metabolism, energy reserve, and starch biosynthesis. ShinyGO enrichment analysis produced very different results for our 183 downregulated peptides, mostly invoking chromatin assembly and remodeling, nucleosome assembly and organisation, DNA packaging and conformation change, and protein–DNA complex assembly and organisation (Fig. 6H).

Pathway Tools to retrieve differentially perturbed pathways based on 8-bin profiling

As useful as the programs described above are, they yet do not accommodate quantitative data, unlike Pathway Tools [136]. It was made available online by the Plant Metabolic Network server and curating the PlantCyc databases encapsulating 126 plant and algae species, including BreadwheatCyc [137]. We could thus display protein expression data on pathway diagrams in a dynamic and interactive way. Using the 6,622 TRAES accessions corresponding to the proteins identified in this study and the quantitative data averaged along 8 bins, we mapped 1,432 proteins in the *T. aestivum* Pathway Tools website (Supplementary Fig. S14A).

The change in expression profiles along the 8 bins was recorded and showed that all peptide quantities varied across sample groups with multiple trends throughout the whole cellular overview (Supplementary Video SV1). As previously reported [37], the primary and secondary metabolisms were well covered. Overall quantities of homoeologous wheat proteins involved in TCA and glyoxylate cycles declined along 8-bin expression profiles (Supplementary Fig. S14B).

Also featured was plant hormone biosynthesis (Supplementary Fig. S14C), which was not highlighted in the other exploratory tools, thus demonstrating the superiority of *T. aestivum* Pathway Tools over other databases [37]. The 8-bin profiling hinted at an accumulation of proteins related to auxin, cytokinin, and gibberellin biosynthesis and a reduction of enzymes participating in 5-deoxystrigol, brassinosteroid, and jasmonate synthesis in LMA-rich samples. Hormonal response was flagged as one of the biochemical mechanisms of LMA expression, in particular gibberellin and ABA signalling [18, 21, 138]. Focussing on the ent-kaurene biosynthesis, expression patterns accumulated in low LMA samples at the initial step of the pathway and diminished in high LMA samples at the last step (Supplementary Fig. S14D, E). The first biosynthetic step is controlled by ent-copalyl disphosphate synthase (TaCSP), which was reported to be associated with LMA via a major locus on wheat chromosome 7B, accordingly renamed as LMA-1 [139]. TaCSP (Cluster_22,809 in Supplementary Fig. S13F) was one of our biomarkers. Even though databases such as Pathway Tools mapped TaCSP to the gibberellin metabolism, its function with this phytohormone was recently contested, and it was suggested that high pI alpha-amylase synthesis in the aleurone of developing wheat grains would be independent of gibberellins during LMA response [36]. Other biomarkers matching phytohormone-associated proteins included a cytokinin dehydrogenase whose decreasing pattern picked up in the bin containing all the wheat sample registering high LMA (Cluster_24,683 in Supplementary Fig. S14F) and a responsive to ABA (Rab) protein (Cluster_36,748 in Supplementary Fig. S14F) whose expression profile closely resembled that of Cluster_AAA. Interestingly, Cluster_24,621 with an increasing expression profile belonged to an uncharacterised protein annotated with GO terms "Response to Auxin" and "Response to ethylene" (Supplementary Fig. S14F).

Because Pathway Tools handles quantitative data, it produced lists of differentially perturbed pathways (DPPS) for each cohort of up- and downregulated biomarkers. Pathways characterising wheat grains with high LMA measurements were degradations of aminobutanoate, glutamate, and stachyose, as well as biosynthesis of UDP-galactose, UDP-glucose, and sucrose (Fig. 6E). DPPS-differentiating samples with low LMA activities were AA metabolisms (A, K, T, and M), rubisco shunt, superoxide radical degradation, starch biosynthesis, gluconeogenesis, S-adenosyl-M cycle, and glycolysis (Fig. 6J). Our method study aside [37], we could not find any other wheat gene expression study utilising this impressive PlantCyc database. However, work on other plant species has amply demonstrated its value [140–145].

Circos plot to visualise chromosomal positions, expression profile, and statistics of identified proteins and biomarkers

Invented over a decade ago [146], Circos plots have proven so valuable to efficiently and enticingly represent qualitative and quantitative information that a multitude of emulations have since arisen, including its packaging within the Galaxy server [147], which we took advantage of here. When the IWGSC released theT. *aestivum* genome and published their findings, the genomic features were elegantly and succinctly captured in a circular plot, which highlighted homeologous genes and translocated chromosomal regions [7]. Being infinitely flexible, Circos plots can chart any data as multiple concentric circular layers provided the correct file format is applied. We opted to chart proteins encoded by genes we could locate on the genome (chromosomal positions retrieved from ShinyGO analysis) and overlay their expression profiles, along with some statistics of candidate LMA-responsive biomarkers (Fig. 7).

Proteins identified in this experiment aligned with the full genome, densely covering each chromosome albeit less so around centromeric regions (Fig. 7B). Overall, expression profiles along 8-bin accumulated in bins 1–6 corresponding to wheat samples with low LMA and decreased in bins 7–8 characterised by high LMA samples (Fig. 7C). LMA-related biomarkers were evenly dispersed on all chromosomes (Fig. 7D). Plotting their effect size (fold changes, Fig. 7E) outlined that most genome areas hosted both up- and downregulated biomarkers bar a few exceptions on chromosomes 4, 6, and 7 for all 3 genomes A, B, and D. Only upregulated biomarkers could be seen on chromosome 4A region 300–500 × 10^6 cM and chromosome 7A region 300–480 × 10^6 cM (replicated on genomes B and D). They matched 3 uncharacterised proteins, a 60S ribosomal protein L18a, a glucose-1-phosphate adenyltransferase, a polyadenylate-binding protein, a 14–3–3 protein, and a protein disulfide isomerase (Supplementary Table S5). Conversely, chromosome 6A region 300–410 × 10^6 cM (replicated on genomes B and D) exclusively located downregulated biomarkers matching a glyceraldehyde-3-phosphate dehydrogenase, a glutathione peroxidase, a tripeptidyl-peptidase II, and an uncharacterised protein. Charting biomarker correlation values with LMA as links failed to isolate stretches of genomic areas specific to LMA-responding proteins (Fig. 7I). This could be explained by the fact that LMA expression in our experiment elicited a complex metabolic response involving many gene products independent of their genomic position. LMA is indeed a multigenic trait; associated quantitative trait loci (QTLs) have been located across all 3 genomes and would contribute to the LMA phenotype in an independently effective and additive fashion [35].

Concluding remarks

For the first time, LMA phenotype was explored via proteomics. All the differentially regulated biological processes highlighted in this study by the various data mining means have been condensed into 1 summarising table and organised into both broad and specific functional categories (Supplementary Table S7).

In this work, we observed that grains displaying high alpha-amylase activities had an activated primary metabolism including glycolysis, gluconeogenesis, TCA cycle, mechanisms for DNA and RNA binding, and protein translation. Protein folding activities driven by chaperones and protein disulfide isomerase, as well as protein assembly via dimerisation and complexing, were also featured. Secondary metabolism was mobilised with the upregulation of phytohormones and chemical and defence responses. Furthermore, LMA invoked cellular structures involving ribosomes, microtubules, and chromatin. Finally, LMA expression significantly impacted grain starch and other carbohydrates and upregulated alpha-gliadins and starch metabolism, while downregulating LMW glutenin, stachyose, sucrose, UDP-galactose, and UDP-glucose. This work demonstrates that proteomics deserves to be part of the wheat LMA molecular toolkit and should be adopted by scientists and breeders in the future as part of accelerated testing programs to screen against this defect. More broadly, the workflow and strategies employed in the current work could be adapted to other traits and species as well as sustain proteogenomics endeavours.

Materials and Methods

Wheat cultivation, sampling, and storage

The wheat collection of 858 genotypes used in this study represents a diverse range of cultivars and germplasm sourced through the Australian Grains Genebank and representing worldwide genetic diversity. Wheat was grown in a single location in field trials at Horsham Victoria from 2012 to 2019 and harvested using a mechanical small-plot harvester.

The threshed grain was stored in sealed containers at 20°C. The environmental conditions (rain and temperature) at the trial site were monitored throughout the growing season. No preharvest rainfall was recorded ensuring that any alpha-amylase activity was nongerminative but associated with LMA.

The list of wheat samples is supplied in Supplementary Table S1.

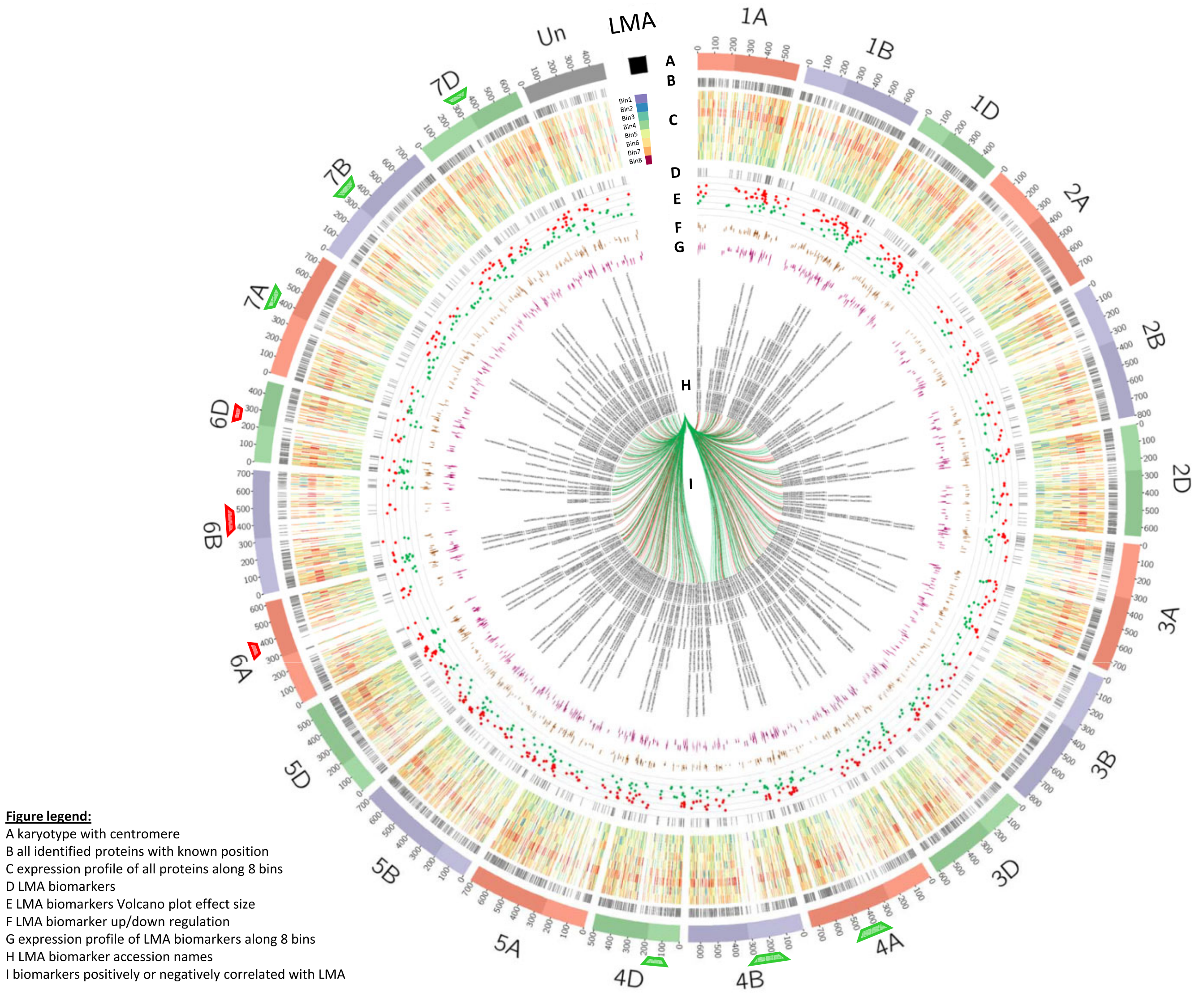


Figure 7: Circos plot of identified proteins and LMA-responsive biomarkers with expression patterns and statistics. (A) *T. aestivum* karyotype with chromosome length marked each 10^6 cM and centromeres indicated by the change in shade. LMA is displayed as a chromosome to portray the trait's 8-bin colour pattern in trace C. (B) Chromosomal positions of all identified proteins as highlights. (C) Profiling of all identified proteins along 8 bins as heatmaps. LMA pattern is provided as a reference. (D) Chromosomal positions of all identified LMA-responsive biomarkers as highlights. (E) Volcano plot effect size of biomarkers as scatterplot. Red denotes downregulation and green denotes upregulation. (F) Profiling of biomarkers along 2 bins as stacked histogram. (G) Profiling of biomarkers along 8 bins as stacked histogram. (H) Biomarker accession IDs as text labels. (I) Positive (green) and negative (red) correlation with LMA as links. Green and red tags under chromosomes 4ABD, 6ABD, and 7ABD denote genomic regions exclusive to biomarkers up- and downregulated, respectively.

LMA assay

The alpha-amylase assay was performed using the Megazyme assay according to the procedure reported by McCleary and Sheehan [148] on 3,773 grain samples (Supplementary Table S1) in parallel to the proteomics workflow.

The distribution of LMA values was plotted as a histogram in Microsoft Excel. Various transformations were performed to achieve a normal distribution such as standardisation, log natural, log 2, inverse, and standardisation of inversed values (data not shown). The transformed values were also plotted as histograms to check for Gaussian distribution.

Wheat grain processing for proteomics analyses

Sample preparation was optimised and thoroughly described [37]. Detailed hereafter are technical considerations essential in efficiently preparing such a large volume of samples. The overall workflow is schematised in Fig. 1. All sample packages were mixed for randomisation and assigned a unique number as they were processed. QR codes on sample bags and tubes were scanned and consigned to the Excel spreadsheet using a handheld barcode scanner (model 1902 GHD-2; Honeywell Australia). All microtubes were prelabelled with unique numbers and sample IDs, both also consigned to a QR code, using a handheld label maker (PT-E550WVP; Brother) controlled by the P-touch editor software (Brother) fitted with 12-mm white laminated tape.

The grains were ground and the QC control was made as specified in [37]. A 20 mg (±0.2 mg) aliquot of flour was used for protein extraction as described in [37].

Two vials of trypsin/Lys-C mix (100 µg, V5078; Promega) were dissolved into 1 mL of the resuspension buffer (50 mM acetic acid) supplied by the manufacturer and kept on ice until use to digest 192 wheat samples at a time. Aliquots of 10 µL protein extracts were transferred into two 96-well plates (Strata 96-well collection plate, 350 µL conical polypropylene; Phenomenex), diluted 6 times with 50 mM ammonium bicarbonate, and digested with 5-µL aliquots of the trypsin/Lys-C solution prepared earlier. Plates were sealed with silicone covers (pierceable sealing mats, 96-square well; Phenomenex) and vortexed for 30 seconds using a rack vortex mixer (MTV1 Multi Tube Vortex Mixer; Ratek) at high speed. Plates were incubated at 37°C for 17 hours. Volumes of 7 µL 10% formic acid (FA)/water were added to stop the digestion. An IS ([Glu1]-fibrinopeptide B human, F3261; Sigma) was added at a final concentration of 1 µg. Protein digests were cleaned, fully evaporated, and reconstituted as described [37].

LC-MS acquisition

All 4,061 wheat and QC samples were processed using the LC-MS method listed below.

Liquid chromatography (LC) was optimised [37]. Our chosen LC method applied a 0.2-mL/min flow rate, 38-minute LC run duration, 6% B for 2.5 minutes, 6–36% B gradient for 30.5 minutes, increased up to 98% B gradient for 0.1 minute, 98% B for 5 minutes, drop down to 3% B in 0.1 minutes, and 6% B for 5 minutes. The LC system and mobile phases were indicated in [37]. The rack types were specified as DeepWell96 in the LC-MS method and the SamplerModule tab of Xcalibur Direct Control software (version 3.0.63; ThermoFisher Scientific) with a 29,000-µm injection depth. Blanks (0.1% FA/water) and QC were injected from two 10-mL vials. Peptides were separated using an RP-LC column (bioZen 1.7 µm Peptide XB-C18,100 Å, LC column 150 × 2.1 mm; Phenomenex) using a 60°C oven temperature. The blank, IS, and QC samples were injected every 48 samples for normalisation purposes. The IS was used to check for mass accuracy (<50 ppm). The LC separation column was changed with a new one when peak resolution degraded (every 1,000 samples or so).

The UHPLC was online with an Orbitrap Velos hybrid ion trap–Orbitrap mass spectrometer (ThermoFisher Scientific) fitted with a heated electrospray ionisation (HESI) source. Every 3 weeks, the instrument was mass calibrated, and the source-sweeping cone and the heated capillary were cleaned. HESI parameters and FTMS spectra acquisition were described in [37].

The sequence lists were prepared in advance in Excel as .cvs files and imported into Xcalibur data acquisition software (version 3.0.63); 5 sequences were needed as Xcalibur only accommodated a maximum of 1,000 lines. Throughout the LC-MS run, the RAW files were individually visualised using the Xcalibur Qual Browser (version 3.0.63). Files that failed to pass our check (loss of peak resolution, incomplete run, no signal, mass accuracy >50 ppm, etc.) were rerun.

LC-MS/MS acquisition

For protein identification, 400 random samples (10% samples) were used following the LC-MS1 analysis. LC, HESI, and full-scan FTMS parameters were as indicated above. MS2 data were acquired using ITMS in positive mode as centroid values and applied various methods summarised below. To maximise the number of peptides sequenced, several passes were performed with inclusion and exclusion lists, with various parameters summarised in Supplementary File SF2.

Pass 1: The minimum signal threshold was 3,000 and the precursor isolation width was 2 *m/z*. No inclusion or exclusion list was used; however, a list of MS2 event was produced by exporting the “Scan Filters” of the RAW file in Xcalibur Qual Browser (ThermoFisher Scientific) and to be used in Pass 2 as an exclusion list containing 2,000 unique *m/z* values (maximum number allowed in Xcalibur). This method was run in duplicate.

Pass 2: Same method as Pass 1, except that the list of MS2 events generated in Pass 1 was uploaded in the Data Dependent Settings as a Reject Mass List. Like in Pass 1, a list of MS2 events was produced by exporting the “Scan Filters” of the RAW file and to be used in Pass 3 as an exclusion list containing 1,997 unique *m/z* values. This method was run in triplicate.

Pass 3: Same method as Pass 2, except that the list of MS2 events generated in Pass 2 was uploaded in the Data Dependent Settings as a Reject Mass List. Like in Pass 2, a list of MS2 events was produced by exporting the “Scan Filters” of the RAW file and to be used in Pass 4 as an exclusion list containing 1,998 unique *m/z* values. This method was run in duplicate.

Pass 4: Same method as Pass 3, except that the list of MS2 events generated in Pass 3 was uploaded in the Data Dependent Settings as a Reject Mass List. This was the last exclusion list used in this study. This method was run in duplicate.

Pass 5: Same method as Pass 1, except that the threshold was dropped to 500 to perform MS2 on peptides of low abundance. This method was run in duplicate.

Pass 6: Same method as Pass 1, except with a Parent Mass List (i.e., an inclusion list) made out of the 2,000 most abundant peptides. This method was run in duplicate.

For Passes 7–11, LC-MS1 reproducible peptides for which intensity exceeded 0.0001 (19,956 peptides in total) were randomised along with retention time (RT) and divided into 10 lists (inclusion lists 1 to 10 containing <2,000 *m/z* values each).

Pass 7: FTMS parameters were as specified above. Using the global MS/MSn method, MS/MS spectra were acquired in non-data-dependent mode. ITMS parameters were as in Pass 5. Inclusion list 1 was uploaded in the inclusion global MS/MS mass list tab of the Global Non-Data Dependent Settings. All remaining 9 parent lists were loaded to individual Pass 7 methods.

Pass 8: FTMS parameters were as specified above. ITMS parameters were as in Pass 5. Inclusion list 1 was uploaded in the parent mass list of the data-dependent settings. All remaining 9 parent lists were loaded to individual Pass 8 methods.

Pass 9: Same method as Pass 8, except that the precursor isolation width was 1 m/z to increase the mass accuracy the m/z values targeted in the parent mass list. All remaining nine parent lists were loaded to individual pass 9 methods.

Pass 10: Same method as Pass 8, except that the precursor isolation width was 0.5 *m/z* to further increase the mass accuracy of the *m/z* values targeted in the parent mass list. All remaining 9 parent lists were loaded to individual Pass 10 methods.

Pass 11: Same method as Pass 8, except that the precursor isolation width was 0.2 *m/z* to target the parent masses as accurately as possible. All remaining 9 parent lists were loaded to individual Pass 11 methods.

All the Xcalibur parameters of the various MS/MS methods can be found in Supplementary File SF1. Exclusion and inclusion lists can be found in Supplementary File SF2. A total of 63 LC-MS2 files were thus acquired; they are available from the MassIVE repository (MSV000090572 [149]).

LC-MS quantitation

The LC–MS RAW files of the 4,061 wheat samples along with the 86 QC and IS replicates (injected once every 48 wheat samples) were processed in the Refiner MS module of Genedata Expressionist 13.0 (Genedata AG). To process all files in 1 batch, a stepwise workflow was devised (Supplementary Fig. S1A, B).

In the first step, a repetition activity was used (processing 1 file at a time) in which the consecutive subactivities were performed: (i) load from file, (ii) RT structure removal with a minimum of 4 scans and *m/z* structure removal with a minimum of 8 points, (iii) chromatogram smoothing using a 3-scan RT window and a moving average estimator, (iv) RT structure removal with a minimum of 5 scans, and (v) save snapshot to export all the processed files individually. The files were individually checked for inconsistencies that would invalidate the subsequent quantitative analyses. Inadequate files were removed from the dataset, leaving 3,990 reproducible wheat files. In the second step (Supplementary Fig. S1C), the activities applied were: (i) load from file on the left for all the samples and on the right for the QCs; (ii) adaptative grid with 10-*m/z* scan counts; (iii) average across experiments (files) using the arithmetic mean; (iv) reference grid joining both sides; (v) chromatogram RT alignment applying a maximum RT shift of 50 scans (30 seconds); (vi) chromatogram peak detection using a 12-scan summation window, minimum peak size of 8 scans, maximum merge distance of 5 points and boundaries merge strategy, 10% gap/peak ratio for peak RT splitting, 3 points for *m/z* smoothing, ascent-based peak detection with 3-point isolation threshold, local maximum centre computation, and maximum curvature boundary determination; (vii) chromatogram isotope clustering with 0.1-minute RT tolerance and 20-ppm *m/z* tolerance, the peptide isotope shaping method with protonation ionisation, minimum charge of 2 and maximum charge of 10, maximum log-ratio distance of 0.8, and variable charge dependency for cluster size restriction; (viii) singleton filter; (ix) metadata import; (x) save snapshot; and (xi) export analyst of the clusters using the integrated maximum intensity.

LC-MS processed quantitative data and metadata (sample description, LMA measurements, sample preparation technical steps, LC-MS sequence, instrument maintenance, etc.) were exported into Genedata Analyst (version 13; Genedata AG) for normalisation purposes (Supplementary Fig. S1D). Data file normalisation with 3 consecutive steps was reported [37]. In brief, first, the quantities were normalised using the flour weights (1% accuracy) to account for sample preparation variation; second, the IS cluster was used to normalise peptide abundances to take into consideration postdigestion technical variation; and third, QCs and injection order were considered to correct instrument variation over time. The normalised quantitative data were exported as a CSV file for further processing. The CSV file contained 44,444 rows (peptide clusters) and 3,990 columns (wheat samples).

Correction of technical biases

The effects of technical biases on the LC-MS spectra were quantified using ANOVA simultaneous component analysis (ASCA), a generalisation of ANOVA that quantifies the variation induced by fixed experimental factors on complex multivariate datasets [150]. The normalised data were imported into R, where clusters containing 100% missing values were removed ($n = 12{,}108$), leaving 32,336 peptide clusters. The resulting dataset was a 3,990 × 32,336 matrix with each row being an individual sample and each column an LC-MS cluster. All remaining missing values were then imputed to a value zero. A separate metadata matrix (3,990 × 4), which contained information on the technical conditions in the LC-MS run for each sample, was compiled. These metadata were (i) LC separation column—categorical variable with 4 levels, (ii) mass calibration—categorical variable with 6 levels, and (iii) source-heated capillary—categorical variable with 2 levels. A total of 3,090 samples had complete data (LC-MS spectra and corresponding metadata). This complete dataset was then analysed using ASCA in MatLab v.R2017b (MathWorks) utilising the PLS Toolbox v. 8.5.2 (Eigenvector Research Inc.) to see which, if any, of the fixed experimental effects had a significant impact on the LC-MS cluster data. The statistical significance of the impact of each fixed experimental effect was estimated by calculating a *P* value from permutation testing with 100 iterations.

The impact of experimental factors with a significant effect on LC-MS cluster data was then accounted for by correcting the data using multiple linear regression in R [151] as described in [45]. The linear model was fitted as follows:

$$\mathrm{Y\ ijkl} = \mathrm{u} + \mathrm{Column\ (i)} + \mathrm{MassCal\ (j)} + \mathrm{Cap\ (k)} + \mathrm{eijk\ (l)}$$

where y is the signal intensity of a given cluster, u is the overall mean, Column is the ith LC column (4 levels), MassCal is the jth mass calibration (6 levels), Cap is kth source-heated capillary (2 levels), and eijkl is the random error term. The "corrected data" was a matrix of the residuals of the above model, which was run iteratively for each of the 32,336 peptide clusters. PCA plots were produced using R [151] and the gg2plot package.

Protein identification

The 63 RAW LC-MS2 files were processed in the Refiner MS module of Genedata Expressionist 13.0 using a stepwise workflow similar to the one described for LC-MS1 data, except for additional activities pertaining to protein database search (Supplementary Fig. S2A–C).

RAW files were searched using the Mascot program (version: 2.6.1; Matrix Science Ltd) within Genedata Refiner. The wheat database searched was retrieved from 3 independent sources. The first source was UniProtKB with 142,969 *T. aestivum* protein sequences (accessed on 26 February 2020, [37]). The second source was the EnsemblPlants repository hosting the *T. aestivum* genome initially sequenced by the IWGSC [7] and containing 143,241 Traes AA sequences. A contaminant database was also retrieved (common Repository of Adventitious Proteins, cRAP). All the FASTA files were combined and redundant sequences removed by following the GalaxyP tutorial "Protein FASTA Database Handling" [152, 153]. The decoy database was created by reversing all the sequences and appending them using the GalaxyP tool "DecoyDatabase." Our Galaxy workflow is available in Supplementary File SF1. The final FASTA file was imported and indexed in Mascot. It contained 286,482 protein sequences and 1,647,476,761 AA residues; its longest sequence bore 5,359 residues. It is available from the MassIVE repository (MSV000090572 [149]).

All MS2 files were searched in 1 batch using Mascot Daemon (version 2.6.1; Matrix Science Ltd) and the following parameters: MS/MS ions search; Mascot generic data format; ESI-TRAP instrument; trypsin enzyme; 9 maximum missed cleavages; carbamidomethyl (C) as fixed modification; guanidyl (K) and oxidation (M) as variable modifications; quantitation none; monoisotopic mass, 2+, 3+, and 4+ peptide charge; 10-ppm peptide tolerance; 0.5-Da MS/MS tolerance; and error-tolerant search (Supplementary Fig. S2D). Results were exported as .csv files into Excel.

The 32,336 peptide clusters from the corrected dataset produced by the LC-MS analyses were matched in R [151] (version 4.1.0-foss-2021a, Supplementary File SF1) to the 29,908 peptide clusters generated by the LC-MS/MS analyses using their respective RT, *m/z*, and mass values with ±0.1 accuracy and then linked to the Mascot identification results. The identification results of the peptide clusters whose RT shifted by more than 1 minute were not included.

Statistical analyses of proteomics data

Out of the 4,061 grains samples processed in this work, 3,990 yielded reproducible LC-MS data for 32,336 peptide clusters. The full quantitative data are available from the MassIVE repository (MSV000090572 [149]). The corrected dataset with Mascot identification results was imported into Genedata Analyst (version 13, Genedata AG). LMA measurements were obtained on 3,773 (out of 3,990) wheat samples. Whilst LMA trait characterised the wheat samples, we also wanted to analyse it along with the peptides to facilitate biomarker discovery. To this end, we used the inverse function to normally distribute the LMA values (Inv(LMA)) and transposed them as a row to incorporate them into the LC-MS dataset under the label "Cluster_AAA" along with all the other 32,336 peptides, thus bringing the total number of clusters to 32,337. This "Cluster_AAA" row was used in the subsequence statistical analyses to isolate peptides displaying profiles similar to that of LMA.

Principal component analysis

A PCA was performed on the full dataset (3,990 samples × 32,336 peptides) in R using the prcomp() function of the stats package. The eigenvalues were plotted using the screeplot() function.

Checking the distribution of LC-MS1 data

To redistribute data normally, the corrected dataset rows (peptides and Cluster_AAA) were z-transformed and plotted as a histogram in R. The hist() function was used to plot the corrected and z-transformed dataset as histograms in R. One-sample Kolmogorov–Smirnov tests were applied to check the normality of the distribution of both corrected and z-transformed datasets using the ks.test() function and "pnorm" argument in R. All the subsequent statistical analyses were performed on the z-transformed dataset.

Subsampling wheat samples to eliminate the bias towards low LMA values

LMA values spanned 0 to 8 U/g with the vast majority (95%) below 0.2 U/g (which corresponded to FN 300 seconds [14]); therefore, the LMA distribution was greatly skewed towards low LMA values. To eliminate this bias, a subset of wheat samples was selected as follows: all the samples bearing a LMA ≥0.17 were selected (467 samples in total) and an equivalent number of samples (467) with LMA <0.17 were randomly selected among the 3,306 remaining wheat samples. This subset of 934 wheat samples was no longer skewed towards low LMA values and is referred as "unbiased samples" hereafter.

PLS to subset LMA-responding peptides

In Genedata Analyst, a PLS 2D plot was created using the 934 unbiased samples and all the 32,346 peptides resolved in this study. The parameters were LMA as a response, 3 latent factors, 10% valid values, and row mean imputation. Both score and loading plots were exported along with the VIP scores. The higher the score, the greater the contribution of the peptide to the PLS and the closer to LMA response. These VIP scores were used to select meaningful subsets of peptides for the subsequent statistical analyses.

Univariate PLS regression to impute LMA missing values

The missing LMA values were predicted using a univariate PLS regression model in Genedata Analyst. First a model was developed using the 934 unbiased samples and 2,996 peptides with PLS high VIP scores (>1.5). Second, among the 934 wheat samples, 179 were randomly chosen so that LMA evenly spanned 0 to 5 and those LMA values were erased. Several PLSR models were tested to accurately predict erased LMA values (data not shown). The most accurate model applied the following parameters: LMA as a response, 20% valid values, and 20 latent factors. The model was then applied to the 217 missing LMA values against the 934 unbiased wheat samples.

SOM clustering

In Genedata Analyst, a SOM was created using the 934 unbiased samples and 7,254 peptides with VIP scores above 1 (including Cluster_AAA) and the following parameters: 6 rows, 8 columns, positive correlation distance, 50 maximum iterations, and 10% valid values.

K-means

In Genedata Analyst, a k-means was performed using the 934 unbiased samples and 7,254 peptides with VIP scores above 1 (including Cluster_AAA) and the following parameters: k = 20, positive correlation distance, mean centroid calculation, 10% valid values, and 50 maximum iterations.

Divisive HCA and agglomerative HCA

A divisive HCA was produced in Genedata Analyst using the 934 unbiased samples and 7,254 peptides with VIP scores above 1 (including Cluster_AAA) and the following parameters: clustering peptides, tree with tile plot, positive correlation distance, Ward linkage, 10% valid values, k-means cluster profile, and split by size. The outcome of this analysis enabled us to sort the peptides based on their accumulation patterns in wheat samples.

Still in Genedata Analyst, we also performed an agglomerative HCA using the all the 934 unbiased samples and 532 LMA-related biomarkers (including Cluster_AAA) and the following parameters: clustering samples, tree, positive correlation distance, Ward linkage, and 50% valid values. The outcome of this analysis allowed us to sort the grain samples according to their LC-MS molecular similarity, which was then exploited in a heatmap.

Correlation

An annotation correlation was performed in Genedata Analyst using the full dataset including Cluster_AAA (3,990 samples × 32,337 peptides) against standardised LMA values. This produced R squared (R^2) values.

Simple linear mixed regression

The full dataset including Cluster_AAA (3,990 samples × 32,337 peptides) was used to run a linear regression in Genedata Analyst with 1 explanatory variable using the following model: $y = \text{Inv(LMA)} + \varepsilon$, in which Inv(LMA) is the normal inverse function of LMA measurements and e the error. The false discovery rates were computed according to the Benjamini–Hochberg estimates as q values.

Peptide expression profiles along 2 or 8 LMA bins

Our data matrix of 3,990 columns by 32,337 rows contained 129,024,630 quantities, which posed representation challenges. We adopted a data reduction strategy involving binning the samples into 8 or 2 arbitrary bins based on their LMA values to produce simpler, more legible graphs for individual peptide profiling.

In the first instance, we sorted all 3,990 wheat samples based on an increasing order of LMA values and then split them into 8 arbitrary bins of 499 samples each. The last bin (0.17132 < LMA < 7.95442) contained all the 266 unsound grains (LMA >0.2).

In the second instance and using the 934 unbiased wheat samples, we created 2 bins based on LMA value threshold of 0.17. The bin containing 467 samples with LMA <0.17 only comprised sound grains. All the 266 unsound grains (LMA >0.2) were comprised in the bin containing 467 samples with LMA ≥0.17.

The peptide quantities were then averaged per bin to produce mean expression profiles along 2 or 8 bins.

T test with effect size and volcano plot

Using the unbiased biomarker dataset (934 samples × 532 peptides including Cluster_AAA), a *t* test was performed with the LMA threshold of 0.17 as a factor and the following parameters: bootstrap with 10 repeats and balanced permutations, effect size based on group means, and 90% valid values. A volcano plot was produced by plotting the effect size against *P* values.

Data mining

The LC-MS2 experiments followed by Mascot search produced identification results for 5,414 peptide clusters, which matched 8,044 protein accessions. These identification results were mined using the databases and tools described below. Resulting outputs were consigned to Supplementary Table S3.

UniProt database and gene ontology (GO)

The list of 8,044 UniProt accessions identified in this study was uploaded in the Retrieve/ID mapping tool of UniProt (accessed in May 2022) [154] to retrieve protein descriptions, FASTA sequences, GO terms, and TRAES accession IDs. Out of the 8,044 UniProt accessions, 5,960 UniProt accessions corresponded to 6,622 TRAES accessions. TRAES accessions were needed to interrogate ShinyGO and BreadwheatCyc databases (described below).

KEGG database and pathway maps

The 8,044 FASTA sequences were uploaded into the Assign KO tool (accessed in May 2022) [155] by specifying the Poaceae family to retrieve KO identifiers. KO identifiers were then mapped using the KEGG Mapper Reconstruct tool (accessed in May 2022) to list pathways, brites, and modules involving identified proteins.

ShinyGO, functional category enrichment, and chromosomal positions

The list of 6,622 TRAES accessions was uploaded into ShinyGO [134] to generate functional category enrichments, dot plots, tree, and networks, as well as retrieve chromosomal positions. Positions were obtained for 4,571 TRAES accessions, which were used in Circos plots (detailed below).

Pathway tools, BreadwheatCyc, and perturbed pathways

The list of 6,622 TRAES accessions along with quantitative data along 8 bins was uploaded into the Pathway Tools software [136] and run online via the BreadwheatCyc database ([156] accessed in June 2022) via the Plant Metabolic Network server [137] using the Omics Dashboard and the Cellular Overview tools to generate Pathway Perturbation Scores (PPS).

The Chrome extension Veed.io was used to create a film capturing the Cellular Overview animation (Supplementary Video SV1).

Circos and chromosomal position

The 4,571 TRAES accessions whose chromosomal positions were known from ShinyGO were charted along a Circos plot invented by Krzywinski and colleagues [146] and recently wrapped in the Galaxy platform by Rasche and colleagues [147, 152, 157]. The details of the various layers are indicated in the figure's legend.

Converting wide to long tables in R and charting using power BI desktop

Most identified peptides matched several UniProt accessions, which corresponded to several TRAES IDs and GO terms. This produced wide tables. In R [151], wide tables were converted to long tables using the pivot_longer() function from the tidyr package. Long tables were merged using the merge() function of the R base package using peptide Cluster IDs as unique references.

Wheat sample metadata, peptide metadata, and quantitative dataset and identities for the biomarkers were imported into Microsoft Power BI Desktop (Version: 2.106.883.0 64-bit June 2022) and linked via the Clusters names to produce dashboards using multiple visuals (word clouds, tree maps, histograms, scatterplots, waterfall plots, pie charts, violin plots, and ribbon charts).

Abbreviations

ABA: abscisic acid; ACN: acetonitrile; AA: amino acid; AMY: amylase; ANOVA: analysis of variance; ASCA: ANOVA simultaneous component analysis; BP: biological process; CC: cellular component; cM: centimorgan; CID: collision-induced dissociation; CSV: comma separated value; cRAP: common Repository of Adventitious Proteins; DPA: day post anthesis; DNA: deoxyribonucleic acid; DPPS: differentially perturbed pathways; TaCSP: ent-copalyl disphosphate synthase from Triticum aestivum; ELISA: enzyme-linked immunosorbent assay; FN: falling number; FA: formic acid; FDR: false discovery rate; FTMS: Fourier transform orbitrap mass analyser; GO: gene ontology; GxE: genetic by environment interaction; GA: gibberellic acid; Gnd-HCl: guanidine hydrochloric acid; HESI: heated electrospray ionisation; HCA: hierarchical clustering analysis; HMW: high molecular weight; HPLC: high performance liquid chromatography; ID: identity; IS: internal standard; IWGSC: International Wheat Genome Sequencing Consortium; ITMS: ion trap orbitrap mass analyser; pI: isoelectric point; IPA: isopropanol; KO: KEGG orthology; kD: kiloDalton; KNN: K-Nearest Neighbours; K-S: Kolmogorov-Smirnov; KEGG: Kyoto Encyclopedia of Genes and Genomes; LMA: late maturity alpha-amylase; LC: liquid chromatography; LMW: low molecular weight; MS or MS1: mass spectrometry; m/z: mass to charge ratio; mRNA: messenger ribonucleic acid; MF: molecular function; MLR: multivariate linear regression; ppm: part per million; PLS: partial least squares; PLSR: partial least squares regression; PTM: post-translational modification; PC: principal component; PCA: principal component analysis; QC: quality control; QTL: quantitative trait locus; QR code: quick response code; RT: retention time; Rab: Responsive to abscisic acid; RO: reverse osmosis; RT-qPCR: reverse transcription quantitative real-time polymerase chain reaction; SOM: self-organising map; SPE: solid phase extraction; MS/MS or MS2: tandem mass spectrometry; 3-D: three-dimensional; TCA: trichloroacetic acid; T. aestivum: Triticum aestivum (common bread wheat); TRAES: Triticum aestivum accession; 2-DE: two-dimensional electrophoresis; 2-D: two-dimensional; UTR: untranslated region; UDP: uridine diphosphate; VIP: variable importance in projection.

Potential Implications

Since proteome evidence is confirmation that the gene is translated to a protein, our work can validate wheat genome annotation. Peptides identified in this study can be mapped against the wheat genome using a proteogenomics strategy. This will confirm the expression at the protein level of not only "high confidence" but also "low confidence" gene models.

Data Availability

The LC-MS1 dataset and raw LC-MS2 data generated and analysed during the current study are available in the MassIVE repository, accession nr. MSV000090572 [149]. All data generated or analysed during this study are included in this published article and its supplementary information files. Supporting data [including Supplementary Fig.s S1-S14, Supplementary Tables S1-S7, Supplementary Video SV1, and Supplementary Files SF1-SF2] is available via the GigaScience repository, GigaDB [158].

Additional Files

Supplementary Fig. S1: Genedata Refiner workflow to process all wheat, IS and QC LCMS1 RAW files and export them to Genedata Analyst.

Supplementary Fig. S2: Genedata Refiner workflow to process all wheat LCMS2 RAW files and export them as quantities to Excel spreadsheet.

Supplementary Fig. S3: LC-MS2 RAW maps for each tandem pass.

Supplementary Fig. S4: Histogram of the number of peptides identified using Mascot algorithm and number of MS2 events in each of the LC-MS2 file.

Supplementary Fig. S5: Histograms and box plots of the number of peptides per accession and number of accessions per peptides.

Supplementary Fig. S6: Distribution of LC-MS1 data across 3,990 wheat samples and 32,336 quantified peptides.

Supplementary Fig. S7: Partial Least Square (PLS) using LMA as a response on the unbiased samples and the unbiased samples and all the quantified peptides.

Supplementary Fig. S8: Partial least square regression (PLSR) to impute LMA missing values.

Supplementary Fig. S9: Binning strategies of wheat samples based on LMA measurements.

Supplementary Fig. S10: Mining identified proteins using Power BI.

Supplementary Fig. S11: Retrieval of protein descriptions and Gene Ontology (GO) terms for Molecular Function (MF), Cellular Component (CC), and Biological Process (BP) from UniProtKB using all 8,044 protein identities.

Supplementary Fig. S12: KEGG output using all 8,044 identified proteins matching 677 KOs.

Supplementary Fig. S13: ShinyGO outputs using all 6,622 TRAES accessions corresponding to the 8,044 UniProt proteins.

Supplementary Fig. S14: Pathway Tools output using 6622 TRAES accessions and quantitative data averaged along 8 bins.

Supplementary Table S1: List wheat samples and LMA measurements.

Supplementary Table S2: List of error-tolerant modifications found by Mascot in the wheat grain proteome.

Supplementary Table S3: List of all wheat grain 5,414 peptides identified by LC-MS2 and matching 8,092 Uniprot protein accessions.

Supplementary Table S4: LC-MS features and statistical results for the 532 AAA-responsive candidate peptides.

Supplementary Table S5: Data mining annotations of the 390 AAA-responsive biomarkers identified by LC-MS2.

Supplementary Table S6: ShinyGO chromosomal positions for all identified proteins.

Supplementary Table S7: List of mechanisms involved in LMA response.

Supplementary Video SV1: Cellular Overview animation of wheat proteome using BreadwheatCyc.

Supplementary File SF1: PDF file including more method details on the high-throughput proteomics workflow of this study, some background information on statistical analyses for big data, all the MS/MS methods parameters, the Galaxy workflow used to create our wheat protein database, and the R script to link LCMS1 and LCMS2 clusters.

Supplementary File SF2: Excel file containing various tabs including a summary of MS parameters across all 11 passes, as well as all the reject, parent and inclusion mass lists used in the ITMS2 methods.

Competing Interests

The authors declare that they have no competing interests.

Funding

This research was funded by the Grains Research and Development Corporation (GRDC), Project DJP2001-008RTX.

Authors' Contributions

Conceptualisation, M.H., H.D., J.P., D.V.; plant materials: J.P., LMA assays: N.R.; grain grinding: D.V., A.B., D.R.; sample processing, D.V., A.B.; LC-MS maintenance: D.V. and V.E.; LC-MS data acquisition: D.V., and A.B.; LC-MS1 and LC-MS2 data acquisition and analysis: D.V.; LC-MS1 matching with LC-MS2 and machine learning in R: S.S.; technical bias removal: T.L.; statistical analyses: D.V. and S.R.; data mining and figures, D.V.; investigation, D.V.; resources, S.R.; data curation, D.V.; submission to MassIVE, D.V.; writing—original draft preparation, D.V.; review and editing, D.V., T.L., J.P., S.R., and H.D.; manuscript submission and revisions; D.V., visualization, D.V.; logistics: D.V.; supervision, S.R.; project administration, D.V., S.R., H.D., and M.H; funding acquisition, M.H. and H.D. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

We thank Mr Pankaj Maharjan for retrieving all wheat samples from storage. We are grateful for advice on MS/MS targeted methods from Drs Aaron Elkins, Priyanka Reddy from AVR, and Dr Enzo Huang from Thermo Scientific. We are grateful to Carl Thomas and Piotr Malicki from AVR for upgrading Genedata and Mascot servers, as well as maintaining the Bioinformatics Advanced Scientific Computing cluster. We thank Dr Gabriel Keeble-Gagnere from AVR for his critical review of the manuscript.

References

1. Hussain B, Akpinar BA, Alaux M, et al. Capturing wheat phenotypes at the genome level. Front Plant Sci 2022;13:851079. https://doi.org/10.3389/fpls.2022.851079.

2. Bacala R, Hatcher DW, Perreault H, et al. Challenges and opportunities for proteomics and the improvement of bread wheat quality. J Plant Physiol 2022;275:153743. https://doi.org/10.1016/j.jplph.2022.153743.
3. Shewry PR. Do ancient types of wheat have health benefits compared with modern bread wheat? J Cereal Sci 2018;79:469–76. https://doi.org/10.1016/j.jcs.2017.11.010.
4. de Sousa T, Ribeiro M, Sabenca C, et al. The 10,000-year success story of wheat!. Foods 2021;10(9):2124. https://doi.org/10.3390/foods10092124.
5. Shewry PR. Wheat. J Exp Bot 2009;60(6):1537–53. https://doi.org/10.1093/jxb/erp058.
6. Venske E, Dos Santos RS, Busanello C, et al. Bread wheat: a role model for plant domestication and breeding. Hereditas 2019;156:16. https://doi.org/10.1186/s41065-019-0093-9.
7. International Wheat Genome Sequencing C, investigators IRp, Appels R, Eversole K, Feuillet C, et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 2018;361:6403. https://doi.org/10.1126/science.aar7191.
8. Guan J, Garcia DF, Zhou Y, et al. The battle to sequence the bread wheat genome: a tale of the three kingdoms. Genomics Proteomics Bioinformatics 2020;18(3):221–9. https://doi.org/10.1016/j.gpb.2019.09.005.
9. International Wheat Genome Sequencing C. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 2014;345(6194):1251788. https://doi.org/10.1126/science.1251788.
10. Zhu T, Wang L, Rimbert H, et al. Optical maps refine the bread wheat triticum aestivum cv. Chinese spring genome assembly. Plant J 2021;107(1):303–14. https://doi.org/10.1111/tpj.15289.
11. Henry RJ, Furtado A and Rangan P. Wheat seed transcriptome reveals genes controlling key traits for human preference and crop adaptation. Curr Opin Plant Biol 2018;45(Pt B):231–6. https://doi.org/10.1016/j.pbi.2018.05.002.
12. Hagberg S. A rapid method for determining alpha-amylase activity. Cereal Chem 1960;37:218–22.
13. Hu Y, Sjoberg SM, Chen CJ, et al. As the number falls, alternatives to the Hagberg-Perten falling number method: a review. Compr Rev Food Sci Food Saf 2022;21(3):2105–17. https://doi.org/10.1111/1541-4337.12959.
14. Steber CM. Avoiding problems in wheat with low falling numbers. Crops & Soils 2017;50(2):22. doi:.
15. Newberry M, Zwart AB, Whan A, et al. Does late maturity alpha-amylase impact wheat baking quality? Front Plant Sci 2018;9:1356. https://doi.org/10.3389/fpls.2018.01356.
16. Neoh GKS, Dieters MJ, Tao K, et al. Late-maturity alpha-amylase in wheat (Triticum aestivum) and its impact on fresh white sauce qualities. Foods 2021;10(2):201. https://doi.org/10.3390/foods10020201.
17. Sjoberg SM, Carter AH, Steber CM, et al. Unraveling complex traits in wheat: approaches for analyzing genotype × environment interactions in a multienvironment study of falling numbers. Crop Sci 2020;60(6):3013–26. https://doi.org/10.1002/csc2.20133.
18. Barrero JM, Mrva K, Talbot MJ, et al. Genetic, hormonal, and physiological analysis of late maturity alpha-amylase in wheat. Plant Physiol 2013;161(3):1265–77. https://doi.org/10.1104/pp.112.209502.
19. Derkx AP, Mares DJ. Late-maturity alpha-amylase expression in wheat is influenced by genotype, temperature and stage of grain development. Planta 2020;251(2):51. https://doi.org/10.1007/s00425-020-03341-1.
20. Mares DJ, Mrva K. Wheat grain preharvest sprouting and late maturity alpha-amylase. Planta 2014;240(6):1167–78. https://doi.org/10.1007/s00425-014-2172-5.
21. Mrva K, Wallwork M, Mares DJ. alpha-amylase and programmed cell death in aleurone of ripening wheat grains. J Exp Bot 2006;57(4):877–85. https://doi.org/10.1093/jxb/erj072.
22. Ainsworth CC, Doherty P, Edwards KG, et al. Allelic variation at alpha-amylase loci in hexaploid wheat. Theor Appl Genet 1985;70(4):400–6. https://doi.org/10.1007/BF00273745.
23. Mrva K, Mares D. Late-maturity alpha-amylase: low falling number in wheat in the absence of preharvest sprouting. J Cereal Sci 2008;47:6–17. https://doi.org/10.1016/j.jcs.2007.01.005.
24. Gale MD, Law CN, Chojecki AJ, et al. Genetic control of alpha-amylase production in wheat. Theor Appl Genet 1983;64(4):309–16. https://doi.org/10.1007/BF00274170.
25. Baulcombe DC, Huttly AK, Martienssen RA, et al. A novel wheat alpha-amylase gene (alpha-Amy3). Mol Gen Genet 1987;209(1):33–40. https://doi.org/10.1007/BF00329833.
26. Whan A, Dielen AS, Mieog J, et al. Engineering alpha-amylase levels in wheat grain suggests a highly sophisticated level of carbohydrate regulation during development. J Exp Bot 2014;65(18):5443–57. https://doi.org/10.1093/jxb/eru299.
27. Mieog JC, Janeček S, Ral JF. New insight in cereal starch degradation: identification and structural characterization of four α-amylases in bread wheat. Amylase 2017;1:35–49. https://doi.org/10.1515/amylase-2017-0004.
28. Ral JP, Whan A, Larroque O, et al. Engineering high alpha-amylase levels in wheat grain lowers falling number but improves baking properties. Plant Biotechnol J 2016;14(1):364–76. https://doi.org/10.1111/pbi.12390.
29. Ral JF, Sun M, Mathy A, et al. A biotechnological approach to directly assess the impact of elevated endogenous a-amylase on Asian white-salted noodle quality. Starch/Stärke 2018;70(1700089):1–10. https://doi.org/10.1002/star.201700089.
30. Cockburn D, Nielsen MM, Christiansen C, et al. Surface binding sites in amylase have distinct roles in recognition of starch structure motifs and degradation. Int J Biol Macromol 2015;75:338–45. https://doi.org/10.1016/j.ijbiomac.2015.01.054.
31. Verity JCK, Hac L, Skerritt JH. Development of a field enzyme-linked immunosorbent assay (ELISA) for detection of α-amylase in preharvest-sprouted wheat. Cereal Chem 1999;76(5):673–81. https://doi.org/10.1094/CCHEM.1999.76.5.673.
32. Mieog JC, Howitt CA, Ral JP. Fast-tracking development of homozygous transgenic cereal lines using a simple and highly flexible real-time PCR assay. BMC Plant Biol 2013;13:71. https://doi.org/10.1186/1471-2229-13-71.
33. McCleary BV. Measurement of polysaccharide degrading enzymes using chromogenic and colorimetric substrates. Chem Aust 1991;58:398–401.
34. McCleary BV, McNally M, Monaghan D, et al. Measurement of alpha-amylase activity in white wheat flour, milled malt, and microbial enzyme preparations, using the Ceralpha assay: collaborative study. J AOAC Int 2002;85(5):1096–102. https://doi.org/10.1093/jaoac/85.5.1096.
35. Cannon AE, Marston EJ, Kiszonas AM, et al. Late-maturity alpha-amylase (LMA): exploring the underlying mechanisms and end-use quality effects in wheat. Planta 2021;255(1):2. https://doi.org/10.1007/s00425-021-03749-3.
36. Mares D, Derkx A, Cheong J, et al. Gibberellins in developing wheat grains and their relationship to late maturity alpha-amylase (LMA). Planta 2022;255(6):119. https://doi.org/10.1007/s00425-022-03899-y.

37. Vincent D, Bui A, Ram D, et al. Mining the wheat grain proteome. Int J Mol Sci 2022;23(2):713. https://doi.org/10.3390/ijms23020713.

38. He M, Wang J, Herold S, et al. A rapid and universal workflow for label-free-quantitation-based proteomic and phosphoproteomic studies in cereals. Curr Protoc 2022;2(6):e425. https://doi.org/10.1002/cpz1.425.

39. Wu Y, Li L. Sample normalization methods in quantitative metabolomics. J Chromatogr A 2016;1430:80–95. https://doi.org/10.1016/j.chroma.2015.12.007.

40. Li H, Han J, Pan J, et al. Current trends in quantitative proteomics – an update. J Mass Spectrom 2017;52(5):319–41. https://doi.org/10.1002/jms.3932.

41. O'Rourke MB, Town SEL, Dalla PV, et al. What is normalization? The strategies employed in top-down and bottom-up proteome analysis workflows. Proteomes 2019;7(3):29. https://doi.org/10.3390/proteomes7030029.

42. Mitra V, Smilde AK, Bischoff R, et al. Tutorial: correction of shifts in single-stage LC-MS(/MS) data. Anal Chim Acta 2018;999:37–53. https://doi.org/10.1016/j.aca.2017.09.039.

43. Mizuno H, Ueda K, Kobayashi Y, et al. The great importance of normalization of LC-MS data for highly-accurate non-targeted metabolomics. Biomed Chromatogr 2017;31(1):e3864. https://doi.org/10.1002/bmc.3864.

44. Poulos RC, Hains PG, Shah R, et al. Strategies to enable large-scale proteomics for reproducible research. Nat Commun 2020;11(1):3793. https://doi.org/10.1038/s41467-020-17641-3.

45. Luke TDW, Pryce JE, Elkins AC, et al. Use of large and diverse datasets for (1)H NMR serum metabolic profiling of early lactation dairy cows. Metabolites 2020;10(5):180. https://doi.org/10.3390/metabo10050180.

46. Mrode RA. Linear Models for the Prediction of Animal Breeding Values. 3rd ed. Wallingford, UK 2014;

47. Lin H, Li M. Introduction to data science. bookdown. 2021; https://scientistcafe.com/ids/index.html

48. Calderon-Celis F, Encinar JR, Sanz-Medel A. Standardization approaches in absolute quantitative proteomics with mass spectrometry. Mass Spectrom Rev 2018;37(6):715–37. https://doi.org/10.1002/mas.21542.

49. Geyer PE, Voytik E, Treit PV, et al. Plasma proteome profiling to detect and avoid sample-related biases in biomarker studies. EMBO Mol Med 2019;11(11):e10427. https://doi.org/10.15252/emmm.201910427.

50. Elias JE, Haas W, Faherty BK, et al. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. Nat Methods 2005;2(9):667–75. https://doi.org/10.1038/nmeth785.

51. Wang G, Wu WW, Zhang Z, et al. Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. Anal Chem 2009;81(1):146–59. https://doi.org/10.1021/ac801664q.

52. Chen Y, Wang Y, Yang J, et al. Exploring the diversity of plant proteome. J Integr Plant Biol 2021;63(7):1197–210. https://doi.org/10.1111/jipb.13087.

53. Min CW, Gupta R, Agrawal GK, et al. Concepts and strategies of soybean seed proteomics using the shotgun proteomics approach. Expert Rev Proteomics 2019;16(9):795–804. https://doi.org/10.1080/14789450.2019.1654860.

54. Adhikari S, Nice EC, Deutsch EW, et al. A high-stringency blueprint of the human proteome. Nat Commun 2020;11(1):5301. https://doi.org/10.1038/s41467-020-19045-9.

55. Burkhart JM, Schumbrutzki C, Wortelkamp S, et al. Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. J Proteomics 2012;75(4):1454–62. https://doi.org/10.1016/j.jprot.2011.11.016.

56. Savitski MM, Kjeldsen F, Nielsen ML, et al. Relative specificities of water and ammonia losses from backbone fragments in collision-activated dissociation. J Proteome Res 2007;6(7):2669–73. https://doi.org/10.1021/pr070121z.

57. Sun S, Yu C, Qiao Y, et al. Deriving the probabilities of water loss and ammonia loss for amino acids from tandem mass spectra. J Proteome Res 2008;7(1):202–8. https://doi.org/10.1021/pr070479v.

58. Yang Y. Intramolecular cyclization side reactions. In: Yang Y, ed. Side Reactions in Peptide Synthesis. Academic Press; 2016; p. 119–61.

59. Ghatak A, Chaturvedi P, Weckwerth W. Cereal crop proteomics: systemic analysis of crop drought stress responses towards marker-assisted selection breeding. Front Plant Sci 2017;8:757. https://doi.org/10.3389/fpls.2017.00757.

60. Kerr ED, Caboche CH, Pegg CL, et al. The post-translational modification landscape of commercial beers. Sci Rep 2021;11(1):15890. https://doi.org/10.1038/s41598-021-95036-0.

61. Gao F, Ayele BT. Functional genomics of seed dormancy in wheat: advances and prospects. Front Plant Sci 2014;5:458. https://doi.org/10.3389/fpls.2014.00458.

62. Komatsu S, Kamal AH, Hossain Z. Wheat proteomics: proteome modulation and abiotic stress acclimation. Front Plant Sci 2014;5:684. https://doi.org/10.3389/fpls.2014.00684.

63. Adegoke TV, Wang Y, Chen L, et al. Posttranslational modification of waxy to genetically improve starch quality in rice grain. Int J Mol Sci 2021;22(9):4845. https://doi.org/10.3390/ijms22094845.

64. Zhou C, Dong Z, Zhang T, et al. Genome-scale analysis of homologous genes among subgenomes of bread wheat (Triticum aestivum L.). Int J Mol Sci 2020;21(8):3015. https://doi.org/10.3390/ijms21083015.

65. Cao H, Duncan O, Islam S, et al. Increased wheat protein content via introgression of an HMW glutenin selectively reshapes the grain proteome. Mol Cell Proteomics 2021;20:100097. https://doi.org/10.1016/j.mcpro.2021.100097.

66. Di Francesco A, Saletti R, Cunsolo V, et al. Qualitative proteomic comparison of metabolic and CM-like protein fractions in old and modern wheat Italian genotypes by a shotgun approach. J Proteomics 2020;211:103530. https://doi.org/10.1016/j.jprot.2019.103530.

67. Maignan V, Bernay B, Geliot P, et al. Biostimulant impacts of Glutacetine(R) and derived formulations (VNT1 and VNT4) on the bread wheat grain proteome. J Proteomics 2021;244:104265. https://doi.org/10.1016/j.jprot.2021.104265.

68. Dimitrova DS, Kaishev VK, Tan S. Computing the Kolmogorov-Smirnov distribution when the underlying CDF is purely discrete, mixed, or continuous. J Stat Softw 2020;95(10):1–42. https://doi.org/10.18637/jss.v095.i10.

69. Lazariv T, Lehmann C. Goodness-of-Fit Tests for Large Datasets. arXiv. 2018; arXiv:1810.09753v1.

70. Banerjee P, Ghosh S, Dutta M, et al. Identification of key contributory factors responsible for vascular dysfunction in idiopathic recurrent spontaneous miscarriage. PLoS One 2013;8(11):e80940. https://doi.org/10.1371/journal.pone.0080940.

71. Rasul G, Glover KD, Krishnan PG, et al. Genetic analyses using GGE model and a mixed linear model approach, and stability analyses using AMMI bi-plot for late-maturity alpha-amylase

activity in bread wheat genotypes. Genetica 2017;145(3):259–68. https://doi.org/10.1007/s10709-017-9962-1.

72. Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17(6):520–5. https://doi.org/10.1093/bioinformatics/17.6.520.

73. Horton NJ, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. The American Statistician 2001;55(3):244–54. https://doi.org/10.1198/000313001317098266.

74. Dixon JK. Pattern recognition with partly missing data. IEEE Trans Syst Man Cybern 1979;9(10):617–21. https://doi.org/10.1109/TSMC.1979.4310090.

75. Wold H. Estimation of principal components and related models by iterative least squares. In: Krishnajah PR, ed. Multivariate Analysis. New York: Academic Press; 1966; p. 391–420.

76. Nguyen DV, Rocke DM. On partial least squares dimension reduction for microarray-based classification: a simulation study. Computational Statistics & Data Analysis 2004;46(3):407–524. https://doi.org/10.1016/j.csda.2003.08.001.

77. Oleszko A, Hartwich J, Wojtowicz A, et al. Comparison of FTIR-ATR and raman spectroscopy in determination of VLDL triglycerides in blood serum with PLS regression. Spectrochim Acta A Mol Biomol Spectrosc 2017;183:239–46. https://doi.org/10.1016/j.saa.2017.04.020.

78. Nengsih TA, Bertrand F, Maumy-Bertrand M, et al. Determining the number of components in PLS regression on incomplete data set. Stat Appl Genet Mol Biol 2019;18(6). https://doi.org/10.1515/sagmb-2018-0059.

79. Sherlock G. Analysis of large-scale gene expression data. Curr Opin Immunol 2000;12(2):201–5. https://doi.org/10.1016/S0952-7915(99)00074-6.

80. Wang K, Wang W, Li M. A brief procedure for big data analysis of gene expression. Animal Model Exp Med 2018;1(3):189–93. https://doi.org/10.1002/ame2.12028.

81. Cresta Morgado P, Carusso M, Alonso Alemany L, et al. Practical foundations of machine learning for addiction research. Part I. Methods and techniques. Am J Drug Alcohol Abuse 2022;48(3):260–71. https://doi.org/10.1080/00952990.2021.1995739.

82. Kohonen T. Essentials of the self-organizing map. Neural Netw 2013;37:52–65. https://doi.org/10.1016/j.neunet.2012.09.018.

83. Liu Z, Dai S, Bones J, et al. A quantitative proteomic analysis of cellular responses to high glucose media in Chinese hamster ovary cells. Biotechnol Prog 2015;31(4):1026–38. https://doi.org/10.1002/btpr.2090.

84. Fankhauser N, Maser P. Identification of GPI anchor attachment signals by a Kohonen self-organizing map. Bioinformatics 2005;21(9):1846–52. https://doi.org/10.1093/bioinformatics/bti299.

85. Yu D, Shen H, Yang J. SOMRuler: a novel interpretable transmembrane helices predictor. IEEE Trans Nanobioscience 2011;10(2):121–9. https://doi.org/10.1109/TNB.2011.2160730.

86. Fraccalvieri D, Tiberti M, Pandini A, et al. Functional annotation of the mesophilic-like character of mutants in a cold-adapted enzyme by self-organising map analysis of their molecular dynamics. Mol Biosyst 2012;8(10):2680–91. https://doi.org/10.1039/c2mb25192b.

87. Madani S, Faez K, Aminghafari M. Identifying similar functional modules by a new hybrid spectral clustering method. IET Syst Biol 2012;6(5):175–86. https://doi.org/10.1049/iet-syb.2010.0066.

88. Tu M, Wang W, Yao N, et al. The transcriptional dynamics during de novo shoot organogenesis of Ma bamboo (Dendrocalamus latiflorus Munro): implication of the contributions of the abiotic stress response in this process. Plant J 2021;107(5):1513–32. https://doi.org/10.1111/tpj.15398.

89. Bednarz H, Roloff N, Niehaus K. Mass spectrometry imaging of the spatial and temporal localization of alkaloids in nightshades. J Agric Food Chem 2019;67(49):13470–7. https://doi.org/10.1021/acs.jafc.9b01155.

90. Wang L, Sun X, Weiszmann J, et al. System-level and Granger network analysis of integrated proteomic and metabolomic dynamics identifies key points of grape berry development at the interface of primary and secondary metabolism. Front Plant Sci 2017;8:1066. https://doi.org/10.3389/fpls.2017.01066.

91. Yu T, Li G, Dong S, et al. Proteomic analysis of maize grain development using iTRAQ reveals temporal programs of diverse metabolic processes. BMC Plant Biol 2016;16(1):241. https://doi.org/10.1186/s12870-016-0878-1.

92. Eisen MB, Spellman PT, Brown PO, et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95(25):14863–8. https://doi.org/10.1073/pnas.95.25.14863.

93. Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A 1999;96(12):6745–50. https://doi.org/10.1073/pnas.96.12.6745.

94. Duncan O, Trosch J, Fenske R, et al. Resource: mapping the triticum aestivum proteome. Plant J 2017;89(3):601–16. https://doi.org/10.1111/tpj.13402.

95. Fercha A, Capriotti AL, Caruso G, et al. Comparative analysis of metabolic proteome variation in ascorbate-primed and unprimed wheat seeds during germination under salt stress. J Proteomics 2014;108:238–57. https://doi.org/10.1016/j.jprot.2014.04.040.

96. Ma C, Zhou J, Chen G, et al. iTRAQ-based quantitative proteome and phosphoprotein characterization reveals the central metabolism changes involved in wheat grain development. Bmc Genomics [Electronic Resource] 2014;15:1029. https://doi.org/10.1186/1471-2164-15-1029.

97. Singh RP, Runthala A, Khan S, et al. Quantitative proteomics analysis reveals the tolerance of wheat to salt stress in response to Enterobacter cloacae SBP-8. PLoS One 2017;12(9):e0183513. https://doi.org/10.1371/journal.pone.0183513.

98. Tasleem-Tahir A, Nadaud I, Chambon C, et al. Expression profiling of starchy endosperm metabolic proteins at 21 stages of wheat grain development. J Proteome Res 2012;11(5):2754–73. https://doi.org/10.1021/pr201110d.

99. Yang M, Gao X, Dong J, et al. Pattern of protein expression in developing wheat grains identified through proteomic analysis. Front Plant Sci 2017;8:962. https://doi.org/10.3389/fpls.2017.00962.

100. He M, Zhu C, Dong K, et al. Comparative proteome analysis of embryo and endosperm reveals central differential expression proteins involved in wheat seed germination. BMC Plant Biol 2015;15:97. https://doi.org/10.1186/s12870-015-0471-z.

101. Molendijk J, Parker BL. Proteome-wide systems genetics to identify functional regulators of complex traits. Cell Syst 2021;12(1):5–22. https://doi.org/10.1016/j.cels.2020.10.005.

102. Chen S, Chen J, Hou F, et al. iTRAQ-based quantitative proteomic analysis reveals the lateral meristem developmental mechanism for branched spike development in tetraploid

wheat (Triticum turgidum L.). Bmc Genomics [Electronic Resource] 2018;19(1):228. https://doi.org/10.1186/s12864-018-4607-z.

103. Guo H, Zhang H, Li Y, et al. Identification of changes in wheat (Triticum aestivum L.) seeds proteome in response to anti-trxs gene. PLoS One 2011;6(7):e22255. https://doi.org/10.1371/journal.pone.0022255.

104. He X, Fang J, Li J, et al. A genotypic difference in primary root length is associated with the inhibitory role of transforming growth factor-beta receptor-interacting protein-1 on root meristem size in wheat. Plant J 2014;77(6):931–43. https://doi.org/10.1111/tpj.12449.

105. Islam N, Woo SH, Tsujimoto H, et al. Proteome approaches to characterize seed storage proteins related to ditelocentric chromosomes in common wheat (Triticum aestivum L.). Proteomics 2002;2(9):1146–55.

106. Kumar RR, Dubey K, Arora K, et al. Characterizing the putative mitogen-activated protein kinase (MAPK) and their protective role in oxidative stress tolerance and carbon assimilation in wheat under terminal heat stress. Biotechnol Rep (Amst) 2021;29:e00597. https://doi.org/10.1016/j.btre.2021.e00597.

107. Li HT, Sartika RS, Kerr ED, et al. Starch granular protein of high-amylose wheat gives innate resistance to amylolysis. Food Chem 2020;330:127328. https://doi.org/10.1016/j.foodchem.2020.127328.

108. Peng Z, Wang M, Li F, et al. A proteomic study of the response to salinity and drought stress in an introgression strain of bread wheat. Mol Cell Proteomics 2009;8(12):2676–86. https://doi.org/10.1074/mcp.M900052-MCP200.

109. Tahir A, Kang J, Choulet F, et al. Deciphering carbohydrate metabolism during wheat grain development via integrated transcriptome and proteome dynamics. Mol Biol Rep 2020;47(7):5439–49. https://doi.org/10.1007/s11033-020-05634-w.

110. Zhao Y, Zhang F, Mickan B, et al. Physiological, proteomic, and metabolomic analysis provide insights into Bacillus sp.-mediated salt tolerance in wheat. Plant Cell Rep 2022;41(1):95–118. https://doi.org/10.1007/s00299-021-02788-0.

111. Yu Z, Islam S, She M, et al. Wheat grain protein accumulation and polymerization mechanisms driven by nitrogen fertilization. Plant J 2018;96(6):1160–77. https://doi.org/10.1111/tpj.14096.

112. Daly DS, Anderson KK, Panisko EA, et al. Mixed-effects statistical model for comparative LC-MS proteomics studies. J Proteome Res 2008;7(3):1209–17. https://doi.org/10.1021/pr070441i.

113. D'Angelo G, Chaerkady R, Yu W, et al. Statistical models for the analysis of isobaric tags multiplexed quantitative proteomics. J Proteome Res 2017;16(9):3124–36. https://doi.org/10.1021/acs.jproteome.6b01050.

114. Goeminne LJ, Argentini A, Martens L, et al. Summarization vs peptide-based models in label-free quantitative proteomics: performance, pitfalls, and data analysis guidelines. J Proteome Res 2015;14(6):2457–65. https://doi.org/10.1021/pr501223t.

115. Klann K, Munch C. PBLMM: peptide-based linear mixed models for differential expression analysis of shotgun proteomics data. J Cell Biochem 2022;123(3):691–6. https://doi.org/10.1002/jcb.30225.

116. Pleil JD, Stiegel MA, Madden MC, et al. Heat map visualization of complex environmental and biomarker measurements. Chemosphere 2011;84(5):716–23. https://doi.org/10.1016/j.chemosphere.2011.03.017.

117. Zhang S, Ghatak A, Bazargani MM, et al. Spatial distribution of proteins and metabolites in developing wheat grain and their differential regulatory response during the grain filling process. Plant J 2021;107(3):669–87. https://doi.org/10.1111/tpj.15410.

118. Ertl P, Rohde B. The molecule cloud – compact visualization of large collections of molecules. J Cheminform 2012;4(1):12. https://doi.org/10.1186/1758-2946-4-12.

119. Khan IK, Bhuiyan M, Kihara D. DextMP: deep dive into text for predicting moonlighting proteins. Bioinformatics 2017;33(14):i83–91. https://doi.org/10.1093/bioinformatics/btx231.

120. Caetano-Anolles G. The compressed vocabulary of microbial life. Front Microbiol 2021;12:655990. https://doi.org/10.3389/fmicb.2021.655990.

121. McConnell P, Johnson K, Lin S. Applications of tree-maps to hierarchical biological data. Bioinformatics 2002;18(9):1278–9. https://doi.org/10.1093/bioinformatics/18.9.1278.

122. Baehrecke EH, Dang N, Babaria K, et al. Visualization and analysis of microarray and gene ontology data with treemaps. BMC Bioinf 2004;5:84. https://doi.org/10.1186/1471-2105-5-84.

123. Supek F, Bosnjak M, Skunca N, et al. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 2011;6(7):e21800. https://doi.org/10.1371/journal.pone.0021800.

124. Daba SD, Liu X, Aryal U, et al. A proteomic analysis of grain yield-related traits in wheat. AoB Plants 2020;12(5):plaa042. https://doi.org/10.1093/aobpla/plaa042.

125. Sharma A, Garg S, Sheikh I, et al. Effect of wheat grain protein composition on end-use quality. J Food Sci Technol 2020;57(8):2771–85. https://doi.org/10.1007/s13197-019-04222-6.

126. Yang M, Liu Y, Dong J, et al. Probing early wheat grain development via transcriptomic and proteomic approaches. Funct Integr Genomics 2020;20(1):63–74. https://doi.org/10.1007/s10142-019-00698-9.

127. Kanehisa M. KEGG bioinformatics resource for plant genomics and metabolomics. Methods Mol Biol 2016;1374:55–70. https://doi.org/10.1007/978-1-4939-3167-5_3.

128. Lv X, Zhang Y, Zhang Y, et al. Source-sink modifications affect leaf senescence and grain mass in wheat as revealed by proteomic analysis. BMC Plant Biol 2020;20(1):257. https://doi.org/10.1186/s12870-020-02447-8.

129. Yadav R, Chakraborty S, Ramakrishna W. Wheat grain proteomic and protein-metabolite interactions analyses provide insights into plant growth promoting bacteria-arbuscular mycorrhizal fungi-wheat interactions. Plant Cell Rep 2022;41(6):1417–37. https://doi.org/10.1007/s00299-022-02866-x.

130. Zhang Y, Pan J, Huang X, et al. Differential effects of a post-anthesis heat stress on wheat (Triticum aestivum L.) grain proteome determined by iTRAQ. Sci Rep 2017;7(1):3468. https://doi.org/10.1038/s41598-017-03860-0.

131. Soldatos TG, Perdigao N, Brown NP, et al. How to learn about gene function: text-mining or ontologies? Methods 2015;74:3–15. https://doi.org/10.1016/j.ymeth.2014.07.004.

132. Canto-Pastor A, Mason GA, Brady SM, et al. Arabidopsis bioinformatics: tools and strategies. Plant J 2021;108(6):1585–96. https://doi.org/10.1111/tpj.15547.

133. Fridrich A, Hazan Y, Moran Y. Too many false targets for MicroRNAs: challenges and pitfalls in prediction of miRNA targets and their gene ontology in model and non-model organisms. Bioessays 2019;41(4):e1800169. https://doi.org/10.1002/bies.201800169.

134. Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics 2020;36(8):2628–9. https://doi.org/10.1093/bioinformatics/btz931.

135. Bobrovskikh AV, Zubairova US, Bondar EI, et al. Transcriptomic data meta-analysis sheds light on high light response in arabidopsis thaliana L. Int J Mol Sci 2022;23(8):4455 . https://doi.org/10.3390/ijms23084455.

136. Karp PD, Latendresse M, Paley SM, et al. Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. Brief Bioinform 2016;17(5):877–90. https://doi.org/10.1093/bib/bbv079.

137. Hawkins C, Ginzburg D, Zhao K, et al. Plant Metabolic Network 15: a resource of genome-wide metabolism databases for 126 plants and algae. J Integr Plant Biol 2021;63(11):1888–905. https://doi.org/10.1111/jipb.13163.

138. Kondhare KR, Hedden P, Kettlewell PS, et al. Quantifying the impact of exogenous abscisic acid and gibberellins on pre-maturity alpha-amylase formation in developing wheat grains. Sci Rep 2014;4:5355. https://doi.org/10.1038/srep05355.

139. Derkx A, Baumann U, Cheong J, et al. A major locus on wheat chromosome 7B associated with late-maturity alpha-amylase encodes a putative ent-copalyl diphosphate synthase. Front Plant Sci 2021;12:637685. https://doi.org/10.3389/fpls.2021.637685.

140. Machicao J, Filho HA, Lahr DJG, et al. Topological assessment of metabolic networks reveals evolutionary information. Sci Rep 2018;8(1):15918. https://doi.org/10.1038/s41598-018-34163-7.

141. Gupta V, Estrada AD, Blakley I, et al. RNA-seq analysis and annotation of a draft blueberry genome assembly identifies candidate genes involved in fruit ripening, biosynthesis of bioactive compounds, and stage-specific alternative splicing. Gigascience 2015;4:5. https://doi.org/10.1186/s13742-015-0046-9.

142. Shi X, Sun H, Chen Y, et al. Transcriptome sequencing and expression analysis of cadmium (Cd) transport and detoxification related genes in Cd-accumulating Salix integra. Front Plant Sci 2016;7:1577. https://doi.org/10.3389/fpls.2016.01577.

143. Nadiya F, Anjali N, Thomas J, et al. Transcriptome profiling of Elettaria cardamomum (L.) Maton (small cardamom). Genom data. 2017;11:102–3. https://doi.org/10.1016/j.gdata.2016.12.013.

144. Sobhani Najafabadi A, Naghavi MR. Mining Ferula gummosa transcriptome to identify miRNAs involved in the regulation and biosynthesis of terpenes. Gene 2018;645:41–7. https://doi.org/10.1016/j.gene.2017.12.035.

145. Ganugi P, Miras-Moreno B, Garcia-Perez P, et al. Concealed metabolic reprogramming induced by different herbicides in tomato. Plant Sci 2021;303:110727. https://doi.org/10.1016/j.plantsci.2020.110727.

146. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. Genome Res 2009;19(9):1639–45. https://doi.org/10.1101/gr.092759.109.

147. Rasche H, Hiltemann S. Galactic Circos: user-friendly Circos plots within the Galaxy platform. Gigascience 2020;9(6):407–25. https://doi.org/10.1093/gigascience/giaa065.

148. McCleary BV, Sheehan H. Measurement of cereal α-amylase: a new assay procedure. J Cereal Sci 1987;6(3):237–51. https://doi.org/10.1016/S0733-5210(87)80061-9.

149. Vincent D. Data for "finding the LMA needle in the wheat proteome haystack" [MassIVE dataset MSV000090572]. 2022. https://massive.ucsd.edu/ProteoSAFe/result.jsp?task=9e8c4c3c9d924de8800237e7e828e1d9&view=advanced_view#%7B%7D.

150. Smilde AK, Jansen JJ, Hoefsloot HC, et al. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. Bioinformatics 2005;21(13):3043–8. https://doi.org/10.1093/bioinformatics/bti476.

151. R Core Team. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2021.

152. Batut B, Hiltemann S, Bagnacani A, et al. Community-driven data analysis training for biology. Cell Syst 2018;6(6):752–8. https://doi.org/10.1016/j.cels.2018.05.012.

153. Sigloch FC, Gruning B. Protein FASTA Database Handling. Galaxy Training 2021. https://training.galaxyproject.org/training-material/topics/proteomics/tutorials/database-handling/tutorial.html.

154. UniProt C. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 2021;49(D1):D480–D9. https://doi.org/10.1093/nar/gkaa1100.

155. Kanehisa M. The KEGG database. Novartis Found Symp 2002;247:91–101; discussion -3, 19-28, 244-52. https://doi.org/10.1002/0470857897.ch8.

156. Network PM. https://pmn.plantcyc.org/organism-summary?object=BREADWHEAT

157. Visualisation with Circos Galaxy Training 2021 https://training.galaxyproject.org/training-material/topics/visualisation/tutorials/circos/tutorial.html

158. Vincent D, Bui A, Ezernieks V, et al. Supporting data for "A community resource to mass explore the wheat grain proteome and its application to the late maturity alpha-amylase (LMA) problem..". GigaScience Database 2023; https://doi.org/10.5524/102436.

Received: April 24, 2023. **Revised:** August 2, 2023. **Accepted:** September 19, 2023