Finding the LMA needle in the wheat haystack proteome

Dr Delphine Vincent

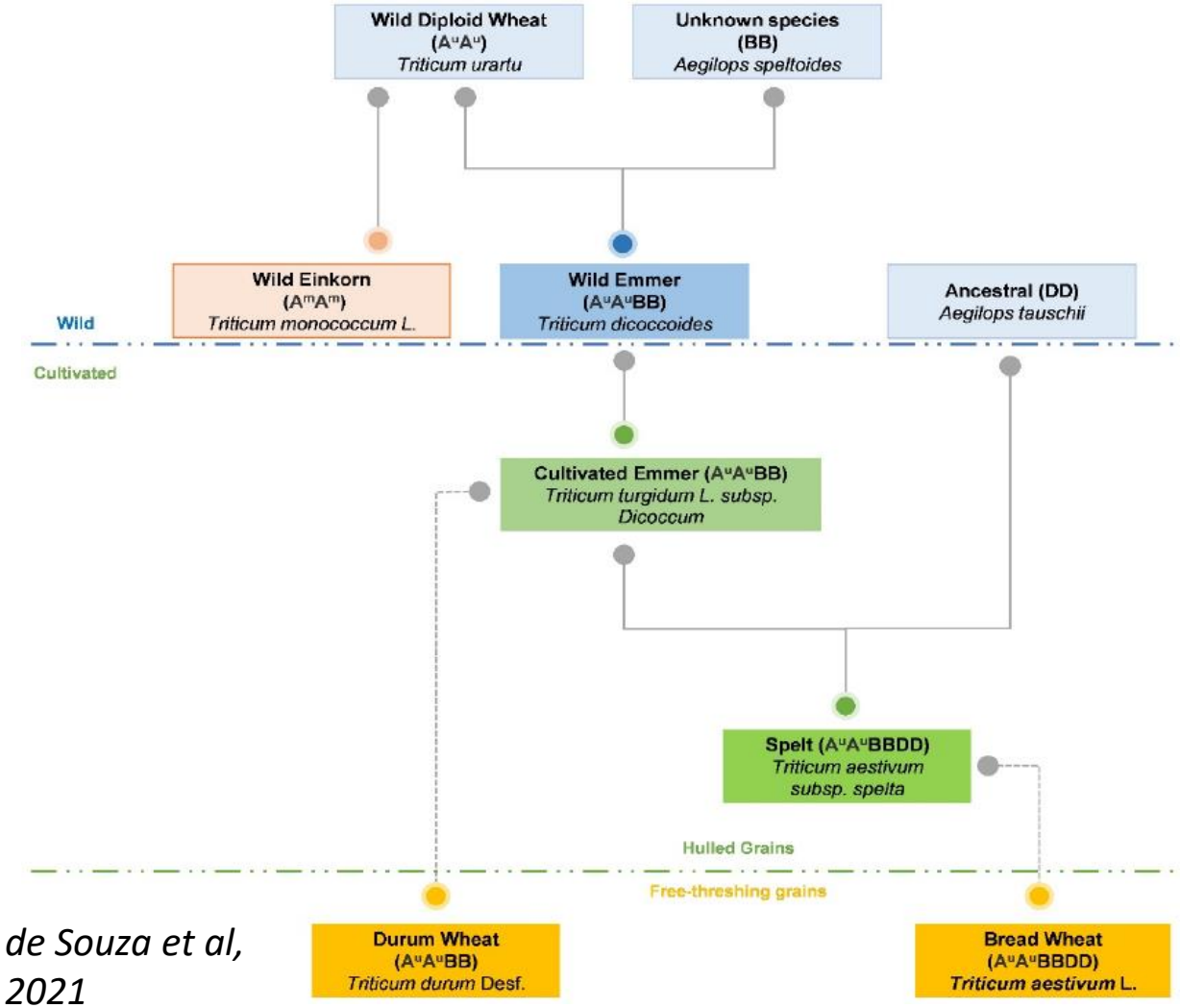
02/12/2022

Meules a Giverny, Monet 1890

Introduction – Wheat

Common bread wheat (Triticum aestivum L.) is the dominant crop in temperate regions.

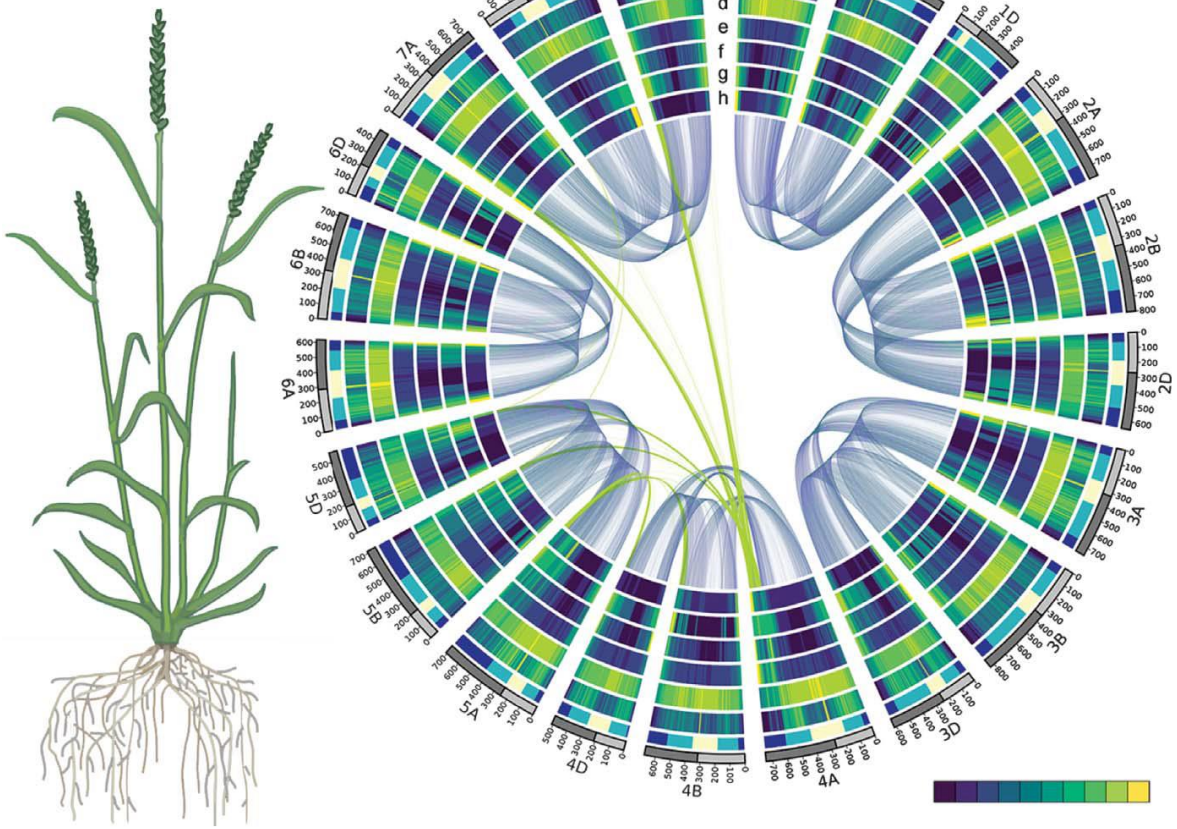
Its hexaploid genome (AABBDD; 2n = 6x = 42) originated from two polyploidization events.



de Souza et al, 2021

The genome was sequenced in 2018 by the International Wheat Genome Sequencing Consortium IWGSC.

The chromosomes from each closely related progenitor are grouped into **homeologous** groups.



IWCSG, 2018

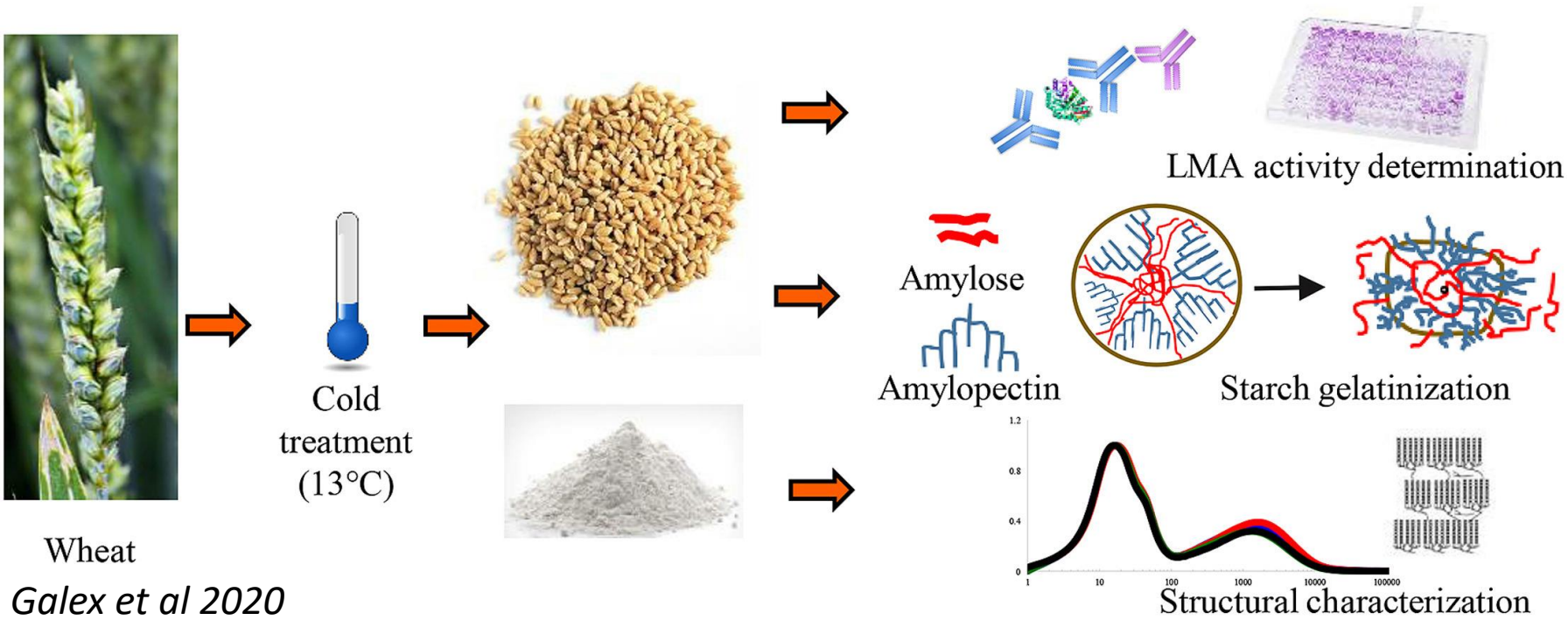
Introduction – LMA: a wheat industry concern

High isoelectric point (pI) α-amylase is normally synthesized after maturity in seeds when they may sprout in response to rain or germinate following sowing the next season's crop.

Late maturity α-amylase (LMA) is a wheat **genetic defect** causing the synthesis of high pI α-amylase in the aleurone as a result of a **temperature shock** during mid-grain development or **prolonged cold** throughout grain development.

In LMA-affected grains, the activated enzyme prematurely degrades the starch leading to grain discount downgraded to animal feed, which incurs a loss of profit for the producers.

Whilst the physiology is well understood, the biochemical mechanisms involved in grain LMA response remain unclear.

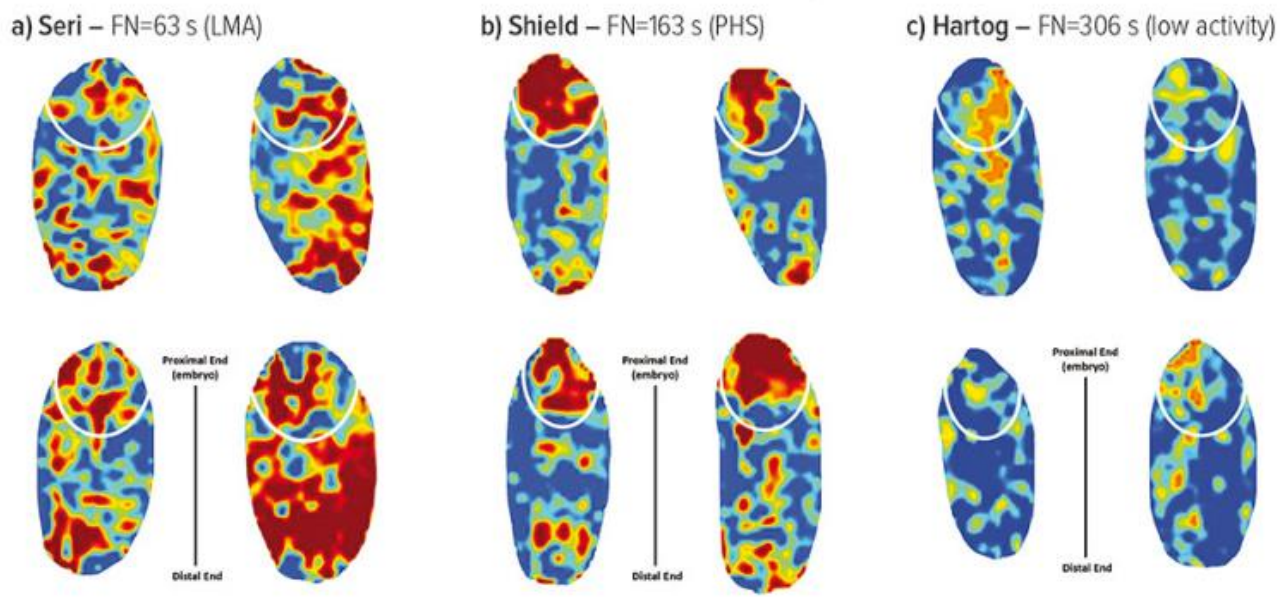


Galex et al 2020

Introduction – LMA: a hidden trait that prevails

Because of its stochastic nature, LMA is difficult to predict and invisible to the naked eye. Measurements involve laborious methods (falling number and Ceralpha assays) but alternatives are being investigated (hyperspectral tools).

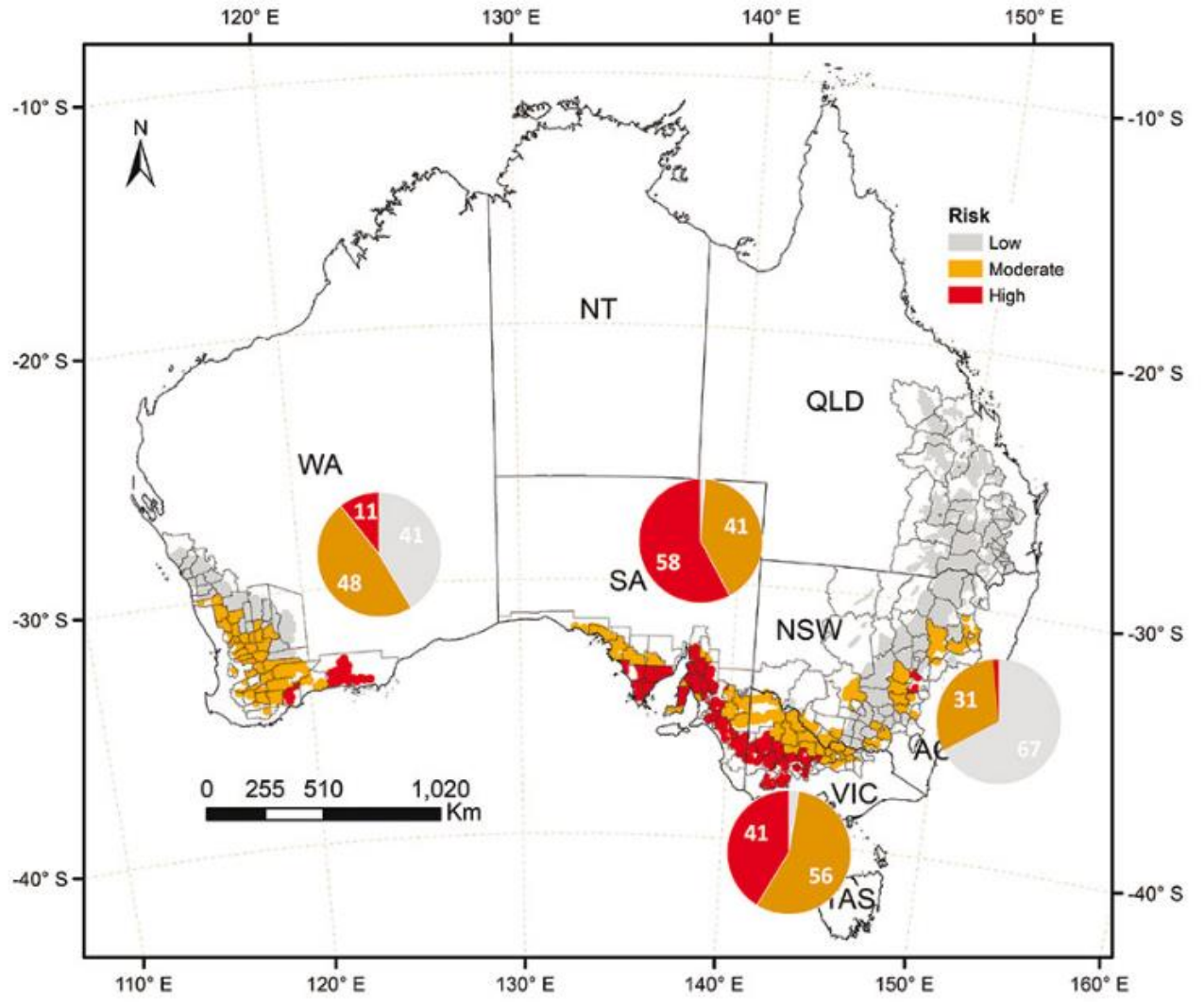
Figure 2: Hyperspectral images showing a) late maturity alpha-amylase enzyme activity, b) pre-harvest grain sprouting enzyme activity and c) little enzyme activity.



Source: Agriculture Victoria *(Panozzo et al)*

LMA is more prevalent than originally thought, with reports arising from North America, Japan, Canada, South Africa, China, Mexico, Germany, and the United Kingdom (Cannon et al, 2022).
In Australia, it mostly affects SW VIC and SE SA (red areas on map).

Figure 1: Risk footprint map for cool-shock temperatures equivalent to official cool-shock LMA testing protocols in the field, during the grain fill window. This is based on 1 May sowing date each year and a quick-maturing wheat variety, simulated using daily weather data from 1901 to 2016. Pie charts indicate percentage of land-use area associated with the broad risk classes.



Source: QAFFI *(Potgieter et al)*

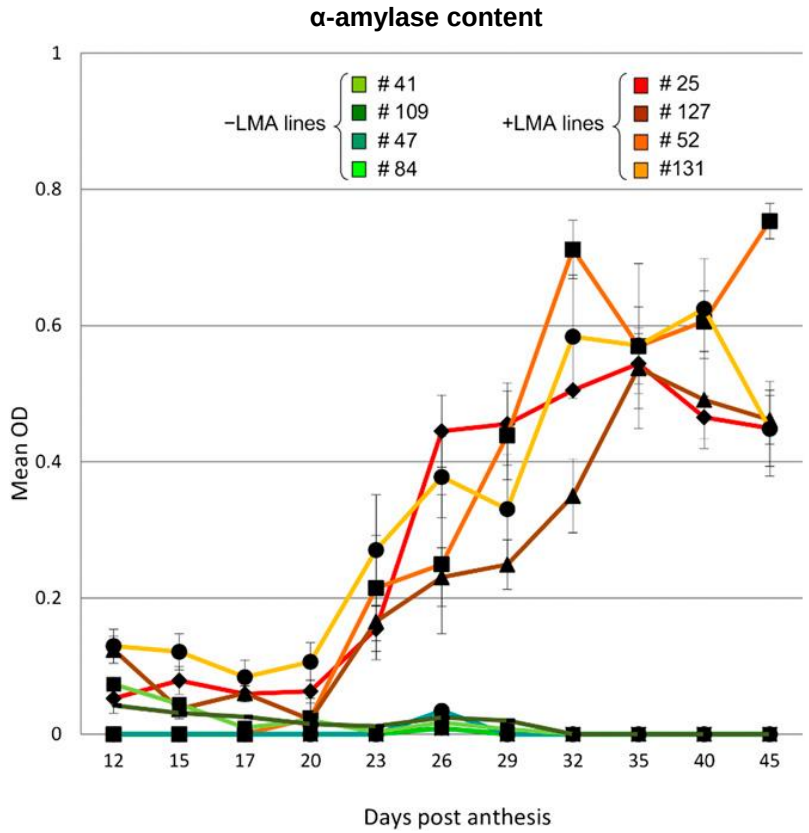
Introduction – LMA: GxE interaction perfect for post-genomics studies

LMA has a genetic (G) component (α-amylase gene required), yet it is only expressed and enzymatically active under particular environmental (E) conditions (temperature shock) at a given developmental stage (mid-grain) making it the product of a GxE interaction.

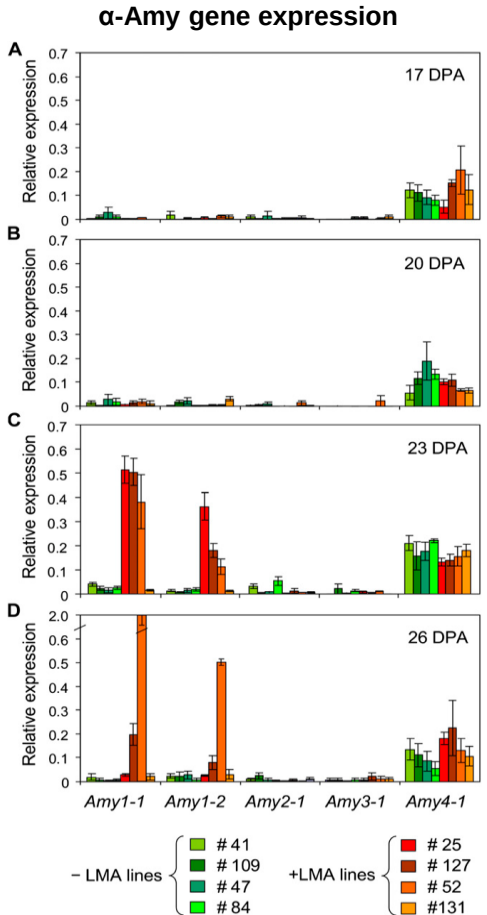
“A comprehensive understanding of LMA from the underlying molecular aspects to the end-use quality effects will greatly benefit the global wheat industry and those whose livelihoods depend upon it.” (Cannon et al, 2022)

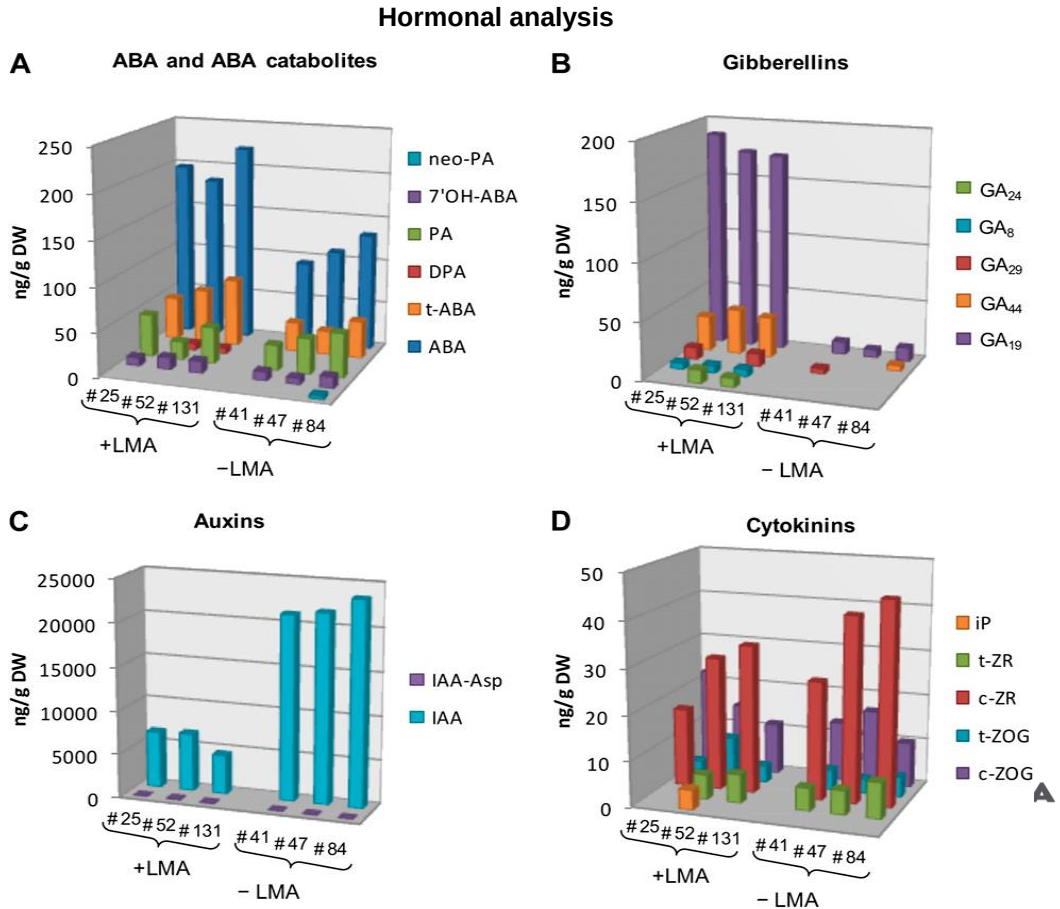
Post-genomics would deliver such deep understanding. Yet, to date, only one transcriptomics study has been published and no proteomics work has been attempted.

In 2014, using microarray technology, Barrero et al. reported that LMA resulted from very narrow and transitory peak of expression of genes encoding high pI α-amylase during grain development and triggered phytohormone responses, in particular elevated gibberellins.



Barrero et al, 2014





Improved phenotyping for late maturity α-amylase (LMA) susceptibility in wheat

Aim: to develop an innovative proteomics screening method with increased throughput, scalability, and repeatability and apply it to wheat germplasm to identify protein biomarkers involved in LMA response.

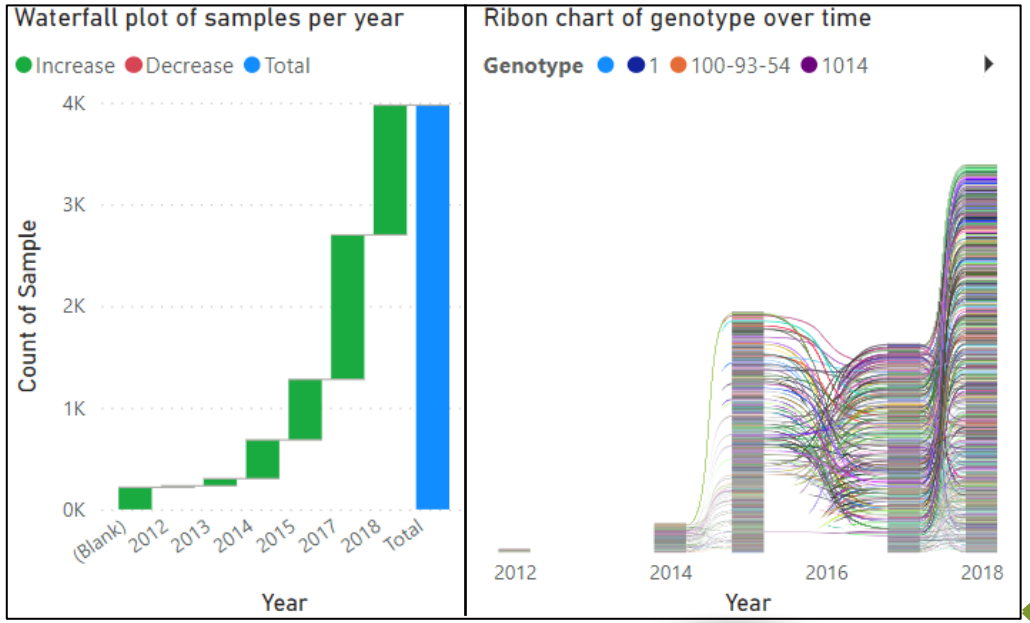
Source: whealbi.eu

"Mass spectrometry (MS)-based proteomics is the most comprehensive approach for the quantitative profiling of proteins, their interactions and modifications.
It is a challenging topic requiring **expertise in biochemistry for sample preparation, analytical chemistry for instrumentation and computational biology for data analysis**."
(Sinha and Mann, 2020)

Particularly when thousands of samples are processed!

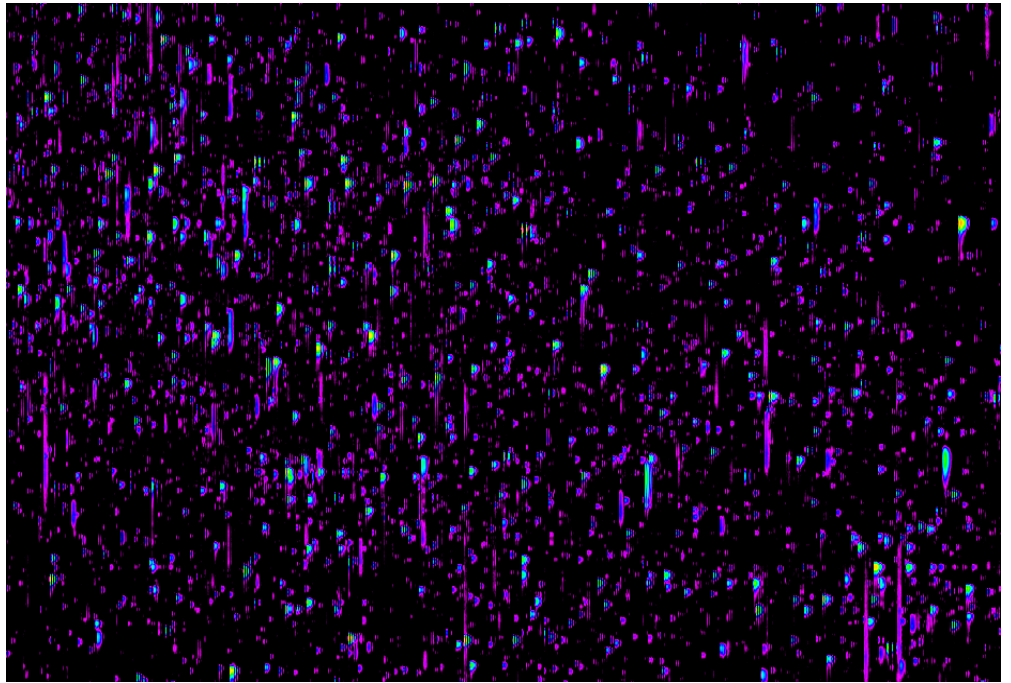
Study design

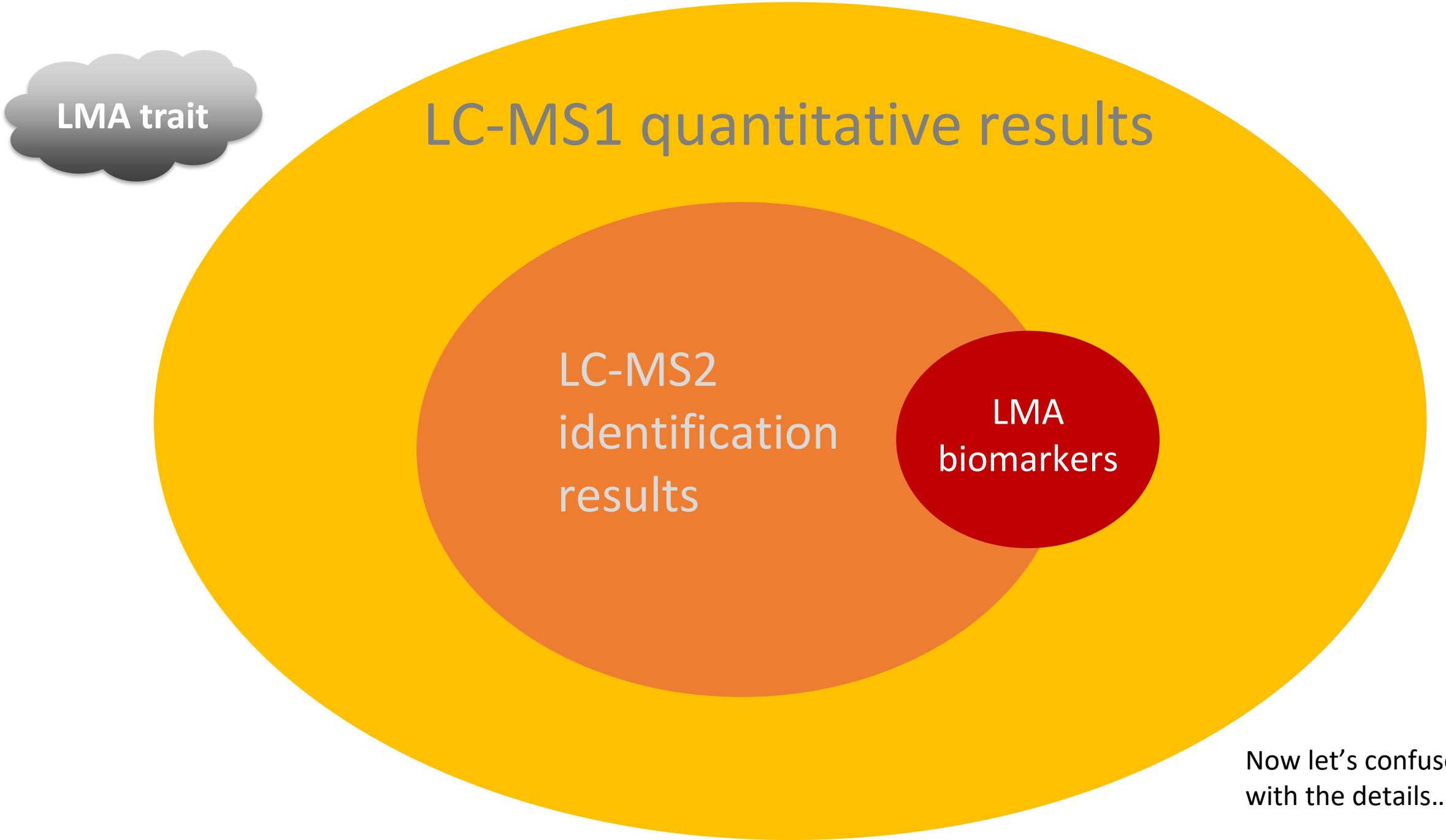
4,061 grain samples from 858 wheat genotypes sourced from all over the world, grown/harvested in Horsham from 2012-2018 and stored in optimal conditions.



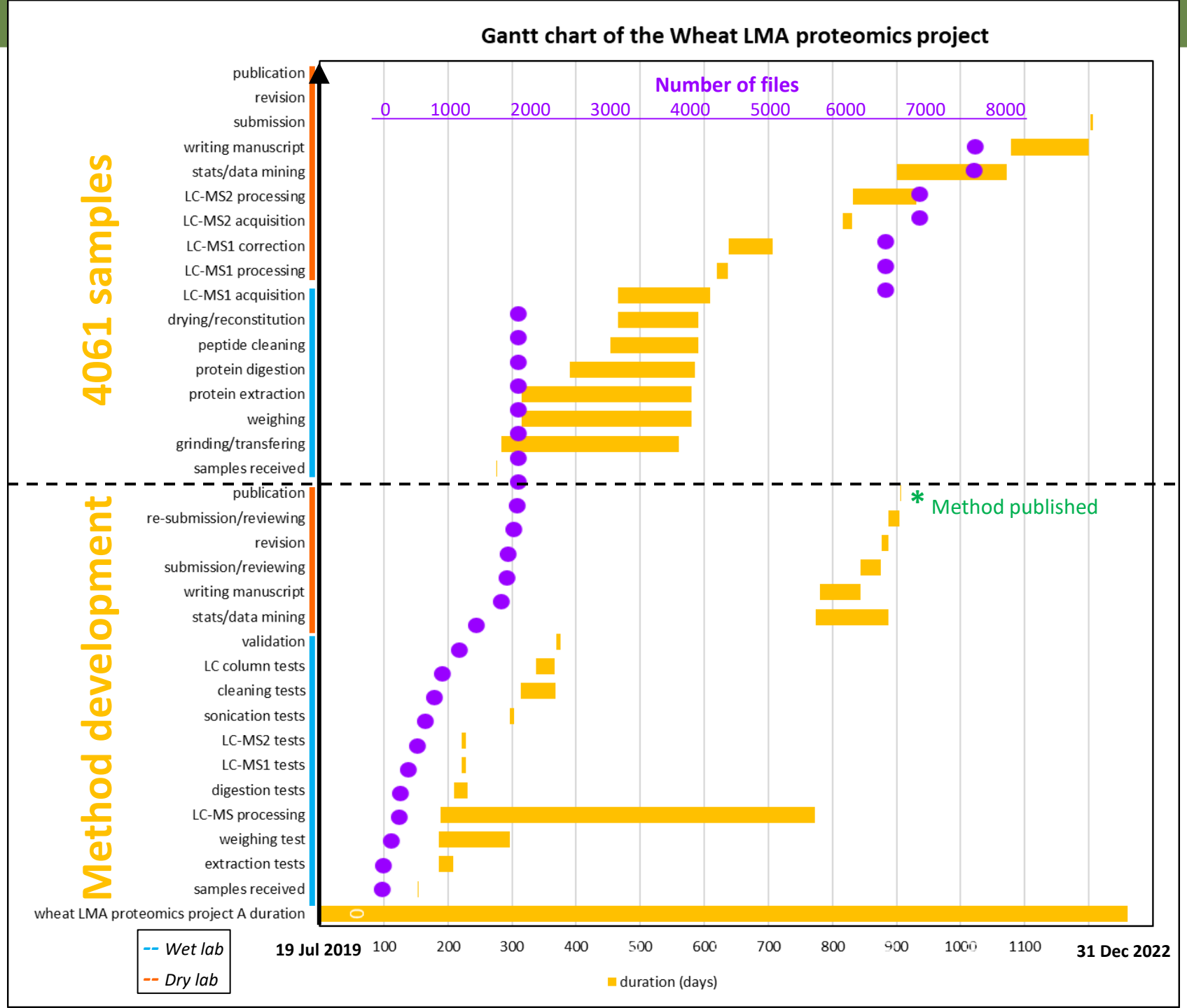
LMA assay

BU Proteomics





Now let's confuse you with the details…



Today's presentation

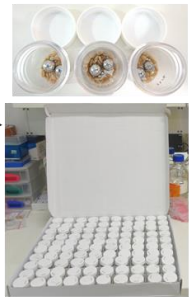
Presented in Oct 2021

*Vincent et al. Mining the Wheat Grain Proteome. Int. J. Mol. Sci. 2022, 23, 713. https://doi.org/10.3390/ijms23020713

High throughput workflow

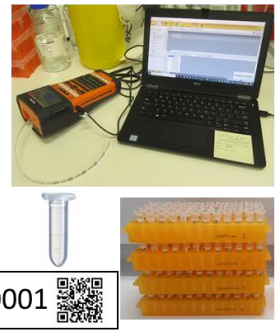
Samples are randomised and packaged in small plastic bags with QR labels. Scan QR to populate the tracking sample spreadsheet.

Pour all the grains in 50mL grinding jar with grinding metal balls.



Pulverise grains using a 2010 GenoGrinder.

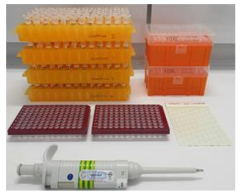
Transfer flour samples into tubes numbered and QR-coded for sample tracking.



Weigh 20.00 mg (+/- 0.02) of flour into 1.5 mL tube.

Extract wheat proteins using 0.5 mL buffer, sonication and heating block for denaturation. Centrifuge to retrieve supernatant.

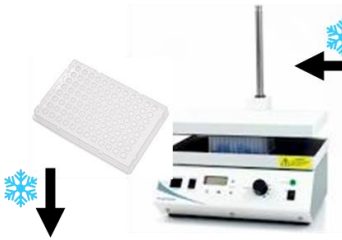
Pipette 10 uL protein extracts into 96-wells plates, spike with internal standard, and digest using trypsin/LysC proteases.



Clean-up wheat peptides using C18 Solid Phase Extraction.

Complete evaporation using SpeeVac fitted with a plate rotor.

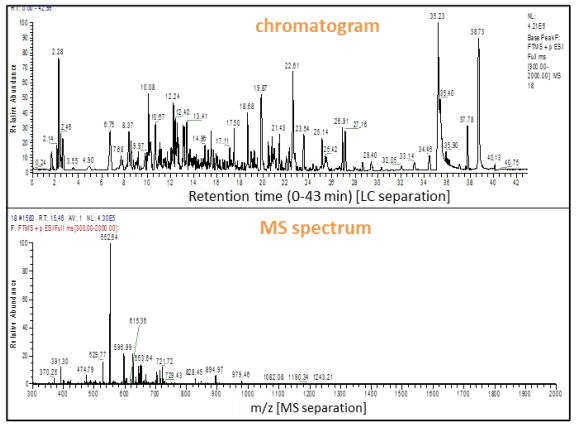
Reconstitution of dry samples in 100 µL mobile phase A using Ratek shaker.



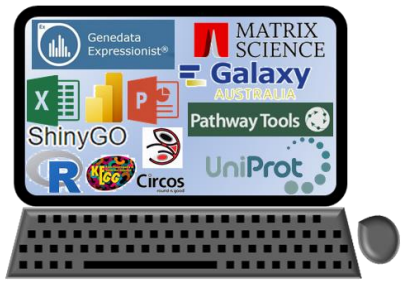
Place 4 x 96-wells plates of wheat peptide samples into Thermo Vanquish UPLC system.

Analyse samples using a Thermo LTQ-orbitrap mass spectrometer in FTMS mode.

Retrieve LC-MS RAW files and check quality. Repeat defective RAW files.



Process all files in Genedata Expressionist to export quantitative data and mine the data using other software.



Inject QC sample (mix of flour samples), IS, and blank every 48 samples.

Mass spectrometer calibrated every 3 weeks, mass accuracy < 50 ppm, heated capillary and LC column changed as required.

Requirements to minimise batch to batch variation:

1/ accurate sample weights

20 (+/- 0.2) mg flour was weighed to maximise sample preparation reproducibility

2/ internal standard (IS)

A pure peptide foreign to wheat spiked into all samples post-digestion and prior to LC-MS analysis

3/ quality control (QC)

An homogenous mix of 100 samples processed the same way as samples and analysed at regular intervals

4/ regular LC-MS maintenance

Consistent peptide separation and mass accuracy over time (5 months of continuous run for LC-MS1 acquisition)

Overall

Dataset: 3990 reproducible wheat samples x 32336 tryptic peptides (clusters) from LC-MS1 analyses
Trait: LMA measurements (with 5.4% (217) missing values)
Protein identification: LC-MS2 experiment performed post-hoc LC-MS1 acquisition.

Challenges:
Big dataset = big file size (2.5 Gb). Couldn't open it with excel. Super slow analysis in Genedata Analyst which often crashed.
Eliminate protein sequence redundancy.
Homoleogous proteins.
Finding appropriate ways to represent/mine the data.

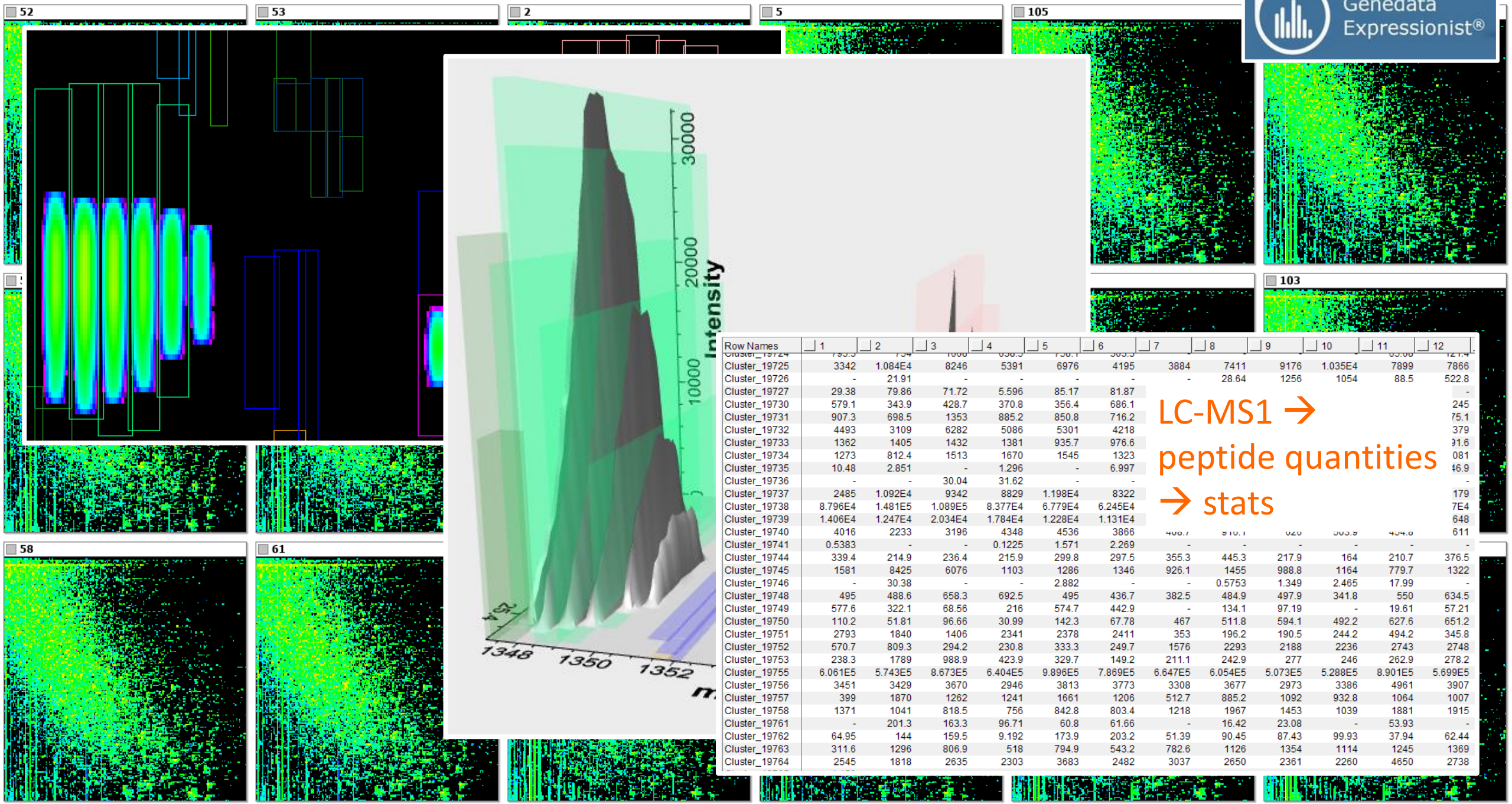
Software/Systems used (in no particular order):
Excel, Word, Powerpoint, Sharepoint, Sharedrive, Teams, OneNote, Endnote, P-Touch Editor, Xcalibur Qual Browser, Orbitrap LTQ Tune, Vanquish UPLC SII Direct Control, Genedata Refiner, Genedata Analyst, Mascot, Galaxy Australia, Veed.io, Circos, X2GO Client, WinSCP, RStudio, UniProt, EnsemblPlants, ShinyGO, Power BI, KEGG, Clustal Omega, Pathway Tools

Data files:
4400 raw files (MS1 and MS2)
>1000 files generated during data analysis/mining (20 Gb)

Aim of the analyses:
1/ raw data correction
2/ checking for normality of trait and corrected data
3/ data reduction
4/ predicting AAA missing values
5/ finding peptide biomarkers responding to AAA measurements
6/ linking MS1 to MS2 data and checking for MS2 method efficacy
7/ data mining tools
8/ finding biomarkers
9/ data representation

LC-MS1 quantitative results

LC-MS1 maps of wheat grains



Row Names	1	2	3	4	5	6	7	8	9	10	11	12
Cluster_19725	3342	1.084E4	8246	5391	6976	4195	3884	7411	9176	1.035E4	7899	7866
Cluster_19726	-	21.91	-	-	-	-	-	28.64	1256	1054	88.5	522.8
Cluster_19727	29.38	79.86	71.72	5.596	85.17	81.87						-
Cluster_19730	579.1	343.9	428.7	370.8	356.4	686.1						245
Cluster_19731	907.3	698.5	1353	885.2	850.8	716.2						75.1
Cluster_19732	4493	3109	6282	5086	5301	4218						379
Cluster_19733	1362	1405	1432	1381	935.7	976.6						91.6
Cluster_19734	1273	812.4	1513	1670	1545	1323						081
Cluster_19735	10.48	2.851	-	1.296	-	6.997						46.9
Cluster_19736	-	-	30.04	31.62	-	-						-
Cluster_19737	2485	1.092E4	9342	8829	1.198E4	8322						179
Cluster_19738	8.796E4	1.481E5	1.089E5	8.377E4	6.779E4	6.245E4						7E4
Cluster_19739	1.406E4	1.247E4	2.034E4	1.784E4	1.228E4	1.131E4						648
Cluster_19740	4016	2233	3196	4348	4536	3866						611
Cluster_19741	0.5383	-	-	0.1225	1.571	2.269	-	-	-	-	-	-
Cluster_19744	339.4	214.9	236.4	215.9	299.8	297.5	355.3	445.3	217.9	164	210.7	376.5
Cluster_19745	1581	8425	6076	1103	1286	1346	926.1	1455	988.8	1164	779.7	1322
Cluster_19746	-	30.38	-	-	2.882	-	-	0.5753	1.349	2.465	17.99	-
Cluster_19748	495	488.6	658.3	692.5	495	436.7	382.5	484.9	497.9	341.8	550	634.5
Cluster_19749	577.6	322.1	68.56	216	574.7	442.9	-	134.1	97.19	-	19.61	57.21
Cluster_19750	110.2	51.81	96.66	30.99	142.3	67.78	467	511.8	594.1	492.2	627.6	651.2
Cluster_19751	2793	1840	1406	2341	2378	2411	353	196.2	190.5	244.2	494.2	345.8
Cluster_19752	570.7	809.3	294.2	230.8	333.3	249.7	1576	2293	2188	2236	2743	2748
Cluster_19753	238.3	1789	988.9	423.9	329.7	149.2	211.1	242.9	277	246	262.9	278.2
Cluster_19755	6.061E5	5.743E5	8.673E5	6.404E5	9.896E5	7.869E5	6.647E5	6.054E5	5.073E5	5.288E5	8.901E5	5.699E5
Cluster_19756	3451	3429	3670	2946	3813	3773	3308	3677	2973	3386	4961	3907
Cluster_19757	399	1870	1262	1241	1661	1206	512.7	885.2	1092	932.8	1064	1007
Cluster_19758	1371	1041	818.5	756	842.8	803.4	1218	1967	1453	1039	1881	1915
Cluster_19761	-	201.3	163.3	96.71	60.8	61.66	-	16.42	23.08	-	53.93	-
Cluster_19762	64.95	144	159.5	9.192	173.9	203.2	51.39	90.45	87.43	99.93	37.94	62.44
Cluster_19763	311.6	1296	806.9	518	794.9	543.2	782.6	1126	1354	1114	1245	1369
Cluster_19764	2545	1818	2635	2303	3683	2482	3037	2650	2361	2260	4650	2738

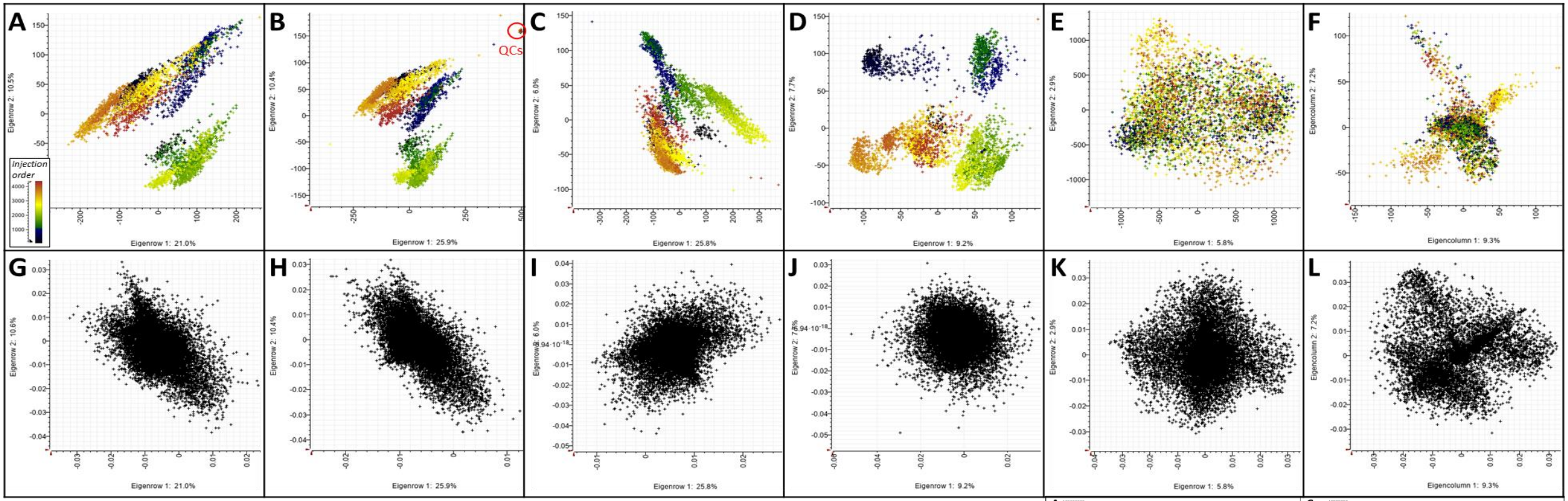
LC-MS1 → peptide quantities → stats

Proteomics data transformation and normalisation

Bag of tricks #1

Serial transformation steps

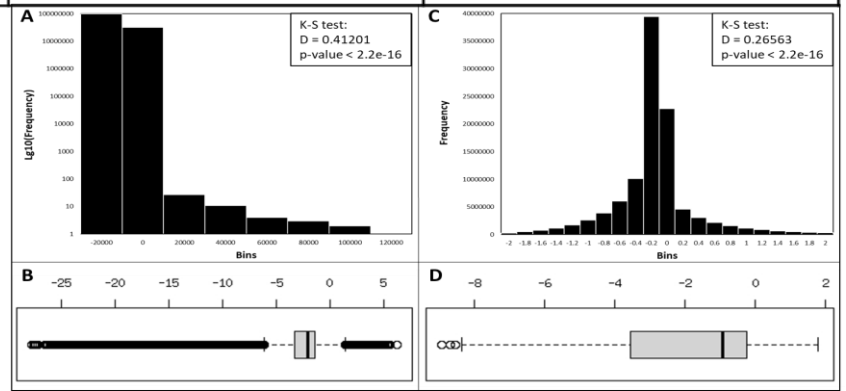
Raw data → *Sample weights* → *IS* → *QCs* → *Linear model* → *z-transformation*

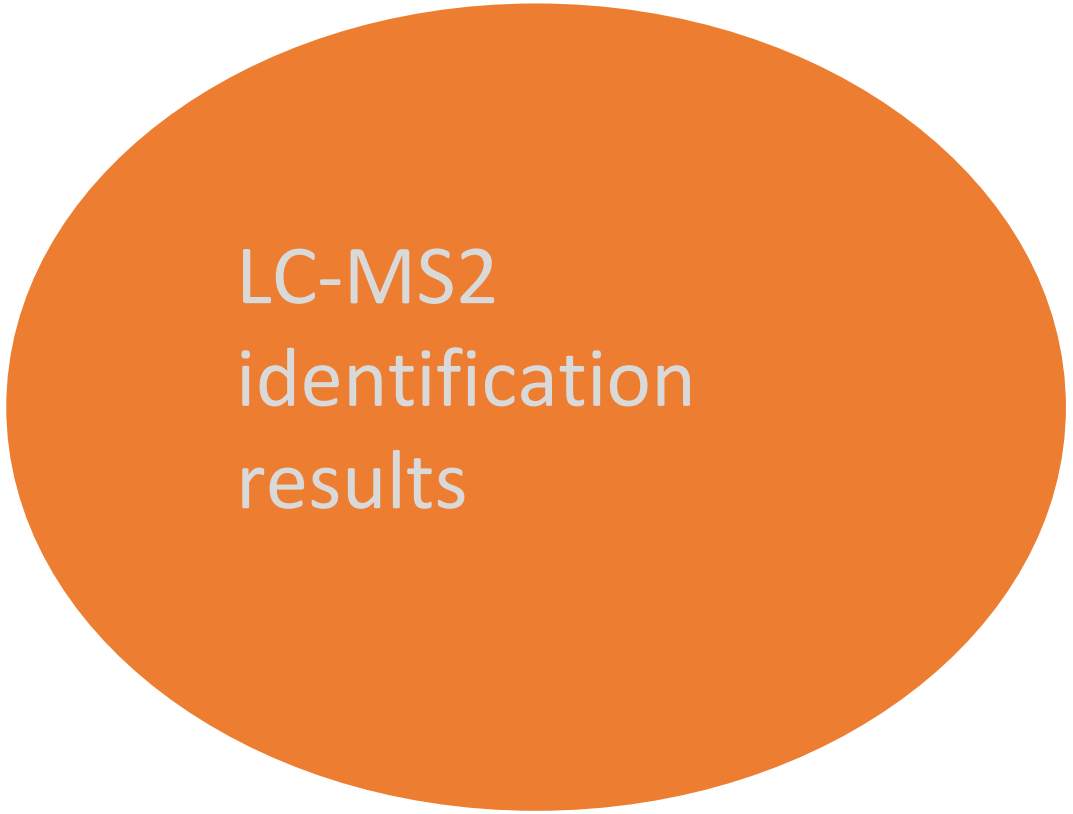


Despite capturing all experimental steps and making provisions to minimise technical variation, the only way to completely get rid of it was achieved by applying a linear regression and retrieving the residuals.

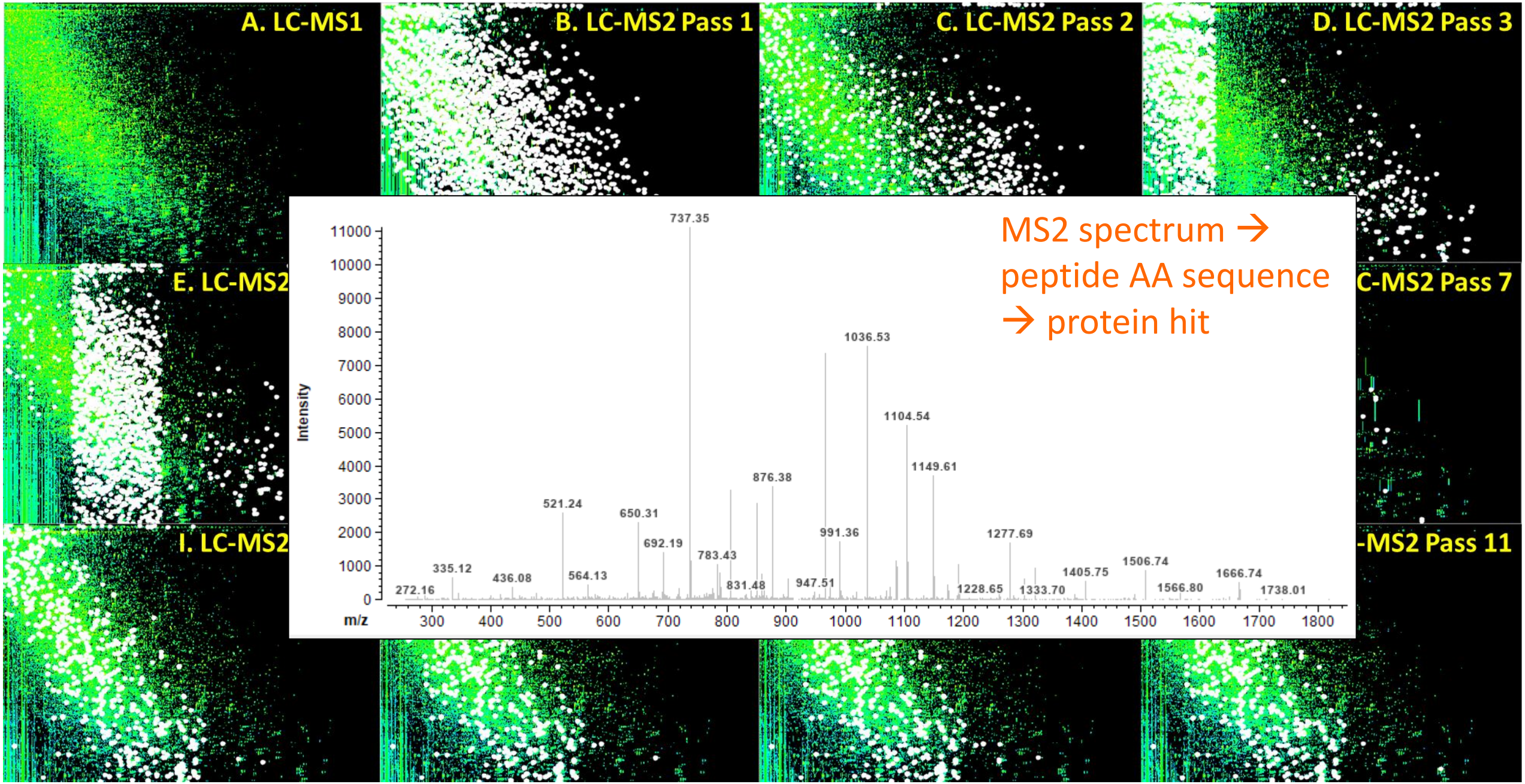
The last z-transformation was needed to achieve data normalisation.

→ **32,336 peptides quantified in 3,990 wheat samples.**




LC-MS2
identification
results

LC-MS2 maps of wheat grain



Wheat non redundant protein database with decoy sequences

Step 1 - 4 fasta files merged

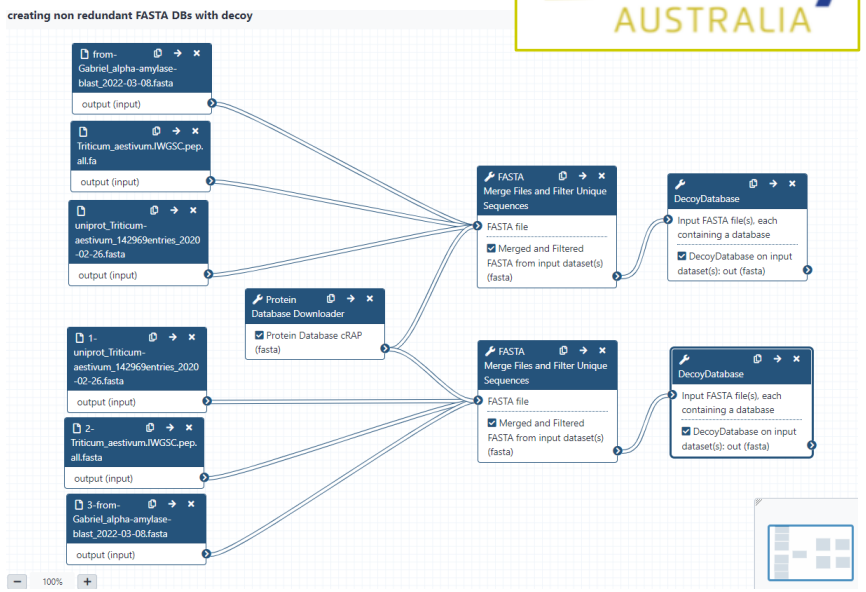
- 142,969 sequences from uniprot_Triticum-aestivum_142969entries_2020-02-26.fasta (UniprotKB)
- 143,241 sequences from Triticum_aestivum.IWGSC.pep.all.fasta (IWGSC wheat consortium)
- 116 sequences from Protein-DB-cRAP.fasta (contaminants)

Step 2 - Redundancy removed based on AA sequence (using Uniprot sequences as a template*-- 143,241 sequences left, none from IWGSC)

Step 3 – Decoy created (reversed sequences indexed with DECOY_)

Step 4 – download final file (286,482 sequences in Triticum-aestivum_Uniprot-Traes-cRAP_non-redundant_with-DECOY_2022-03-17.fasta)

*TRAES annotations refer to UniprotKB. Therefore Uniprot sequences were used as a template to eliminate the redundancy.



Example of Uniprot protein sequence: Alpha-amylase

>sp|P08117|AMY3_WHEAT Alpha-amylase AMY3 OS=Triticum aestivum OX=4565 GN=AMY1.1 PE=2 SV=1
MGKHSATLCGLLVVVLCLASSLAQAQILFQGFNWESWKTQGGWYKFMQGKVEEIASTGATHVWLPPPSQSVSPEGYLPGQLYNLNSKYGSGADLKSLIQAFRGKNISCVADIV
INHRCADKKDGRGVYCIFEGGTSDNRLDWGPDEICSDDTKYSNGRGHRDTGGGFDAAPDIDHLNPRVQRELSAWLNWLKTDLGFDGWRLDFAKGYSAAMAKIYVDNSKPAF
VVGELYDRDRQLLANWVRGVGGPATAFDFPTKGVLQEAVQGDLGRMRGSDGKAPGMIGWMPEKTVTFIDNHDTGSTQRLWPFPSDKVMQGYAYILTHPGIPCIFYDHVFD
WKLKQEITALATVRSRNGIHPGSTLDILKAEGDLYVAKIGGKVITKIGSRYNIGDNVIPSGFKIAAKGNNYCVWEKSGL

Example of TRAES protein sequence: Alpha-amylase

>TraesCS5A02G464500.1 pep chromosome:IWGSC:5A:643924798:643926446:-1 gene:TraesCS5A02G464500 transcript:TraesCS5A02G464500.1 gene_biotype:protein_coding transcript_biotype:protein_coding gene_symbol:AMY1.1 description:Alpha-amylase AMY3 [Source:UniProtKB/Swiss-Prot;Acc:P08117]
MGKHSATLCGLLVVVLCLASSLAQAQILFQGFNWESWKTQGGWYKFMQGKVEEIASTGATHVWLPPPSQSVSPEGYLPGQLYNLNSKYGSGADLKSLIQAFRGKNISCVADIV
INHRCADKKDGRGVYCIFEGGTSDNRLDWGPDEICSDDTKYSNGRGHRDTGGGFDAAPDIDHLNPRVQRELSAWLNWLKTDLGFDGWRLDFAKGYSAAMAKIYVDNSKPAF
VVGELYDRDRQLLANWVRGVGGPATAFDFPTKGVLQEAVQGDLGRMRGSDGKAPGMIGWMPEKTVTFIDNHDTGSTQRLWPFPSDKVMQGYAYILTHPGIPCIFYDHVFD
WKLKQEITALATVRSRNGIHPGSTLDILKAEGDLYVAKIGGKVITKIGSRYNIGDNVIPSGFKIAAKGNNYCVWEKSGL

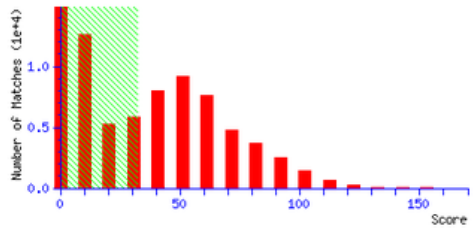
Mascot search for protein identification

MATRIX SCIENCE

Search parameters

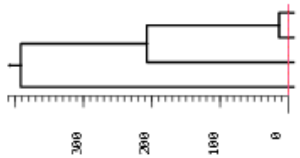
Type of search	: MS/MS Ion Search
Error tolerance	: Error tolerant search of all significant protein families
Enzyme	: Trypsin
Fixed modifications	: Carbamidomethyl (C)
Mass values	: Monoisotopic
Protein mass	: Unrestricted
Peptide mass tolerance	: ± 10 ppm
Fragment mass tolerance	: ± 0.5 Da
Max missed cleavages	: 9
Instrument type	: ESI-TRAP
Number of queries	: 97,195

Score distribution



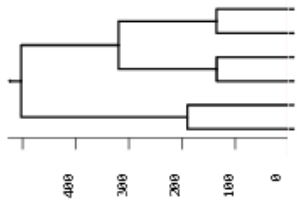
Modification	Site	Above thr.	ET	Total matches
Carbamidomethyl	C	19380	0	19380
Non-specific cleavage	-	0	4653	4653
Ammonia-loss	N-term	0	849	849
Gln->pyro-Glu	N-term	0	698	698
Ethyl	N-term	0	248	248
Ser->Gly	S	0	203	203
Gln->Lys	Q	0	178	178
Guanidinyl	N-term	0	134	134
Deamidated	Q	0	130	130
Gln->Ala	Q	0	101	101
Asp->Asn	D	0	98	98
Arg->GluSA	R	0	92	92
Thr->Ala	T	0	87	87
Glu->pyro-Glu	N-term	0	79	79
Val->Glu	V	0	79	79
Oxidation	M	0	79	79
Amidated	C-term	0	74	74
Ser->Pro	S	0	70	70
Thr->Gly	T	0	67	67
Cation:Na	C-term	0	50	50
Asn->His	N	0	50	50
Asn->Lys	N	0	50	50
Deoxy	T	0	45	45
Trimethyl	K	0	45	45
Xle->Ser	I	0	43	43
Ammonia-loss	N	0	43	43
Trioxidation	C	0	40	40
Asn->Ser	N	0	39	39
Deoxy	S	0	39	39
Ser->Xle	S	0	37	37
Label:13C(6)15N(4)	R	0	36	36
Arg->Ser	R	0	36	36
Thr->His	T	0	36	36
Cation:Na	D	0	35	35
Glu->Lys	E	0	35	35
ICPL:13C(6)	N-term	0	35	35
Cation:Na	E	0	33	33
Xle->Pro	I	0	33	33
Val->Ala	V	0	32	32
Label:13C(9)15N(1)	F	0	30	30
Carbamidomethyl	N-term	0	29	29
Methyl	N-term	0	29	29
Glu->Gln	E	0	28	28
Gly->Ala	G	0	27	27
Acetyl	N-term	0	27	27
Lys->Glu	K	0	26	26
	I	0	23	23
	L	0	23	23
	L	0	22	22

10 log(*P*), where *P* is
a random event.
identity threshold and
d for **97,195** queries.
beyond green
mology (p<0.05).



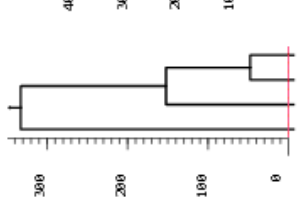
1	1	tr|Q5UHH7|Q5UHH7_...	68346
	4	tr|A0A3B6ECF3|A0A...	7573
	2	tr|Q5MD68|Q5MD68_...	50395
	3	tr|Q4U195|Q4U195_...	22397

tr|Q5UHH7|Q5UHH7_WHEAT 0.19 dimeric alpha-amylase inhibitor (Fragment) OS=Triticum aestivum OX=4565 PE=4 SV=1
tr|A0A3B6ECF3|A0A3B6ECF3_WHEAT AAI domain-containing protein OS=Triticum aestivum OX=4565 PE=4 SV=1
tr|Q5MD68|Q5MD68_WHEAT 0.19 dimeric alpha-amylase inhibitor (Fragment) OS=Triticum aestivum OX=4565 PE=4 SV=1
tr|Q4U195|Q4U195_WHEAT Dimeric alpha-amylase inhibitor OS=Triticum aestivum OX=4565 PE=4 SV=1



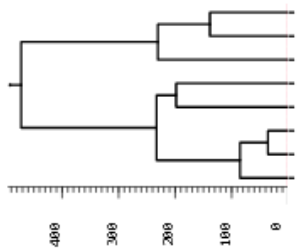
2	1	tr|I6QQ39|I6QQ39_...	66676
	4	tr|A0A3B6JBZ2|A0A...	31971
	2	tr|A0A3B6ILV9|A0A...	52860
	3	tr|A0A3B6IJ76|A0A...	48825
	6	tr|B7U6L5|B7U6L5_...	15150
	5	tr|A0A3B6JER7|A0A...	26409

tr|I6QQ39|I6QQ39_WHEAT Globulin-3A OS=Triticum aestivum OX=4565 GN=Glo-3A PE=2 SV=1
tr|A0A3B6JBZ2|A0A3B6JBZ2_WHEAT Uncharacterized protein OS=Triticum aestivum OX=4565 PE=4 SV=1
tr|A0A3B6ILV9|A0A3B6ILV9_WHEAT Uncharacterized protein OS=Triticum aestivum OX=4565 PE=4 SV=1
tr|A0A3B6IJ76|A0A3B6IJ76_WHEAT Uncharacterized protein OS=Triticum aestivum OX=4565 PE=4 SV=1
tr|B7U6L5|B7U6L5_WHEAT Globulin 3B OS=Triticum aestivum OX=4565 GN=glo-3B PE=4 SV=1
tr|A0A3B6JER7|A0A3B6JER7_WHEAT Uncharacterized protein OS=Triticum aestivum OX=4565 PE=4 SV=1



3	1	tr|A0A3B6KSH4|A0A...	53881
	4	tr|M1MQ51|M1MQ51_...	10356
	2	tr|A0A3B6TZ67|A0A...	51745
	3	tr|A0A3B6IYD4|A0A...	46579

tr|A0A3B6KSH4|A0A3B6KSH4_WHEAT Beta-amylase OS=Triticum aestivum OX=4565 PE=3 SV=1
tr|M1MQ51|M1MQ51_WHEAT Beta-amylase (Fragment) OS=Triticum aestivum OX=4565 GN=BMY1 PE=2 SV=1
tr|A0A3B6TZ67|A0A3B6TZ67_WHEAT Beta-amylase OS=Triticum aestivum OX=4565 PE=3 SV=1
tr|A0A3B6IYD4|A0A3B6IYD4_WHEAT Beta-amylase OS=Triticum aestivum OX=4565 PE=3 SV=1



4	1	tr|A0A3B6MWJ...
	2	sp|Q41593|SPZ...
	3	sp|Q9ST58|SPZ...
	7	tr|A0A3B6IPZ0
	8	tr|A0A3B6TLW2
	4	tr|A0A3B6KQL2
	5	sp|Q9ST57|SPZ...
	6	sp|P93692|SPZ...

tr|Q5UHH7|Q5UHH7_WHEAT 0.19 dimeric alpha-amylase inhibitor (Fragment) OS=Triticum aestivum OX=4565 PE=4 SV=1

Database:	Triticum-aestivum_Uniprot-Traes
Score:	68346
Monoisotopic mass (M_r):	13899
Calculated pI:	6.66

Sequence similarity is available as an NCBI BLAST search of tr|Q5UHH7|Q5UHH7_WHEAT against nr.

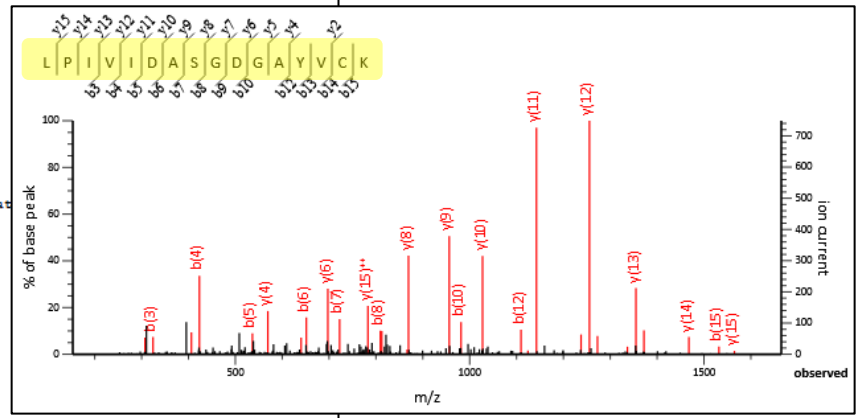
Search parameters

MS data file:	/genedata/runtime/13.5/refiner_ms/var/cache/refiner_ms/13.5/disk1/refiner_ms/13.5/temp/genedat
Enzyme:	Trypsin: cuts C-term side of KR unless next residue is P.
Fixed modifications:	Carbamidomethyl (C)

Protein sequence coverage: 94%

Matched peptides shown in ***bold red***.

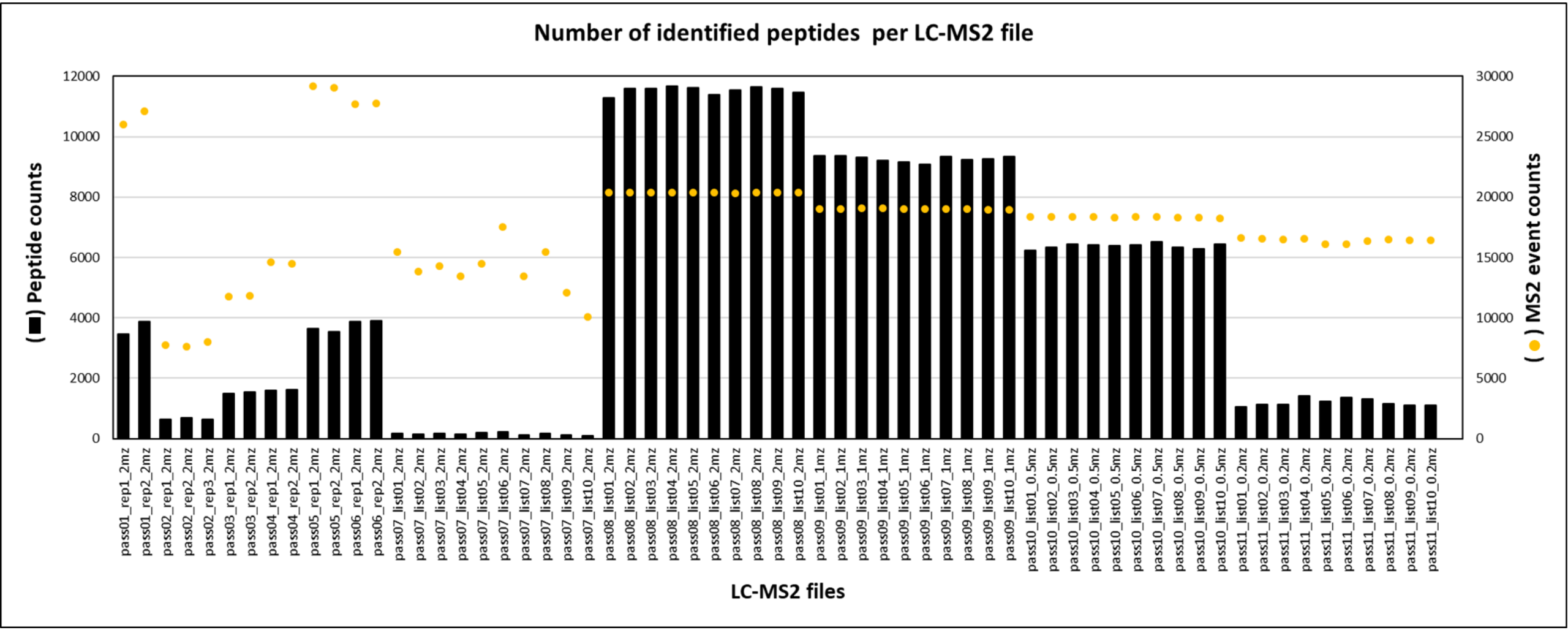
```
  1 SGPWMCYPGQ AFQVPALPAC RPLLRLQCNG SQVPEAVLRD CCQQLAHISE
 51 WCRCGALYSM LDSMYKEHGA QEGQAGTGAF PRCRREVVKL TAASITAVCR
101 LPIVVDASGD GAYVCKDVAA YPDA
```



Identification success rate

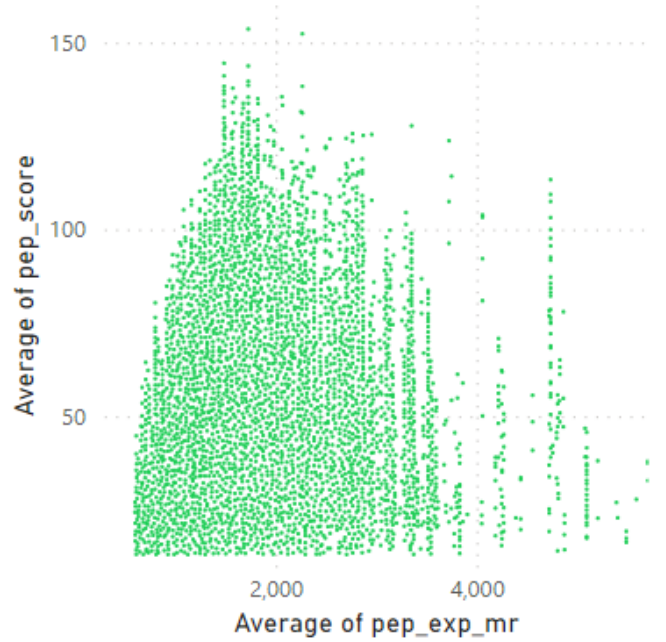
LC-MS2 experiment not performed on all samples; random subsampling and pooling of 400 samples.
Multiple rounds (63) of MS2 methods on the single pooled sample for deeper proteome coverage.
Hit yield varied greatly from method to method.

→ **6,550 unique peptides identified.**

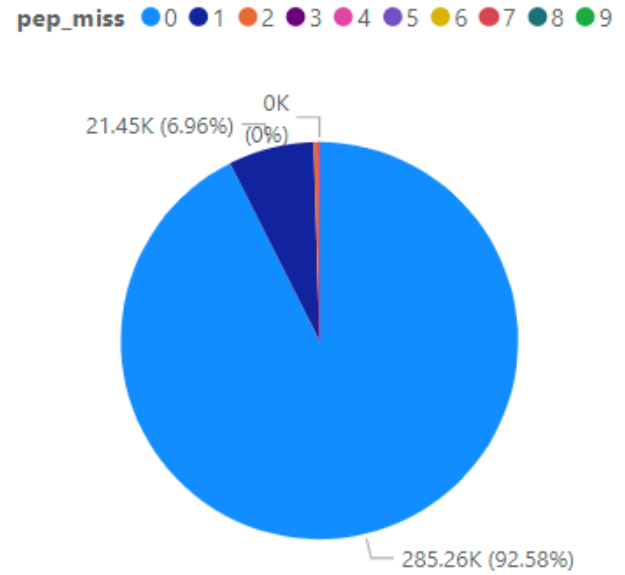


Identities charted in Power BI

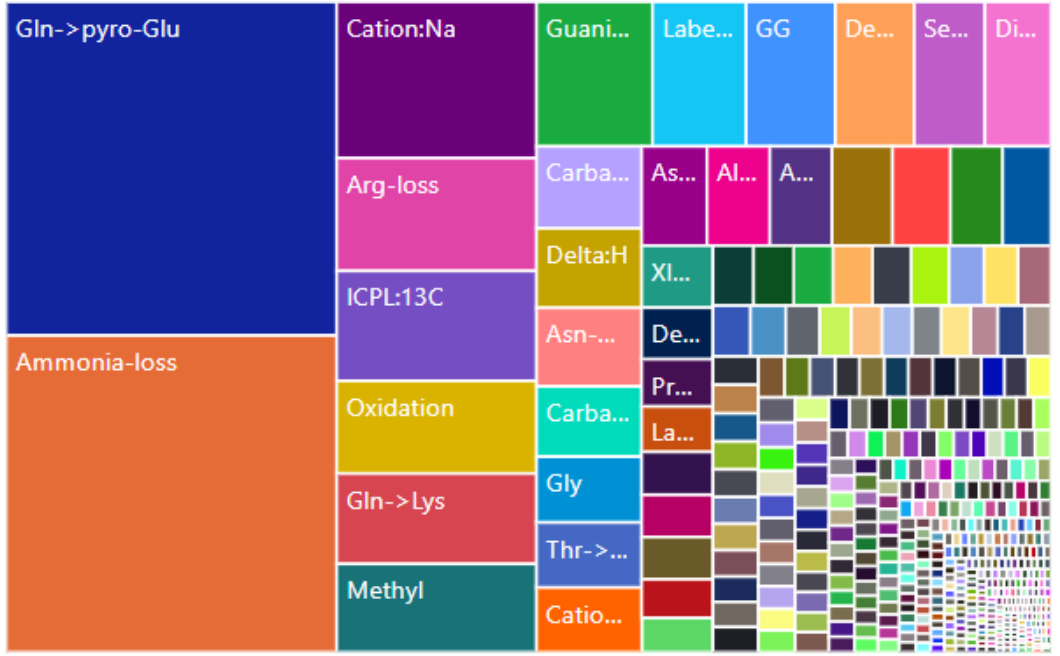
Peptide mass by score



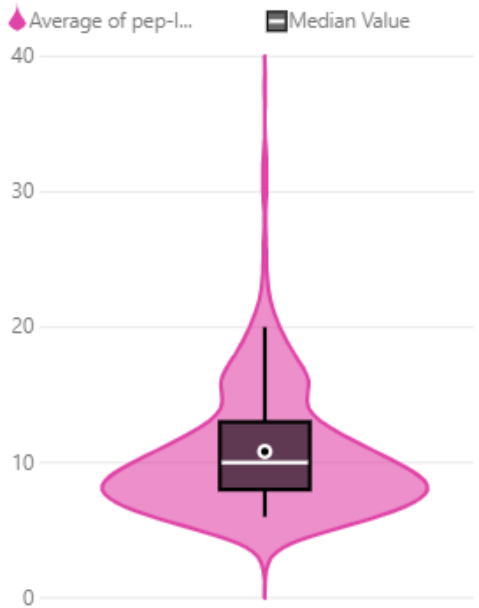
Missed cleavage



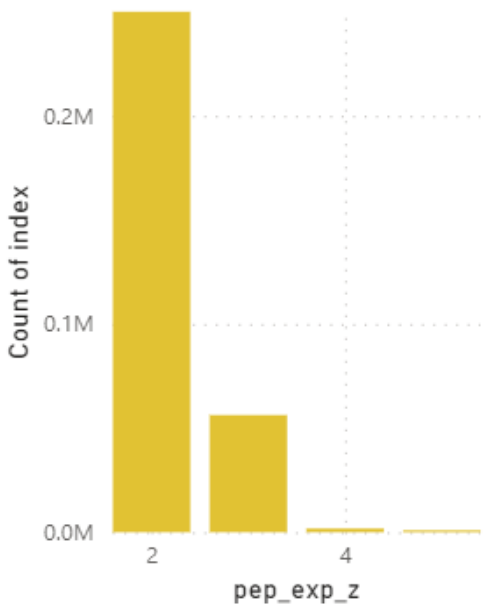
Peptide modifications



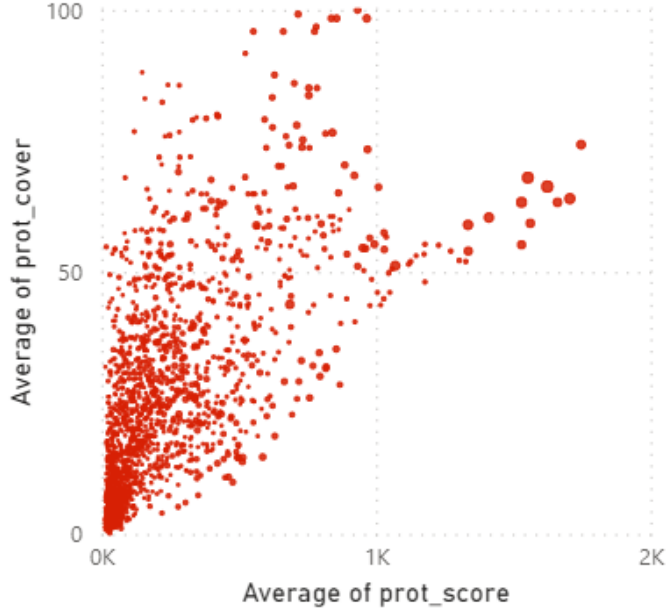
Peptide length (AA)



Peptide charge



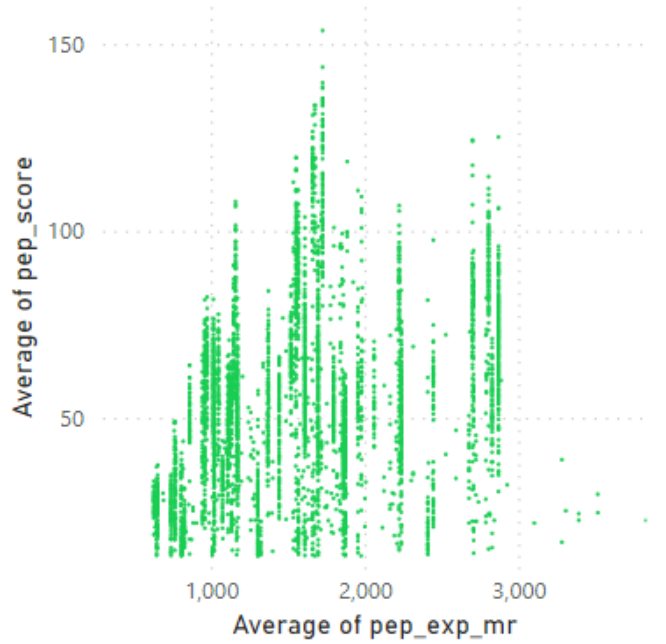
Protein score by sequence coverage



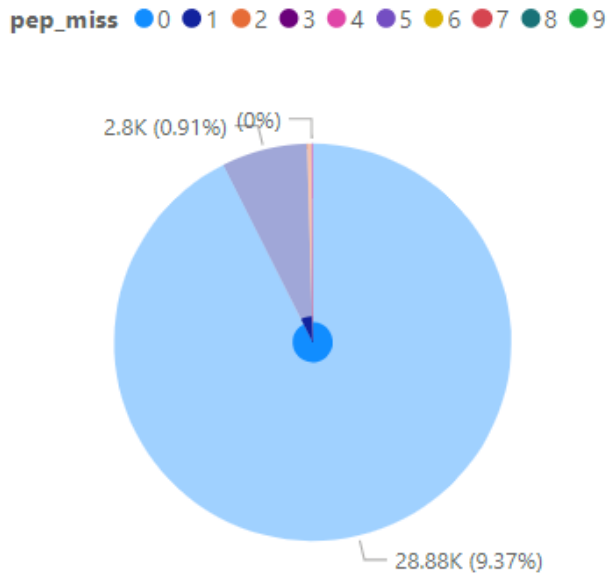
Protein description

Identities charted – drill in on inhibitors

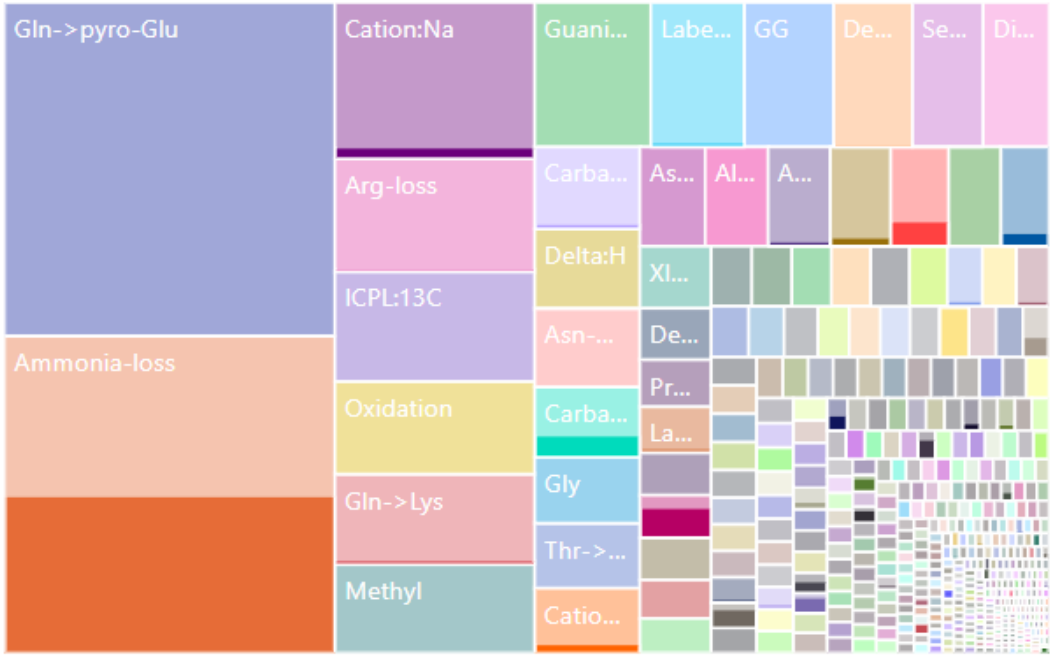
Peptide mass by score



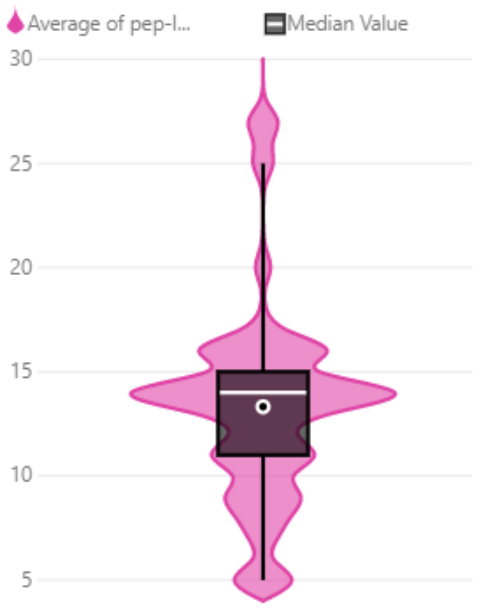
Missed cleavage



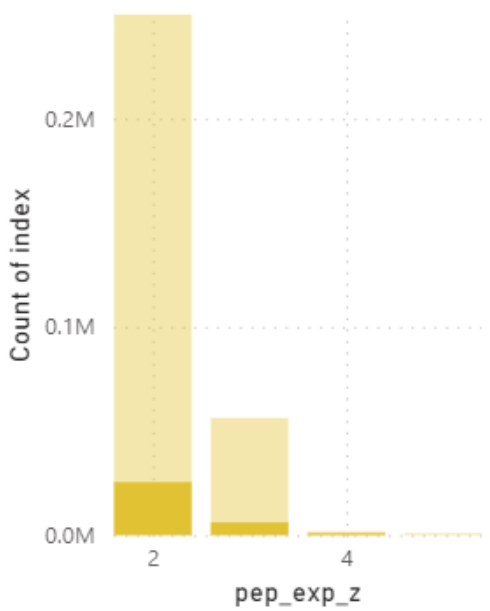
Peptide modifications



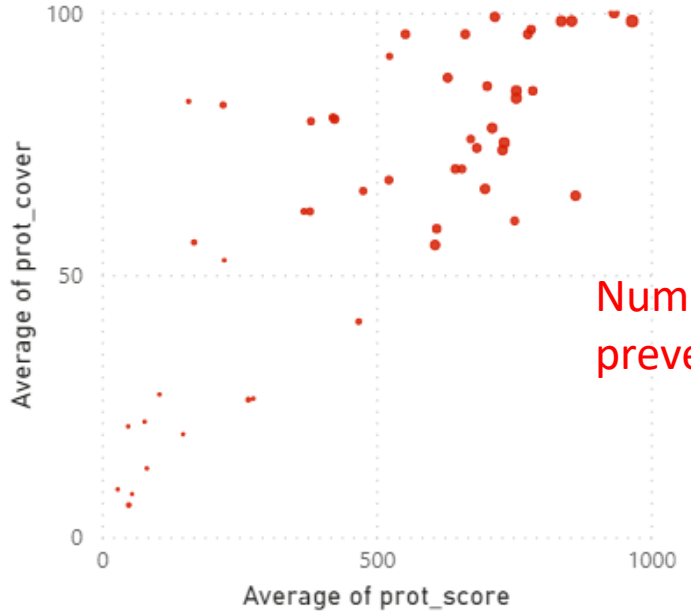
Peptide length (AA)



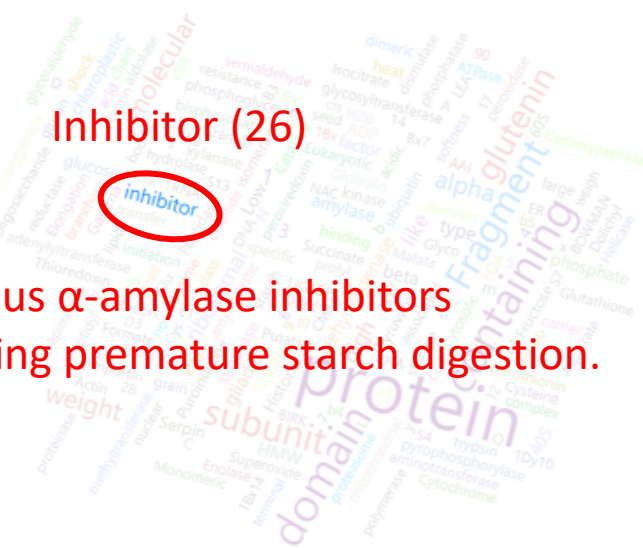
Peptide charge



Protein score by sequence coverage



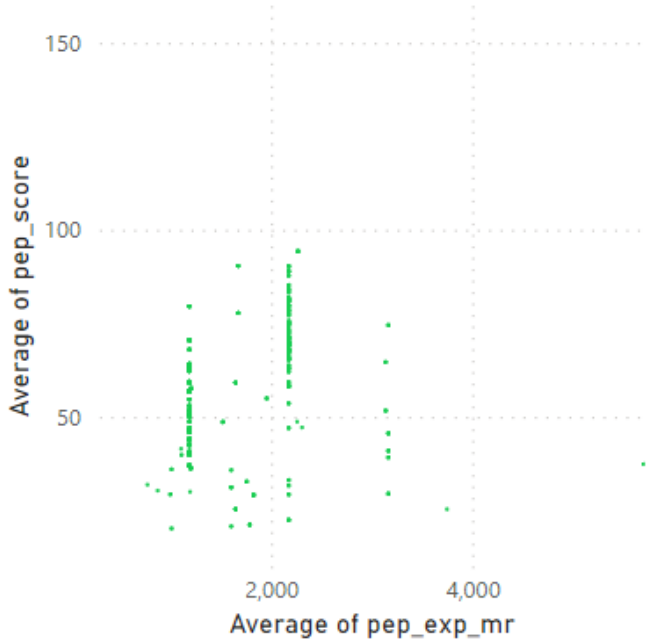
Protein description



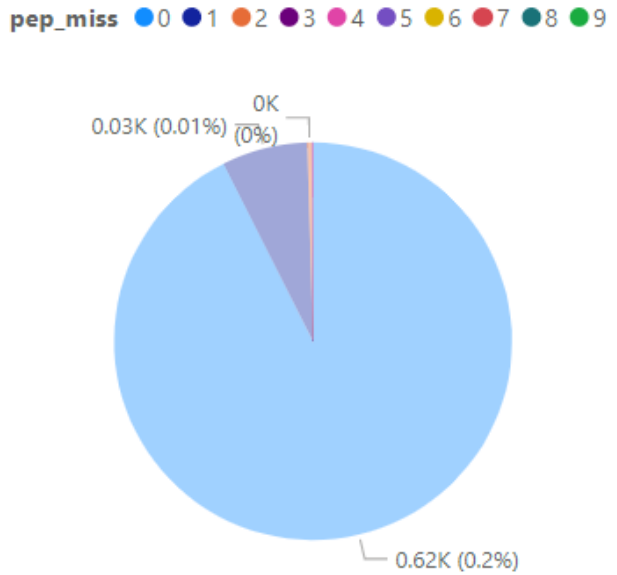
Numerous α-amylase inhibitors preventing premature starch digestion.

Identities charted – drill in on deamidation

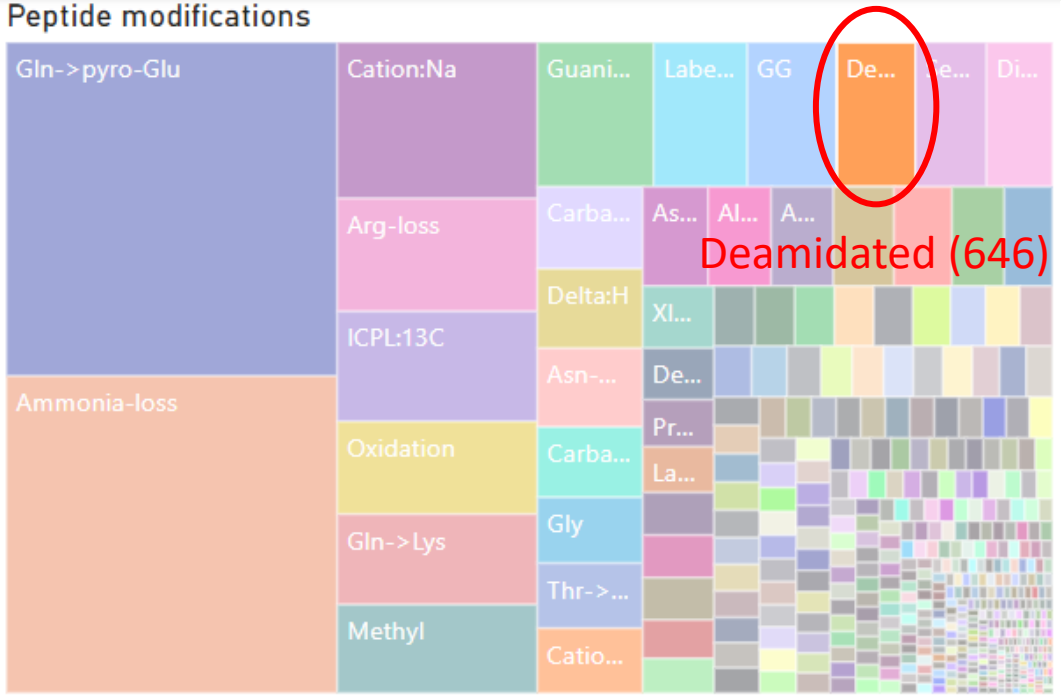
Peptide mass by score



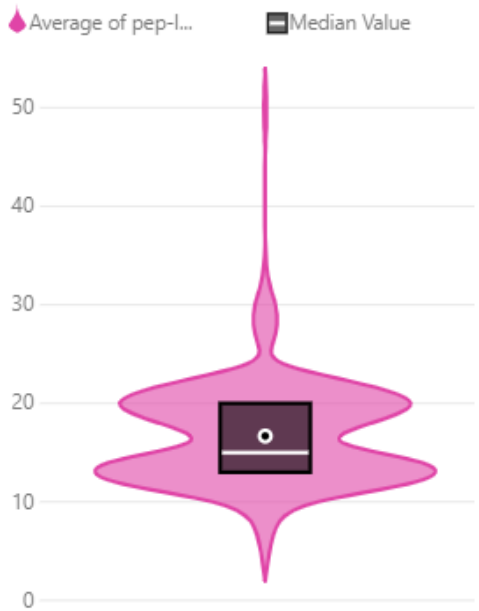
Missed cleavage



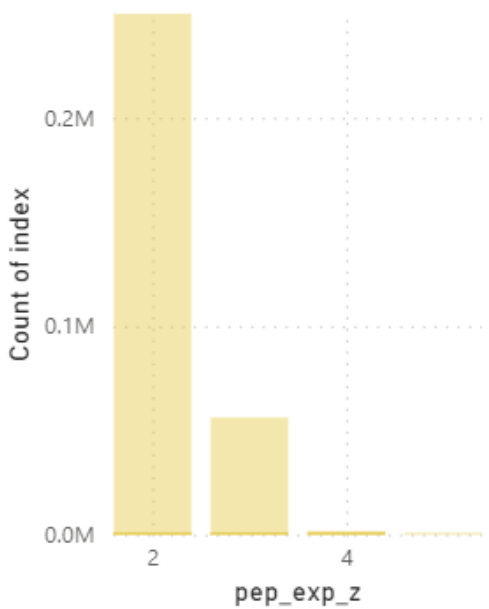
Peptide modifications



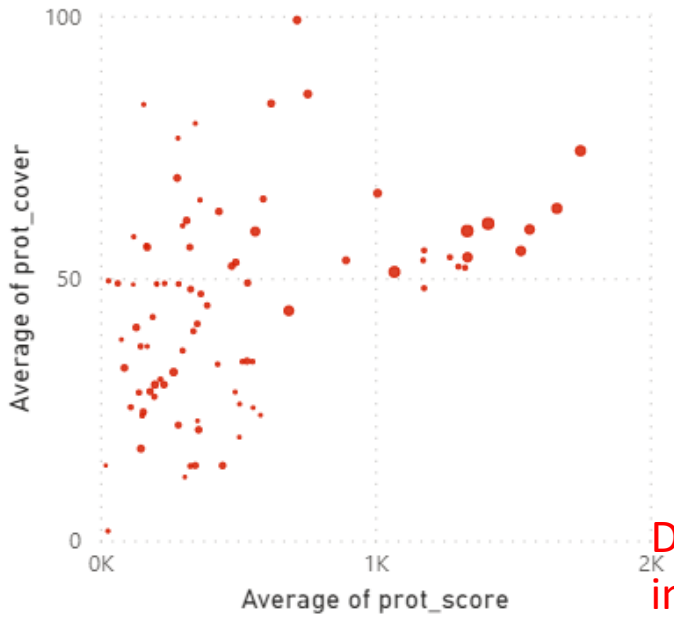
Peptide length (AA)



Peptide charge

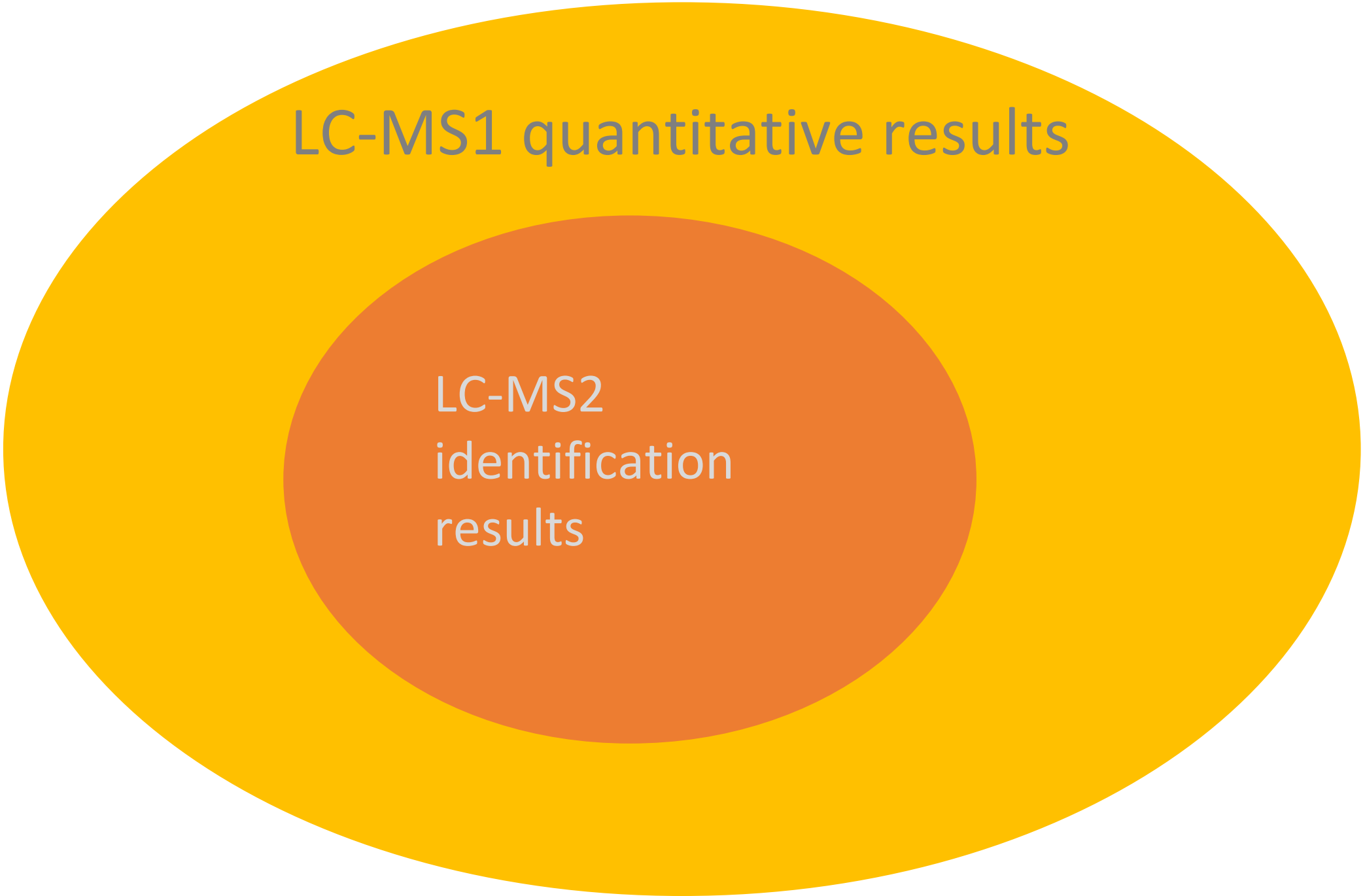


Protein score by sequence coverage



Protein description

Deamidations of glutamine residues in glutenins have been reported.


LC-MS1 quantitative results
LC-MS2 identification results

Linking quantities to identities

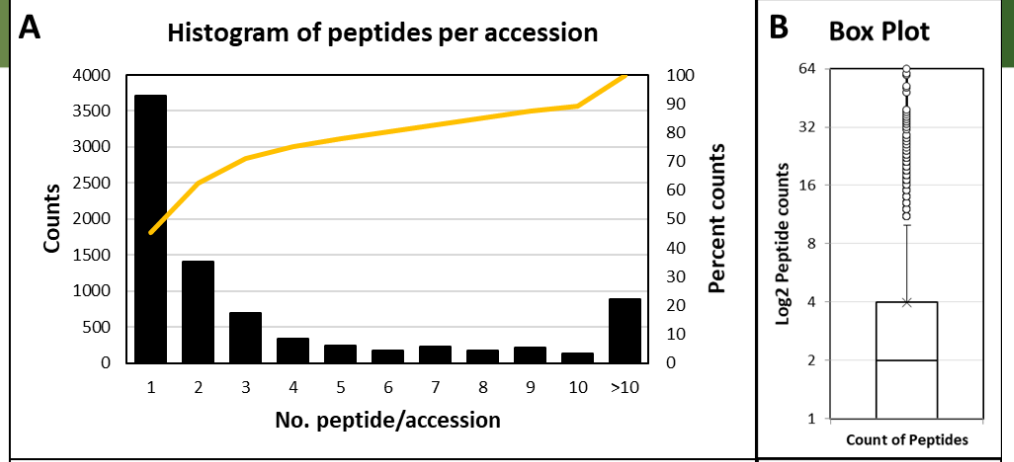
29,908 clusters from 63 LC-MS2 files had to be matched to 32,336 clusters from 3990 LC-MS1 files

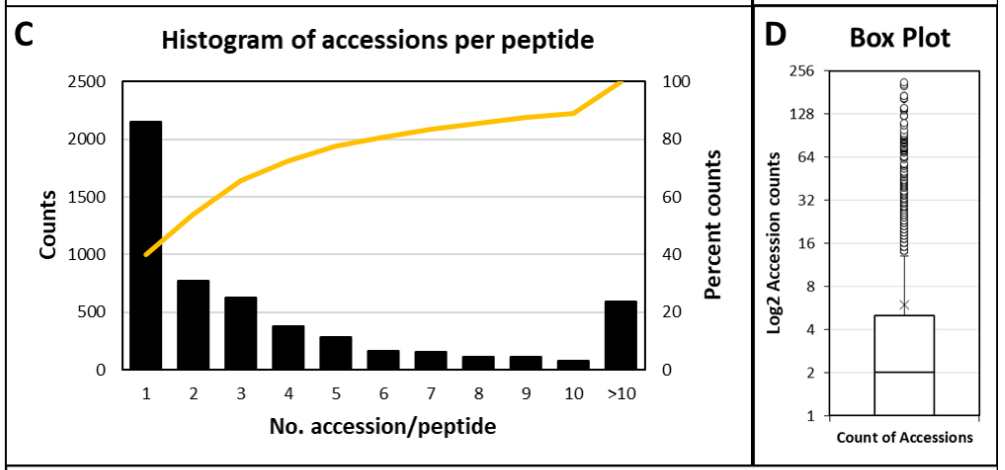
Parameters used:

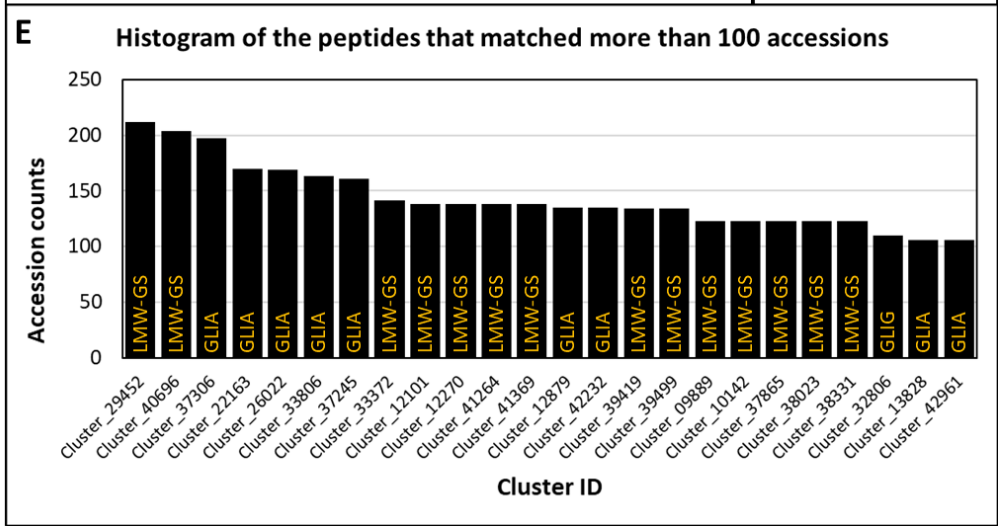
- m/z (20 ppm tolerance)
- Mass (20 ppm tolerance)
- Retention time (RT, 1 min tolerance)

→ 5,414 (17%) clusters with peptide IDs could be linked; they belonged to 8,044 protein accessions

Items quantified	Occurences
Number of wheat genotypes	858
Number of wheat samples	4061
Sampling years	8 (2012-2019)
Trait (LMA)	1
Digestion types	1
Number of reproducible LC-MS1 files	3990
Number of LC-MS1 peaks	137669
Number of reproducible LC-MS1 clusters	32336
Cluster size range	2 - 10
Cluster charge range	2 - 7
Cluster m/z range	300.13 - 1921.55
Cluster mass range	598.26 - 6527.06
Base peak range	120 - 520083
Number of clusters with peptide identity	5414
Number of identified accessions	8044
Range of peptides/accession	1 - 64
Range of accessions/peptide	1 - 212



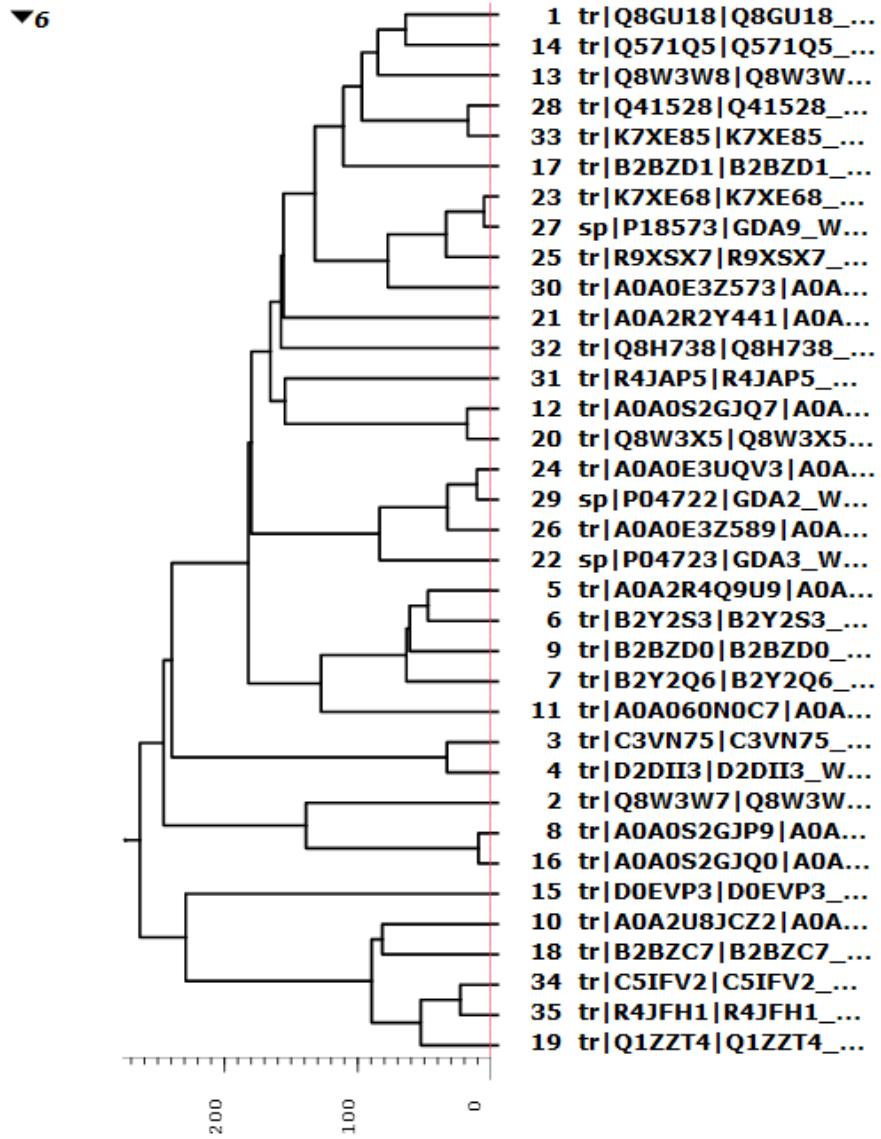




Homeologous protein accessions – protein level

Lots of homeologous protein accessions due to genome polyploidisation.

Example: gliadin/glutenin.



1	tr|Q8GU18|Q8GU18_...	44511	tr|Q8GU18|Q8GU18_WHEAT Low molecular weight glutenin subunit OS=Triticum aestivum OX=4565 GN=lmw-gs PE=4 SV=1
14	tr|Q571Q5|Q571Q5_...	11792	tr|Q571Q5|Q571Q5_WHEAT Putative LMW-glutenin subunit OS=Triticum aestivum OX=4565 GN=glu PE=2 SV=1
13	tr|Q8W3W8|Q8W3W...	14878	tr|Q8W3W8|Q8W3W8_WHEAT Low-molecular-weight glutenin subunit group 3 type II (Fragment) OS=Triticum aestivum OX=4565 GN=lmw-gs PE=4 SV=1
28	tr|Q41528|Q41528_...	1544	tr|Q41528|Q41528_WHEAT Alpha-gliadin OS=Triticum aestivum OX=4565 PE=4 SV=1
33	tr|K7XE85|K7XE85_...	310	tr|K7XE85|K7XE85_WHEAT Alpha-gliadin OS=Triticum aestivum OX=4565 GN=gli-2 PE=4 SV=1
17	tr|B2BZD1|B2BZD1_...	9560	tr|B2BZD1|B2BZD1_WHEAT LMW-D8 OS=Triticum aestivum OX=4565 GN=Glu-D3 PE=4 SV=1
23	tr|K7XE68|K7XE68_...	2471	tr|K7XE68|K7XE68_WHEAT Alpha-gliadin OS=Triticum aestivum OX=4565 GN=gli-2 PE=4 SV=1
27	sp|P18573|GDA9_W...	1731	sp|P18573|GDA9_WHEAT Alpha/beta-gliadin MM1 OS=Triticum aestivum OX=4565 PE=1 SV=1
25	tr|R9XSX7|R9XSX7_...	1891	tr|R9XSX7|R9XSX7_WHEAT Alpha-gliadin OS=Triticum aestivum OX=4565 GN=gli-2 PE=4 SV=1
30	tr|A0A0E3Z573|A0A...	1443	tr|A0A0E3Z573|A0A0E3Z573_WHEAT Alpha-gliadin (Fragment) OS=Triticum aestivum OX=4565 PE=4 SV=1
21	tr|A0A2R2Y441|A0A...	6573	tr|A0A2R2Y441|A0A2R2Y441_WHEAT CSB_alpha gliadin 9 OS=Triticum aestivum OX=4565 GN=Gli-2 PE=4 SV=1
32	tr|Q8H738|Q8H738_...	430	tr|Q8H738|Q8H738_WHEAT Gamma-gliadin (Fragment) OS=Triticum aestivum OX=4565 PE=2 SV=1
31	tr|R4JAP5|R4JAP5_...	468	tr|R4JAP5|R4JAP5_WHEAT Low-molecular-weight glutenin subunit (Fragment) OS=Triticum aestivum OX=4565 GN=LMW-GS PE=4 SV=1
12	tr|A0A0S2GJQ7|A0A...	15749	tr|A0A0S2GJQ7|A0A0S2GJQ7_WHEAT Low-molecular-weight glutenin subunit OS=Triticum aestivum OX=4565 GN=LMW-GS PE=4 SV=1
20	tr|Q8W3X5|Q8W3X5...	6621	tr|Q8W3X5|Q8W3X5_WHEAT Low-molecular-weight glutenin subunit group 2 type I OS=Triticum aestivum OX=4565 GN=lmw-gs PE=4 SV=1
24	tr|A0A0E3UQV3|A0A...	2084	tr|A0A0E3UQV3|A0A0E3UQV3_WHEAT Alpha-gliadin (Fragment) OS=Triticum aestivum OX=4565 PE=4 SV=1
29	sp|P04722|GDA2_W...	1477	sp|P04722|GDA2_WHEAT Alpha/beta-gliadin A-II OS=Triticum aestivum OX=4565 PE=2 SV=1
26	tr|A0A0E3Z589|A0A...	1773	tr|A0A0E3Z589|A0A0E3Z589_WHEAT Alpha-gliadin (Fragment) OS=Triticum aestivum OX=4565 PE=4 SV=1
22	sp|P04723|GDA3_W...	3520	sp|P04723|GDA3_WHEAT Alpha/beta-gliadin A-III OS=Triticum aestivum OX=4565 PE=2 SV=1
5	tr|A0A2R4Q9U9|A0A...	19487	tr|A0A2R4Q9U9|A0A2R4Q9U9_WHEAT Low-molecular-weight glutenin subunit OS=Triticum aestivum OX=4565 GN=LMW-GS PE=4 SV=1
6	tr|B2Y2S3|B2Y2S3_...	18849	tr|B2Y2S3|B2Y2S3_WHEAT Low molecular weight glutenin subunit OS=Triticum aestivum OX=4565 GN=GluB3-2 PE=4 SV=1
9	tr|B2BZD0|B2BZD0_...	17826	tr|B2BZD0|B2BZD0_WHEAT LMW-s glutenin subunit 0359D24-S OS=Triticum aestivum OX=4565 GN=Glu-D3 PE=4 SV=1
7	tr|B2Y2Q6|B2Y2Q6_...	18123	tr|B2Y2Q6|B2Y2Q6_WHEAT LMW-B2 OS=Triticum aestivum OX=4565 GN=GluB3-2 PE=4 SV=1
11	tr|A0A060N0C7|A0A...	16900	tr|A0A060N0C7|A0A060N0C7_WHEAT Low molecular weight glutenin subunit (Fragment) OS=Triticum aestivum OX=4565 GN=glu PE=2 SV=1
3	tr|C3VN75|C3VN75_...	23258	tr|C3VN75|C3VN75_WHEAT Low molecular weight glutenin OS=Triticum aestivum OX=4565 GN=Glu-A3 PE=4 SV=1
4	tr|D2DII3|D2DII3_W...	19585	tr|D2DII3|D2DII3_WHEAT Low-molecular-weight glutenin subunit OS=Triticum aestivum OX=4565 GN=GluA3-16 PE=4 SV=1
2	tr|Q8W3W7|Q8W3W...	34685	tr|Q8W3W7|Q8W3W7_WHEAT Low-molecular-weight glutenin subunit group 3 type II (Fragment) OS=Triticum aestivum OX=4565 GN=lmw-gs PE=4 SV=1
8	tr|A0A0S2GJP9|A0A...	17855	tr|A0A0S2GJP9|A0A0S2GJP9_WHEAT Low-molecular-weight glutenin subunit OS=Triticum aestivum OX=4565 GN=LMW-GS PE=4 SV=1
16	tr|A0A0S2GJQ0|A0A...	10651	tr|A0A0S2GJQ0|A0A0S2GJQ0_WHEAT Low-molecular-weight glutenin subunit OS=Triticum aestivum OX=4565 GN=LMW-GS PE=4 SV=1
15	tr|D0EVP3|D0EVP3_...	11274	tr|D0EVP3|D0EVP3_WHEAT LMW-m glutenin subunit OS=Triticum aestivum OX=4565 PE=4 SV=1
10	tr|A0A2U8JCZ2|A0A...	17108	tr|A0A2U8JCZ2|A0A2U8JCZ2_WHEAT LMW-B3 OS=Triticum aestivum OX=4565 GN=Glu-3 PE=4 SV=1
18	tr|B2BZC7|B2BZC7_...	8408	tr|B2BZC7|B2BZC7_WHEAT LMW-m glutenin subunit 0154A5-M OS=Triticum aestivum OX=4565 GN=Glu-D3 PE=4 SV=1
34	tr|C5IFV2|C5IFV2_...	85	tr|C5IFV2|C5IFV2_WHEAT Low molecular weight protein (Fragment) OS=Triticum aestivum OX=4565 PE=4 SV=1
35	tr|R4JFH1|R4JFH1_...	42	tr|R4JFH1|R4JFH1_WHEAT Low-molecular-weight glutenin subunit (Fragment) OS=Triticum aestivum OX=4565 GN=LMW-GS PE=4 SV=1
19	tr|Q1ZZT4|Q1ZZT4_...	8183	tr|Q1ZZT4|Q1ZZT4_WHEAT Low-molecular-weight glutenin subunit OS=Triticum aestivum OX=4565 PE=4 SV=1

Homeologous protein accessions – peptide level

Example:
Peptide “VLQQLNPCK” (Cluster_29452) matches 212 different protein accessions of “Low Molecular Weight Glutenin Subunit”.

However, anything goes as far as naming is concerned
from vey concise
“**LMWGS1**” (6 characters)
to lengthy
“**LMW-D7 (LMW-glutenin P3-42) (LMW-m glutenin subunit 0275P20-M) (Low molecular weight glutenin) (Low-molecular-weight glutenin subunit) (Low-molecular-weight glutenin subunit group 7 type IV**” (190 characters).

Protein annotations in databases ought to be tidied up and a naming convention agreed upon.

Protein description	No characters
Glutenin, low molecular weight subunit PTDUCD1	46
Low-molecular-weight glutenin subunit	37
HMW glutenin i-type subunit 3A	30
Low molecular weight glutenin subunit P-14	42
Low molecular weight glutenin subunit P-13	42
Low molecular weight glutenin subunit P-15	42
Low molecular weight glutenin subunit A3-8	42
Low molecular weight glutenin subunit A3-3	42
Low molecular weight glutenin subunit A3-6	42
LMWGS1	6
LMWGS2	6
LMW-glutenin	12
Low molecular weight glutinin subunit	37
Low molecular weight glutenin	29
LMW-m glutenin subunit	22
LMW-m glutenin subunit (Low-molecular-weight glutenin subunit)	62
Low molecular weight glutinen subunit	37
Low molecular weight glutenin subunit A3-4 (Low-molecular-weight glutenin subunit)	82
LMW-m glutenin subunit (Fragment)	33
Low molecular weight glutenin subunit A3-2	42
Low molecular weight glutenin subunit D3-7	42
Low molecular weight glutenin subunit A3-2 (Low-molecular-weight glutenin subunit) (Fragment)	93
LMW glutenin subunit Glu-D3 (Fragment)	38
Low-molecular-weight glutenin subunit (Fragment)	48
Low-molecular-weight glutenin subunit LMW-H6-5-2	48
LMW-m glutenin subunit 9	24
LMW-m glutenin subunit 14	25
LMW-m glutenin subunit 18	25
LMW-GS (Fragment)	17
Low molecular weight glutenin subunit (Fragment)	48
Low molecular weight glutenin subunit t128	42
Low molecular weight glutenin subunit t128 (Fragment)	53
Low-molecular-weight glutenin subunit (Low-molecular-weight glutenin subunit Glu-B3)	84
Low molecular weight glutenin subunit LMW-10	44
Low molecular weight glutenin subunit	37
LMM glutenin 1 (Fragment)	25
Low-molecular-weight glutenin storage protein	45
Low-molecular-weight glutenin storage protein (Low-molecular-weight glutenin subunit)	85
LMW-D6 (LMW-GS) (LMW-GS P-11) (LMW-m glutenin subunit 3) (LMW-m glutenin subunit 51) (Low molecular weight glutenin subunit D3-6) (Low-molecular-weight glutenin subunit)	169
Low molecular weight glutenin subunit K1	40
Low molecular weight glutenin subunit GF-1	42
LMW-glutenin P3-41 (Low-molecular-weight glutenin subunit)	58
Low-molecular-weight glutenin subunit Glu-A3 (Fragment)	55
Low molecular weight glutenin subunit Glu-A3 (Fragment)	55
(T.aestivum) gamma-gliadin class B-I (Fragment)	47
Low molecular weight glutenin (Fragment)	40
LMW-GS (LMW-GS P-12) (LMW-m glutenin subunit 1238L16-M) (Low molecular weight glutenin) (Low-molecular-weight glutenin subunit)	127
LMW-i glutenin pGH3.1	21
LMW-GS (LMW-i glutenin subunit 1594F5-I) (Low molecular weight glutenin) (Low-molecular-weight glutenin subunit)	112
LMW-GS (LMW-i glutenin subunit 19) (Low molecular weight glutenin) (Low-molecular-weight glutenin subunit) (Low-molecular-weight glutenin subunit Glu-A3)	153

Conversion of wide tables into long tables

Our strategy was to consider all 8,044 protein hits identified from the 5,414 sequenced peptides irrespective of their homology. Wide table (5414 identified peptides x up to 212 accessions) converted into a long table (32347 protein accessions).

	A	B	C	D	E	F	G	H	I	J	K
1	Cluster	Peptide	Accession1	Accession2	Accession3	Accession4	Accession5	Accession6	Accession7	Accession8	Accession9
2	Cluster_3547	APGDHITRDE	tr|A0A3B6BZY3|A0A3B6BZY3_WHEAT								
3	Cluster_3667	DLGTGAALVA	tr|A0A3B6ECI4|A0A3B6ECI4_WHEAT								
4	Cluster_2701	HASARGVRR	tr|A0A3B6NY58|A0A3B6NY58_WHEAT								
5	Cluster_2794	GPTSGSSASR	tr|A0A3B6QLI9|A0A3B6QLI9_WHEAT								
6	Cluster_3985	LAYVALDYEQE	tr|A0A067YN	tr|A0A075D1	tr|A0A075D1	tr|A0A1D5X5	tr|A0A3B5XT	tr|A0A3B5XY	tr|A0A3B6HZ	tr|Q5EWZ1|C	tr|Q6TDU4|C
7	Cluster_0780	MTRHRNLIK	tr|A0A3B6LRZ1|A0A3B6LRZ1_WHEAT								
8	Cluster_2867	RSAAKISASVA	tr|A0A3B6AVI2|A0A3B6AVI2_WHEAT								
9	Cluster_2681	GPGADFGK	tr|A0A3B6JDM3|A0A3B6JDM3_WHEAT								
10	Cluster_2614	SIAGIVTER	tr|A0A3B6HX	tr|A0A3B6INI	tr|A0A3B6JDG4|A0A3B6JDG4_WHEAT						
11	Cluster_0874	HGEGEREEEQ	tr|B7U6L4|B	tr|I6QQ39|I6QQ39_WHEAT							
12	Cluster_1140	RRDYSSAAASF	tr|A0A3B6A1V2|A0A3B6A1V2_WHEAT								
13	Cluster_0813	KSNGDNTDHF	tr|A0A3B5ZVM1|A0A3B5ZVM1_WHEAT								
14	Cluster_0205	EMQDTALAYIK	tr|A0A3B6MZ81|A0A3B6MZ81_WHEAT								
15	Cluster_1317	LLAGVTIAHGG	sp|P02275|H	sp|Q43213|H	sp|Q43214|H	tr|A0A1D6AV	tr|A0A3B5Z2	tr|A0A3B6EF	tr|A0A3B6KS	tr|A0A3B6LV	tr|A0A3B6MZ
16	Cluster_3213	YGMDDYLEIK	tr|A0A3B6NL	tr|A0A3B6NN	tr|A0A3B6NP	tr|A0A3B6PJI	tr|A0A3B6PJI	tr|A0A3B6PL	tr|A0A3B6QC	tr|A0A3B6QE	tr|A0A3B6QF
17	Cluster_0403	AWGLDKSRQI	tr|A0A3B6R9W5|A0A3B6R9W5_WHEAT								
18	Cluster_3598	ANNSSNSAAA	tr|A0A3B6I248|A0A3B6I248_WHEAT								
19	Cluster_4147	SSYYFSNKTILD	tr|A0A077RP	tr|A0A3B6EE41|A0A3B6EE41_WHEAT							
20	Cluster_3015	TARRTSSSR	tr|A0A3B6PK	tr|A0A3B6PKA6|A0A3B6PKA6_WHEAT							
21	Cluster_1525	NRQPEITSLKR	tr|A0A3B6LT11|A0A3B6LT11_WHEAT								
22	Cluster_4368	GFYTYDAFVAA	tr|A0A3B6SU01|A0A3B6SU01_WHEAT								
23	Cluster_0351	GFYTYDAFVAA	tr|A0A3B6TZD9|A0A3B6TZD9_WHEAT								
24	Cluster_1152	DQAGAGALLH	tr|A0A1D5UZ	tr|A0A3B6BZ	tr|A0A3B6BZ	tr|A0A3B6C1	tr|A0A3B6D6	tr|A0A3B6D9H7|A0A3B6D9H7_WHEAT			
25	Cluster_3251	NELESTGSHAL	tr|A0A3B6ER	tr|A0A3B6ET98|A0A3B6ET98_WHEAT							
26	Cluster_1813	TNNTSKSVMR	tr|A0A3B6TS	tr|A0A3B6TY	tr|A0A3B6TYR2|A0A3B6TYR2_WHEAT						
27	Cluster_3624	ATVPAPAAETA	tr|W5FWF5|W5FWF5_WHEAT								
28	Cluster_0856	QQMADAVTAI	tr|A0A1D5ZL	tr|A0A3B6LV	tr|A0A3B6LW	tr|A0A3B6LW	tr|A0A3B6LW	tr|A0A3B6N0	tr|Q07810|Q07810_WHEAT		
29	Cluster_1856	ADPGRLLPHR	tr|A0A3B6PH00|A0A3B6PH00_WHEAT								
30	Cluster_2744	ALAFPQQAR	tr|A0A3B6LU	tr|A0A3B6MYZ0|A0A3B6MYZ0_WHEAT							
31	Cluster_3033	VGSSTDLIAQR	tr|A0A096US	tr|A0A3B6KP	tr|A0A3B6N018|A0A3B6N018_WHEAT						
32	Cluster_2690	ARALLPQR	tr|A0A3B6KNG3|A0A3B6KNG3_WHEAT								
33	Cluster_3264	GQCSDAYSYPI	tr|A0A3B6HX	tr|A0A3B6IM	tr|A0A3B6JDF8|A0A3B6JDF8_WHEAT						
34	Cluster_3254	NTYGWSISVD	tr|A0A3B6HV	tr|A0A3B6IQ	tr|A0A3B6JHB6|A0A3B6JHB6_WHEAT						
35	Cluster_3575	CGCAVPCPGG	sp|P30569|EC1_WHEAT								
36	Cluster_4064	GMNADYGAP	tr|A0A3B6ML	tr|A0A3B6MN60|A0A3B6MN60_WHEAT							
37	Cluster_1061	VRTFPDLSSTA	tr|A0A3B6MXB0|A0A3B6MXB0_WHEAT								
38	Cluster_1270	KGSEERLMVLI	tr|A0A2Z6ERU2|A0A2Z6ERU2_WHEAT								
39	Cluster_1325	LLAGITIAHGG	tr|A0A3B6NK	tr|A0A3B6PK	tr|A0A3B6QBP3|A0A3B6QBP3_WHEAT						
40	Cluster_1044	TAEKEPKSPKK	tr|A0A3B6LW24|A0A3B6LW24_WHEAT								
41	Cluster_3373	YAGQEVFSGST	tr|A0A3B6RA	tr|A0A3B6TEU3|A0A3B6TEU3_WHEAT							
42	Cluster_1044	YFEVILVDVAH	tr|A0A1D5UY	tr|A0A3B6AV	tr|W5FEX7|W5FEX7_WHEAT						
43	Cluster_0996	LLPGGGASELT	tr|A0A3B6RH	tr|A0A3B6SG	tr|A0A3B6SM	tr|A0A3B6TG	tr|A0A3B6TLG9|A0A3B6TLG9_WHEAT				
44	Cluster_2577	IMDYFIK	tr|A0A3B6NJ	tr|R9W6A6|R	tr|R9W924|R9W924_WHEAT						
45	Cluster_4036	HRLLGLRTAM	tr|A0A3B6AUT2|A0A3B6AUT2_WHEAT								
46	Cluster_1408	LTSPQQSGQG	sp|P08489|GLT4_WHEAT								

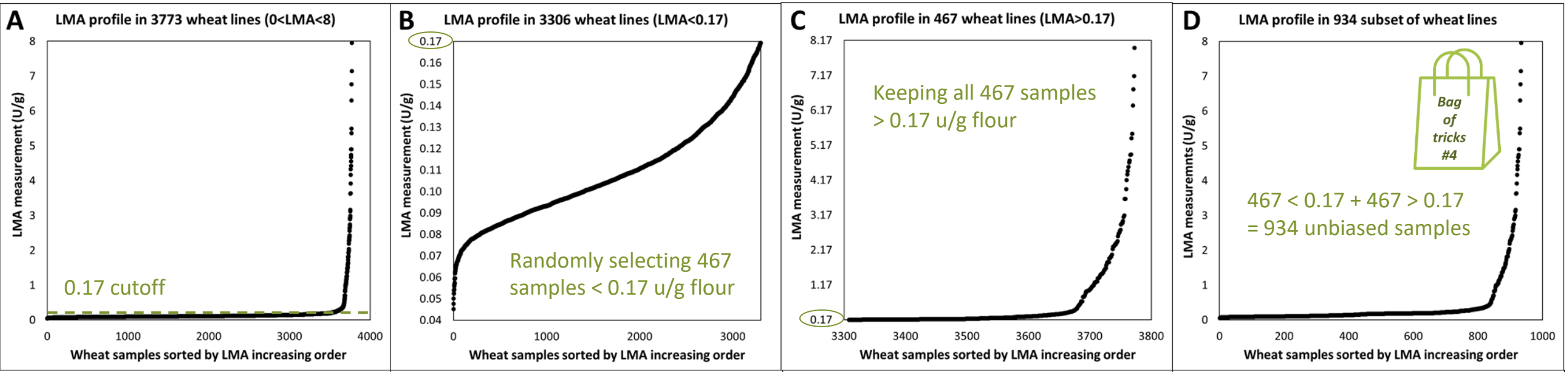
	A	B	C
1	Cluster	Peptide	Accessions
2	Cluster_35477'	APGDHITRDETK	tr|A0A3B6BZY3|A0A3B6BZY3_WHEAT
3	Cluster_36671'	DLGTGAALVAGLLAMK	tr|A0A3B6ECI4|A0A3B6ECI4_WHEAT
4	Cluster_27016'	HASARGVRR	tr|A0A3B6NY58|A0A3B6NY58_WHEAT
5	Cluster_27945'	GPTSGSSASR	tr|A0A3B6QLI9|A0A3B6QLI9_WHEAT
6	Cluster_39858'	LAYVALDYEQELETAK	tr|A0A067YNJ5|A0A067YNJ5_WHEAT
7	Cluster_39858'	LAYVALDYEQELETAK	tr|A0A075D1A6|A0A075D1A6_WHEAT
8	Cluster_39858'	LAYVALDYEQELETAK	tr|A0A075D1U0|A0A075D1U0_WHEAT
9	Cluster_39858'	LAYVALDYEQELETAK	tr|A0A1D5X5C6|A0A1D5X5C6_WHEAT
10	Cluster_39858'	LAYVALDYEQELETAK	tr|A0A3B5XT63|A0A3B5XT63_WHEAT
11	Cluster_39858'	LAYVALDYEQELETAK	tr|A0A3B5XYX6|A0A3B5XYX6_WHEAT
12	Cluster_39858'	LAYVALDYEQELETAK	tr|A0A3B6HZ58|A0A3B6HZ58_WHEAT
13	Cluster_39858'	LAYVALDYEQELETAK	tr|Q5EWZ1|Q5EWZ1_WHEAT
14	Cluster_39858'	LAYVALDYEQELETAK	tr|Q6TDU4|Q6TDU4_WHEAT
15	Cluster_39858'	LAYVALDYEQELETAK	tr|W5AH12|W5AH12_WHEAT
16	Cluster_07804'	MTRHRNLIK	tr|A0A3B6LRZ1|A0A3B6LRZ1_WHEAT
17	Cluster_28671'	RSAAKISASVA	tr|A0A3B6AVI2|A0A3B6AVI2_WHEAT
18	Cluster_26819'	GPGADFGK	tr|A0A3B6JDM3|A0A3B6JDM3_WHEAT
19	Cluster_26146'	SIAGIVTER	tr|A0A3B6HXM3|A0A3B6HXM3_WHEAT
20	Cluster_26146'	SIAGIVTER	tr|A0A3B6INL0|A0A3B6INL0_WHEAT
21	Cluster_26146'	SIAGIVTER	tr|A0A3B6JDG4|A0A3B6JDG4_WHEAT
22	Cluster_08748'	HGEGEREEEQGR	tr|B7U6L4|B7U6L4_WHEAT
23	Cluster_08748'	HGEGEREEEQGR	tr|I6QQ39|I6QQ39_WHEAT
24	Cluster_11409'	RRDYSSAAASPER	tr|A0A3B6A1V2|A0A3B6A1V2_WHEAT
25	Cluster_08139'	KSNGDNTDHR	tr|A0A3B5ZVM1|A0A3B5ZVM1_WHEAT
26	Cluster_02053'	EMQDTALAYIKGQFSWK	tr|A0A3B6MZ81|A0A3B6MZ81_WHEAT
27	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	sp|P02275|H2A1_WHEAT
28	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	sp|Q43213|H2A5_WHEAT
29	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	sp|Q43214|H2A6_WHEAT
30	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A1D6AWG3|A0A1D6AWG3_WHEAT
31	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A3B5Z2B2|A0A3B5Z2B2_WHEAT
32	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A3B6EFC8|A0A3B6EFC8_WHEAT
33	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A3B6KS31|A0A3B6KS31_WHEAT
34	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A3B6LVY4|A0A3B6LVY4_WHEAT
35	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A3B6MZ32|A0A3B6MZ32_WHEAT
36	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A3B6NJA7|A0A3B6NJA7_WHEAT
37	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A3B6NJB2|A0A3B6NJB2_WHEAT
38	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A3B6NJJ7|A0A3B6NJJ7_WHEAT
39	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A3B6NK66|A0A3B6NK66_WHEAT
40	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A3B6NKD7|A0A3B6NKD7_WHEAT
41	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A3B6NL63|A0A3B6NL63_WHEAT
42	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A3B6NL75|A0A3B6NL75_WHEAT
43	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A3B6PDZ2|A0A3B6PDZ2_WHEAT
44	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A3B6PEU0|A0A3B6PEU0_WHEAT
45	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A3B6PFD2|A0A3B6PFD2_WHEAT
46	Cluster_13174'	LLAGVTIAHGGVIPNINSVLLF	tr|A0A3B6PFE4|A0A3B6PFE4_WHEAT

One trait: LMA

LMA activity and creating an unbiased dataset

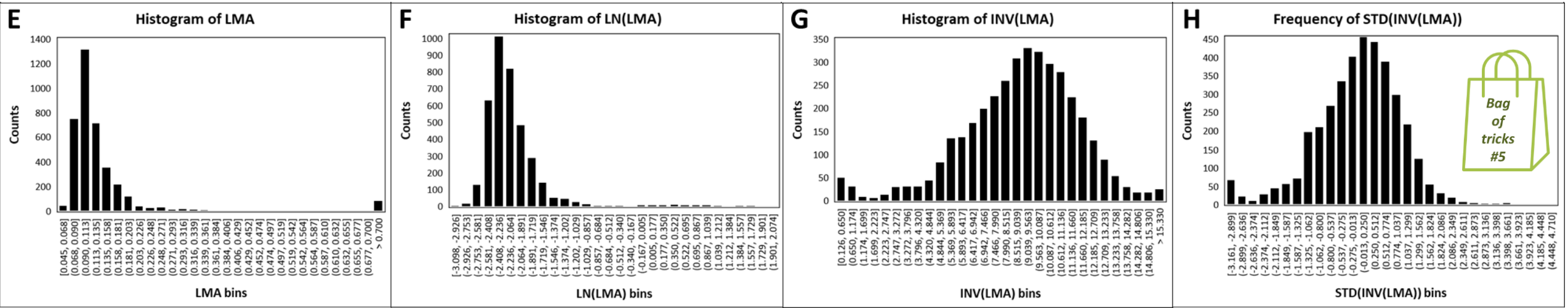
0 < LMA activity < 8 u/g flour but heavily biased towards very small values. Arbitrary cutoff of 0.17 u/g flour to subsample data.

→ **934 samples chosen to create a balanced dataset.**



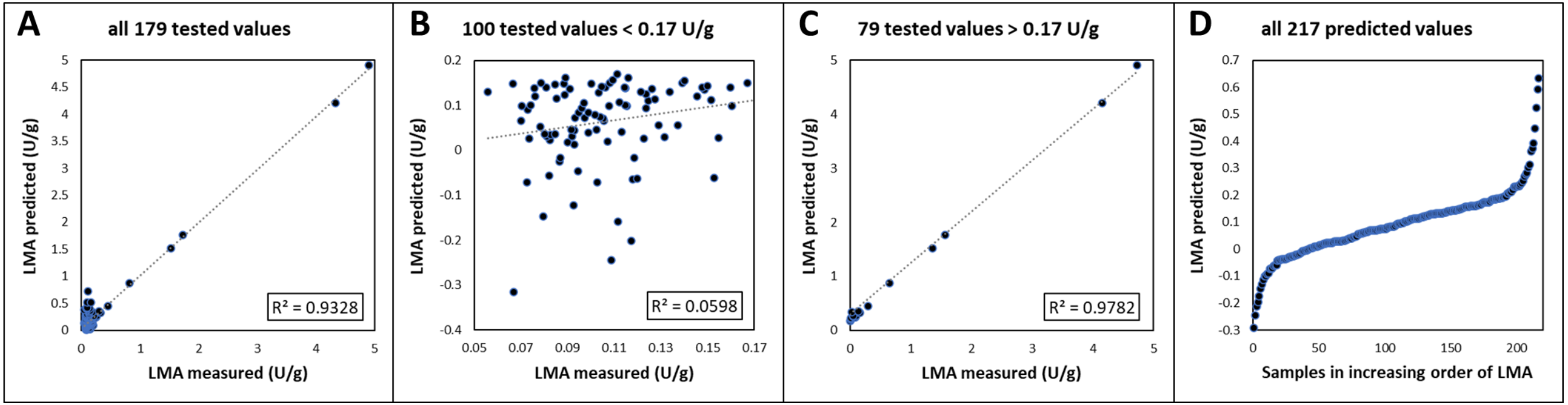
Some stats require normal data.

→ **Inverse or standardised of inverse function to normalise LMA distribution.**



Predicting LMA missing values

217 samples without LMA measurements.
Predicted using a univariate partial least square (PLS) regression model.
→ **LMA missing values predicted with 93% accuracy (98% accuracy for values > 0.17 u/g).**



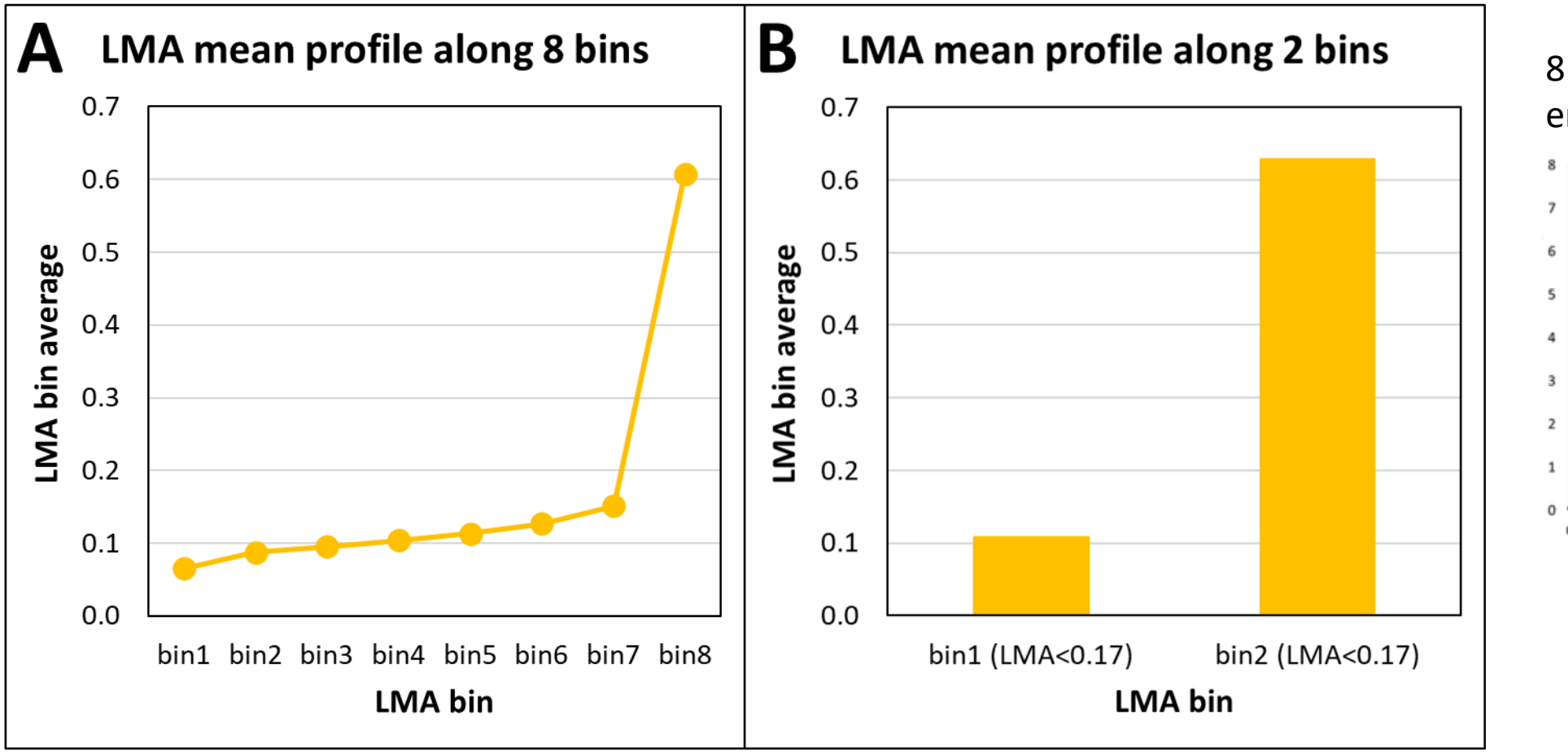
Bag of tricks #6

Incorporating LMA trait at the peptide level for biomarker discovery to detect peptides that behave similarly or conversely to LMA trait (**Cluster_AAA**)
→ **assessed the relevance of the statistical tests carried out by validating anticipated results**
→ **improved biomarker discovery**

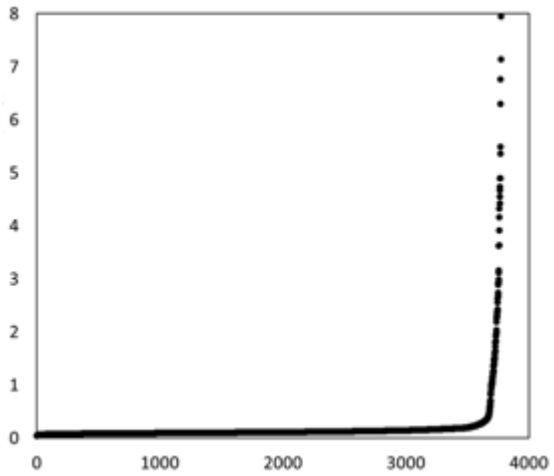
Binning samples to improve LMA representation

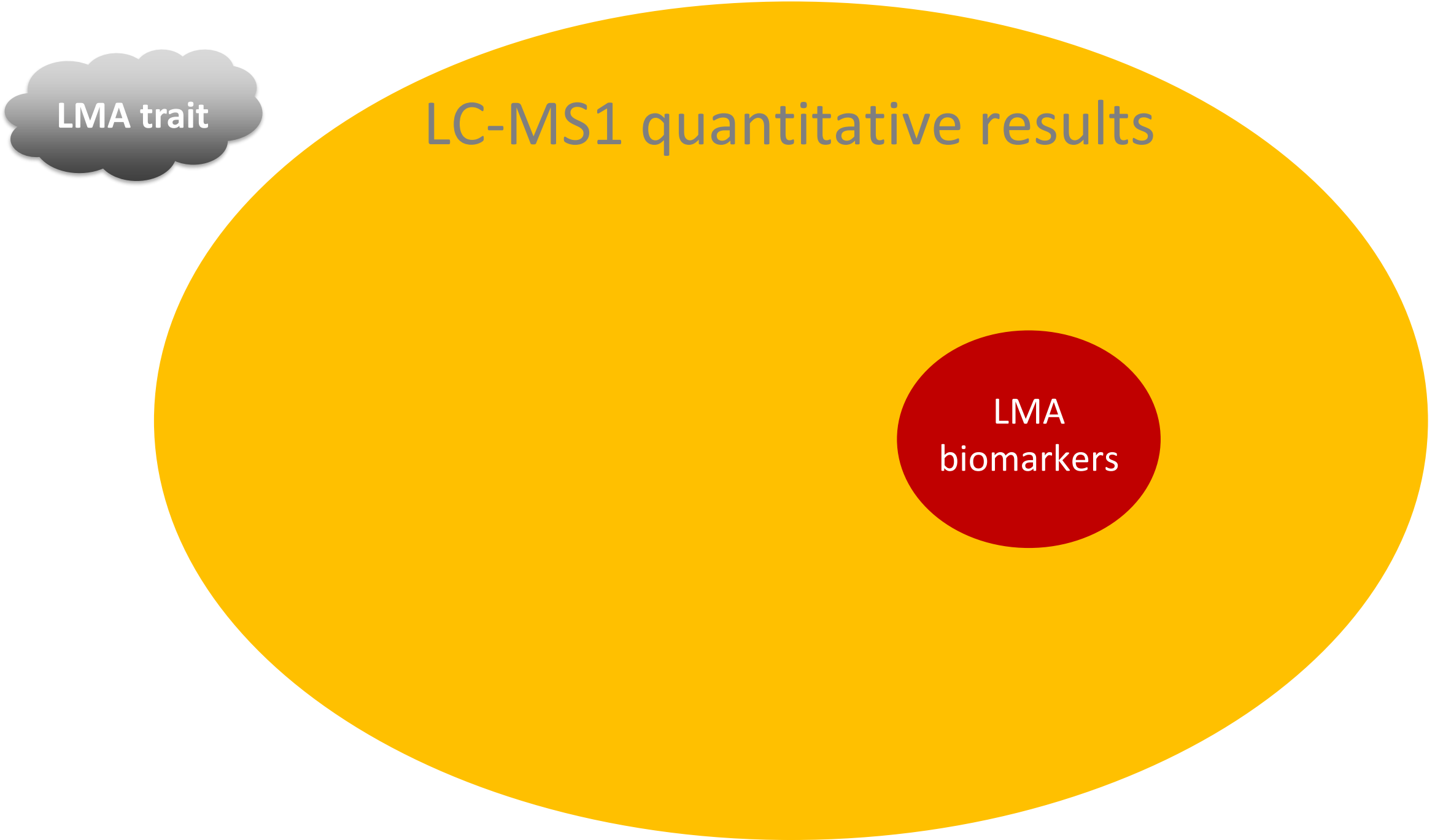
Our data matrix of 3,990 columns by 32,337 rows contained 129,024,630 quantities which posed representation challenges.
We adopted a data reduction strategy involving binning the samples into 8 or 2 arbitrary bins based on their LMA values in order to produce simpler more legible graphs for individual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

- 8 bins = 499 samples each (total 3,990 samples = full dataset)
- 2 bin = 467 samples each (total 934 samples = unbiased dataset)

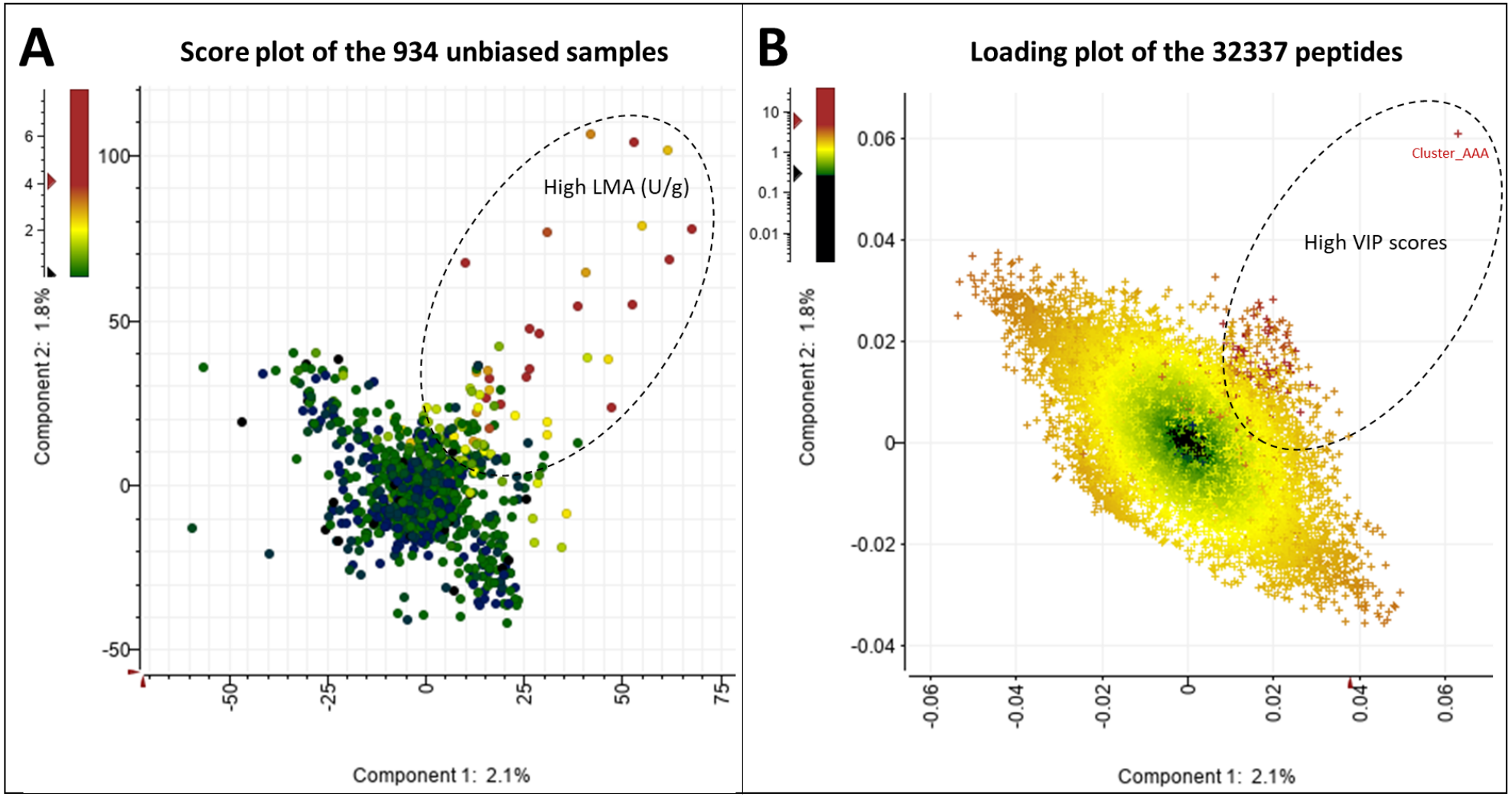


8 bins are enough to emulate LMA distribution.




LMA trait
LC-MS1 quantitative results
LMA biomarkers

Partial Least Square (PLS) analysis with LMA as a response and variable importance in projection (VIP)



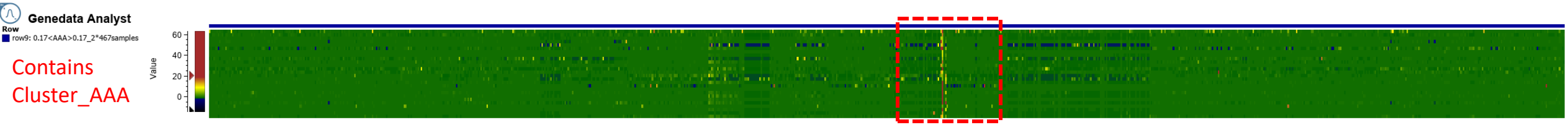
The higher the VIP score, the greater the contribution of the peptide to the PLS and the closer to LMA response.

By setting up 3 VIP score thresholds of increasing stringency, we created three subsets of peptides of decreasing sizes that could be used in more computationally demanding processes.

PLS VIP scores > 1.5 = 2996 (9.3%) peptides used for regression
PLS VIP scores > 1.0 = 7254 (22.4%) peptides used for SOM, HCA, K-Means
PLS VIP scores > 0.5 = 14490 (44.8%) peptides used for linear models and correlation

Self Organising Map (SOM) and heat map

A SOM (Kohonen 1982) is an unsupervised machine learning technique that generates a low-dimensional representation (2 dimensions are usually enough) of a higher dimensional dataset while preserving the topological structure of the data. Heat maps offer an efficient method of visualizing complex quantitative data sets organized as matrices (Key 2012).



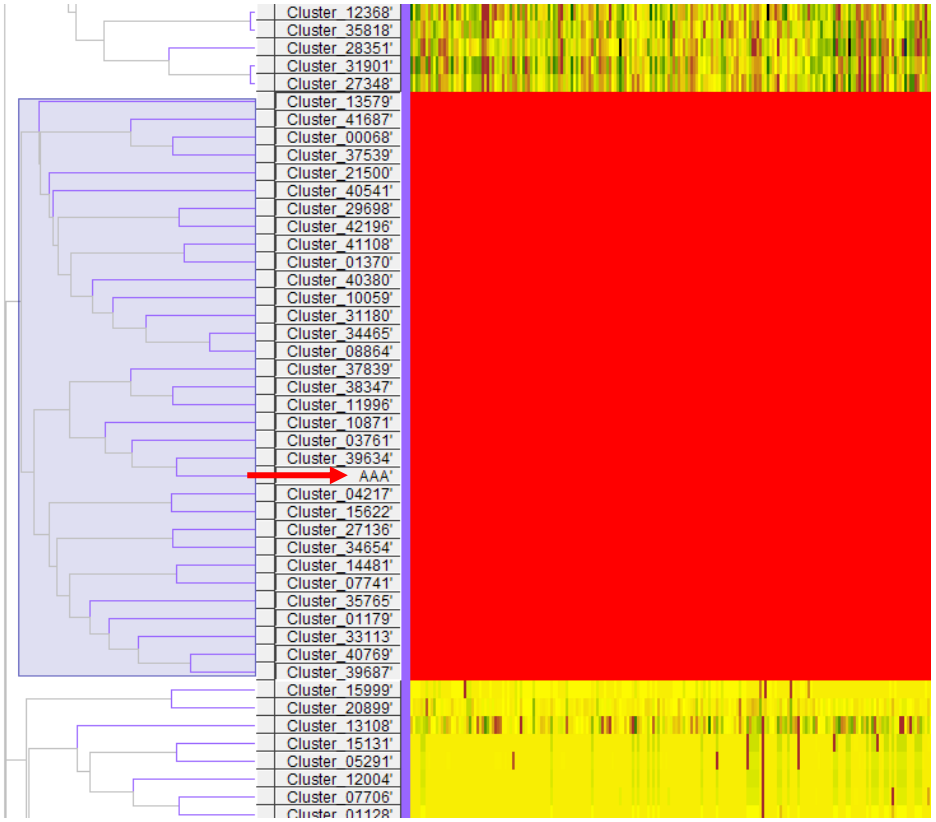
Contains Cluster_AAA

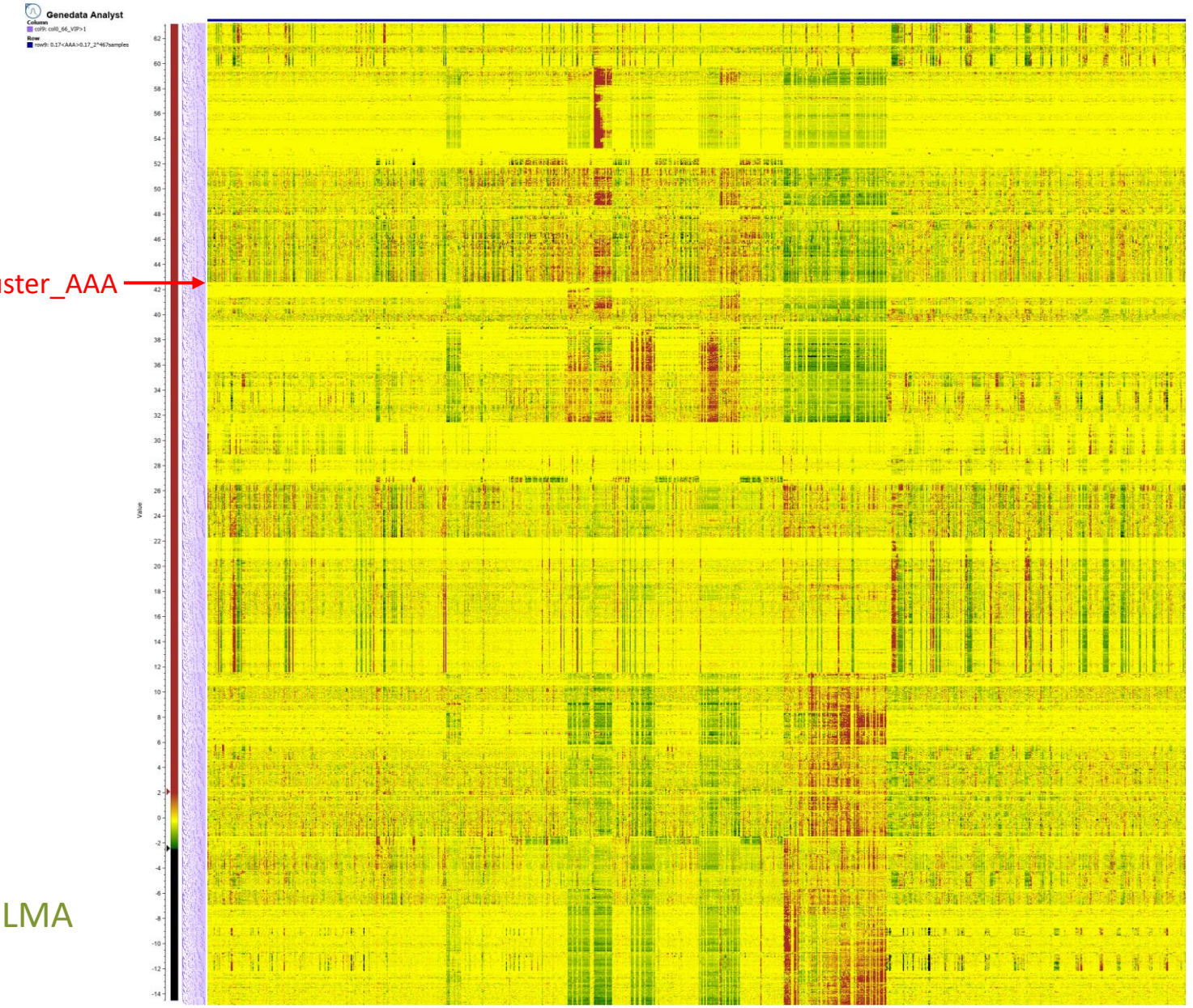
Cluster (4,3) with 26 peptides included Cluster_AAA

→ All these peptides are candidate biomarkers for LMA response

Divisive Hierarchical Clustering Analysis (HCA) with heat map

In divisive HCA, two vectors are randomly initialized and assigned each peptide using a probability function; the vectors are iteratively recalculated to form the centroids of two clusters (Sherlock 2000).





Group with clusters ordered from 1915-1947 (33 peptides) contains Cluster_AAA.

→ All these peptides are candidate biomarkers for LMA response

K-means clustering

K-means is another unsupervised machine learning method. Its ease of implementation, simplicity, computational efficiency, and empirical success make it one of the most popular clustering methods (Jain 2010). n observations are segmented into k clusters in which each observation belongs to the cluster with the nearest mean, thereby minimising within-cluster variances.

77: K-Means_col9_row9_columns_17-20_pos-cor_50iterations_10%valid

k = 17: Cluster Info | k = 18: Cluster Info | k = 19: Cluster Info | k = 20: Cluster Info | Info
Table | Clustering Scores | Cluster Assignments

Column	k = 20: Cluster	k = 20: Distance
Cluster_15481'	14	0.1622
Cluster_15521'	14	0.7928
Uncharacterized protein OS=Triticum aestivum OX=4565 ...	14	0.703
Uncharacterized protein OS=Triticum aestivum OX=4565 ...	14	0.5578
Cluster_16242'	14	0.7187
Cluster_16684'	14	0.4326
Cluster_16741'	14	0.8351
Cluster_17010'	14	0.8866
Cluster_17212'	14	0.8886
Cluster_17593'	14	0.5464
Cluster_17753'	14	0.6505
Cluster_17888'	14	0.7956
Cluster_18319'	14	0.2402
Cluster_19044'	14	0.3499
Cluster_20920'	14	0.5211
Cluster_22157'	14	0.6107
Cluster_22882'	14	0.1242
Cluster_24396'	14	0.3861
Cluster_25103'	14	0.3657
Histone H2B OS=Triticum aestivum OX=4565 GN=CAMPL...	14	0.4274
Cluster_26295'	14	0.5233
Cluster_27044'	14	0.6757
Cluster_27687'	14	0.5688
Cluster_28259'	14	0.2441
Cluster_28653'	14	0.8878
Cluster_29276'	14	0.3371
Cluster_29511'	14	0.52
Cluster_29595'	14	0.7203
Cluster_30417'	14	0.4342
Cluster_30587'	14	0.4544
Cluster_30774'	14	0.7582
Gama-gliadin OS=Triticum aestivum OX=4565 GN=Gli-Ts...	14	0.8936
Cluster_31564'	14	0.3789
Cluster_31999'	14	0.3317
Cluster_32241'	14	0.1552
Cluster_33492'	14	0.1573
Globulin 3 OS=Triticum aestivum OX=4565 GN=glo-3A PE...	14	0.7656
Cluster_34052'	14	0.6972
Cluster_34667'	14	0.5251
Gamma-gliadin OS=Triticum aestivum OX=4565 GN=qp91...	14	0.6047
Cluster_35920'	14	0.3907
Cluster_36524'	14	0.437
27K protein (Fragment) OS=Triticum aestivum OX=4565 G...	14	0.7043
Cluster_37406'	14	0.4142
Cluster_37880'	14	0.4019
Cluster_38087'	14	0.3357
Cluster_38991'	14	0.5964
Cluster_39225'	14	0.3541
Cluster_39518'	14	0.2133
Cluster_39926'	14	0.8983
Cluster_40208'	14	0.1989
Cluster_41099'	14	0.9175
Cluster_41184'	14	0.1632
Cluster_41218'	14	0.4825
Cluster_41939'	14	0.6078
Cluster_42380'	14	0.1623
Uncharacterized protein OS=Triticum aestivum OX=4565	14	0.1617

77: K-Means_col9_row9_columns_17-20_pos-cor_50iterations_10%valid

Table | Clustering Scores | Cluster Assignments | k = 17: Cluster Info | k = 18: Cluster Info | k = 19: Cluster Info | k = 20: Cluster Info | Info

Cluster ID	Cluster Type	Size	Variance
Cluster 1	Column	435	0.2982
Cluster 2	Column	602	0.3169
Cluster 3	Column	240	0.3661
Cluster 4	Column	357	0.3641
Cluster 5	Column	630	0.154
Cluster 6	Column	398	0.2217
Cluster 7	Column	392	0.3599
Cluster 8	Column	491	0.259
Cluster 9	Column	193	0.3787
Cluster 10	Column	485	0.2608
Cluster 11	Column	340	0.327
Cluster 12	Column	188	0.2001
Cluster 13	Column	127	0.1781
Cluster 14	Column	93	0.3507
Cluster 15	Column	295	0.3279
Cluster 16	Column	207	0.2455
Cluster 17	Column	499	0.09764
Cluster 18	Column	226	0.3535
Cluster 19	Column	646	0.2689
Cluster 20	Column	410	0.331

Contains Cluster_AAA

77: K-Means_col9_row9_columns_17-20_pos-cor_50iterations_10%vali

Table | Clustering Scores | Cluster Assignments | k = 17: Cluster Info | k = 18:

Number of Clusters	Profile Type	Explained Variance
k = 17	Column	0.7008
k = 18	Column	0.702
k = 19	Column	0.7085
k = 20	Column	0.7121

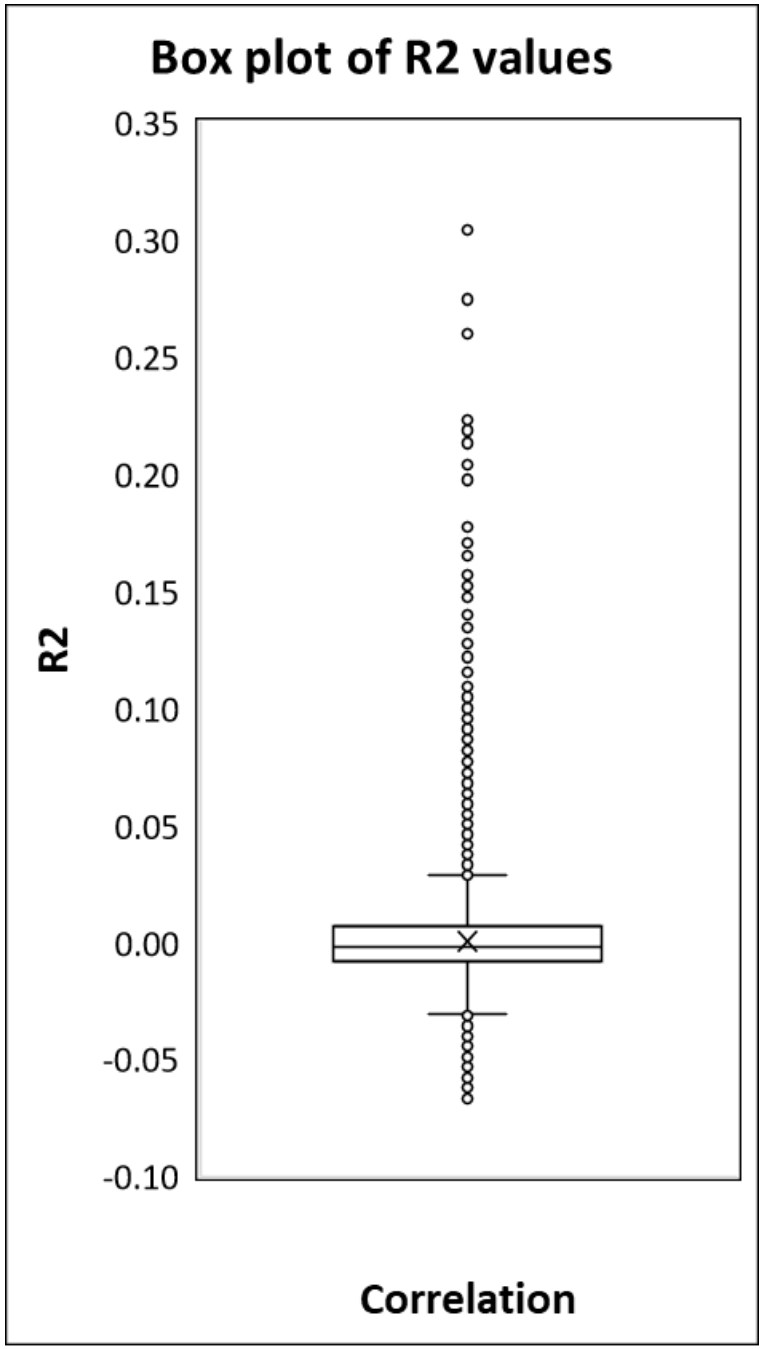
K=20: Cluster (14) with 93 peptides including Cluster_AAA

→ All these peptides are candidate biomarkers for LMA response

Correlation with LMA

Formulated in 1895 (Pearson 1895), correlation analysis aims at measuring an association or linear relationship between two continuous variables, by which the change in the magnitude of one variable (LMA) is associated with a change in the magnitude of the other variable (peptide abundance), either in the same (positive correlation) or in the opposite (negative correlation) direction (Schober, Boer, and Schwarte 2018).

Correlation coefficients (R2 or r) are scaled from -1 to +1, where 0 denotes no association and the relationship gets stronger towards an absolute value of 1.



Correlation spans -7% < R2 < 30%
R2(Cluster_AAA)=1

28 peptides with R2>15%

→ All these peptides are candidate biomarkers for LMA response

Linear Model (LM)

Simple linear regression dates from 1805 (Legendre 1805) and determines the relationship between a quantitative outcome and a single quantitative explanatory variable (Seltman 2018), such as LMA.

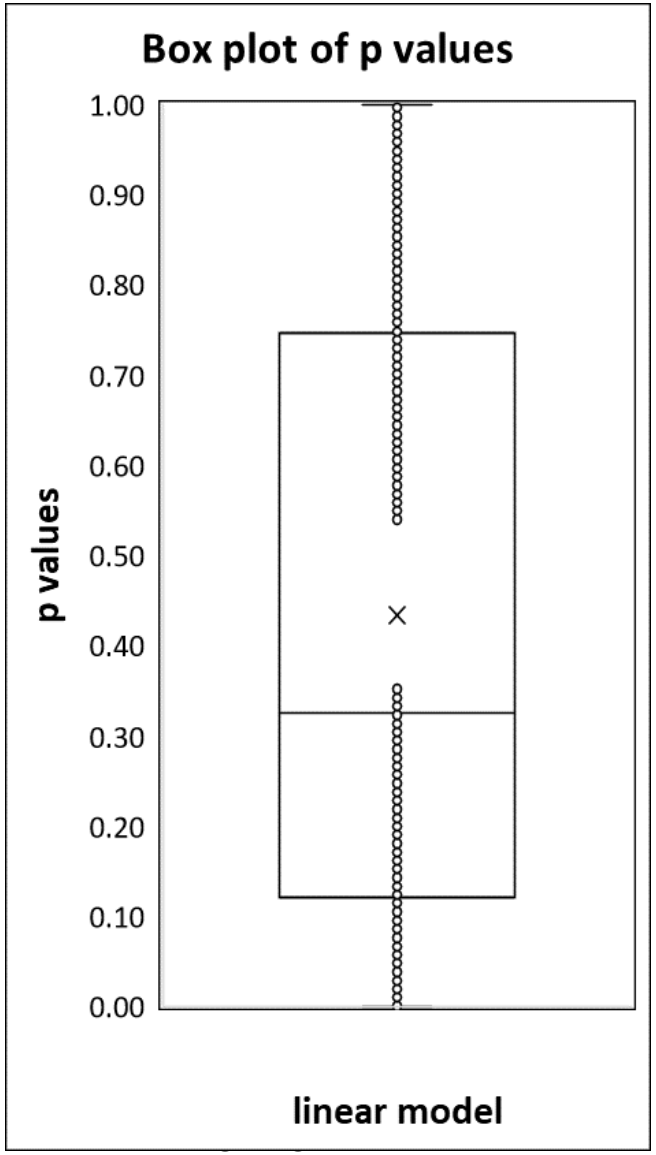
One factor linear model: **y = Inv(AAA) + e**

False discovery rates (FDR) using Benjamini-Hochberg method to compute q-values.

q-values spans $6x10^{-8}$ < q values < 1
q-value(INV(AAA))=0

494 peptides with q values < 0.05

→ All these peptides are candidate biomarkers for LMA response



Cluster_AAA →

79: Linear Model_col9_row9_columns_y=INV(AAA)

Table | P-Values | F-Ratio | SSE | AIC | Estimates | Design Matrix | Info

Column	INV(AAA)	INV(AAA) (BH-Q)
AAA'	1.000E-35	2.687E-31
Cluster_27763'	5.960E-8	0.0008008
Cluster_21745'	1.192E-7	0.001068
Cluster_36604'	2.980E-7	0.002002
Cluster_27914'	4.172E-7	0.002242
Cluster_10117'	5.364E-7	0.002403
Cluster_10847'	1.132E-6	0.004347
Cluster_03799'	1.311E-6	0.004405
Cluster_05224'	1.550E-6	0.004514
"""Granule-bound starch ...	1.848E-6	0.004514
Cluster_36089'	1.788E-6	0.004514
Cluster_36671'	2.146E-6	0.004805
Cluster_40484'	2.325E-6	0.004805
Cluster_02536'	4.470E-6	0.008581
Cluster_09759'	6.139E-6	0.01031
SET domain-containing ...	5.960E-6	0.01031
Cluster_27945'	7.391E-6	0.01168
Cluster_12529'	1.001E-5	0.01495
Cluster_28902'	1.073E-5	0.01517
Cluster_09116'	1.335E-5	0.01708
Actin (Fragment) OS=Trit...	1.305E-5	0.01708
Cluster_30545'	1.687E-5	0.0206
Aminotran_1_2 domain-...	1.806E-5	0.0211
DUF295 domain-contain...	2.187E-5	0.02351
Cluster_30587'	2.104E-5	0.02351
Cluster_00321'	2.396E-5	0.02418
Cluster_01403'	2.629E-5	0.02418
Cluster_09552'	2.766E-5	0.02418
"""GT75-3 OS=Triticum a...	2.789E-5	0.02418
Cluster_29209'	2.480E-5	0.02418
Cluster_40266'	2.658E-5	0.02418
Cluster_18053'	3.165E-5	0.02658
Cluster_03170'	3.815E-5	0.02699
Cluster_03206'	3.958E-5	0.02699
Uncharacterized protein ...	4.017E-5	0.02699
Globulin 3 OS=Triticum a...	3.761E-5	0.02699
Clathrin heavy chain OS...	3.821E-5	0.02699
Cluster_12385'	3.779E-5	0.02699
Cluster_13391'	4.005E-5	0.02699
Uncharacterized protein ...	3.344E-5	0.02699
Cluster_28803'	4.131E-5	0.02707
Cluster_31750'	4.280E-5	0.02738
Cluster_02048'	4.721E-5	0.02758
Uncharacterized protein ...	4.613E-5	0.02758
Cluster_10993'	4.584E-5	0.02758
Cluster_35933'	4.655E-5	0.02758
Cluster_02086'	5.001E-5	0.02859
Histone H2A OS=Triticu...	5.388E-5	0.03017
Cluster_03209'	6.270E-5	0.03024
Phosphoglucomutase (F...	6.825E-5	0.03024
Cluster_07494'	7.367E-5	0.03024
Cluster_08210'	6.092E-5	0.03024
Cluster_09917'	6.062E-5	0.03024
Cluster_10519'	7.427E-5	0.03024
Cluster_12093'	7.313E-5	0.03024

LMA biomarkers found!

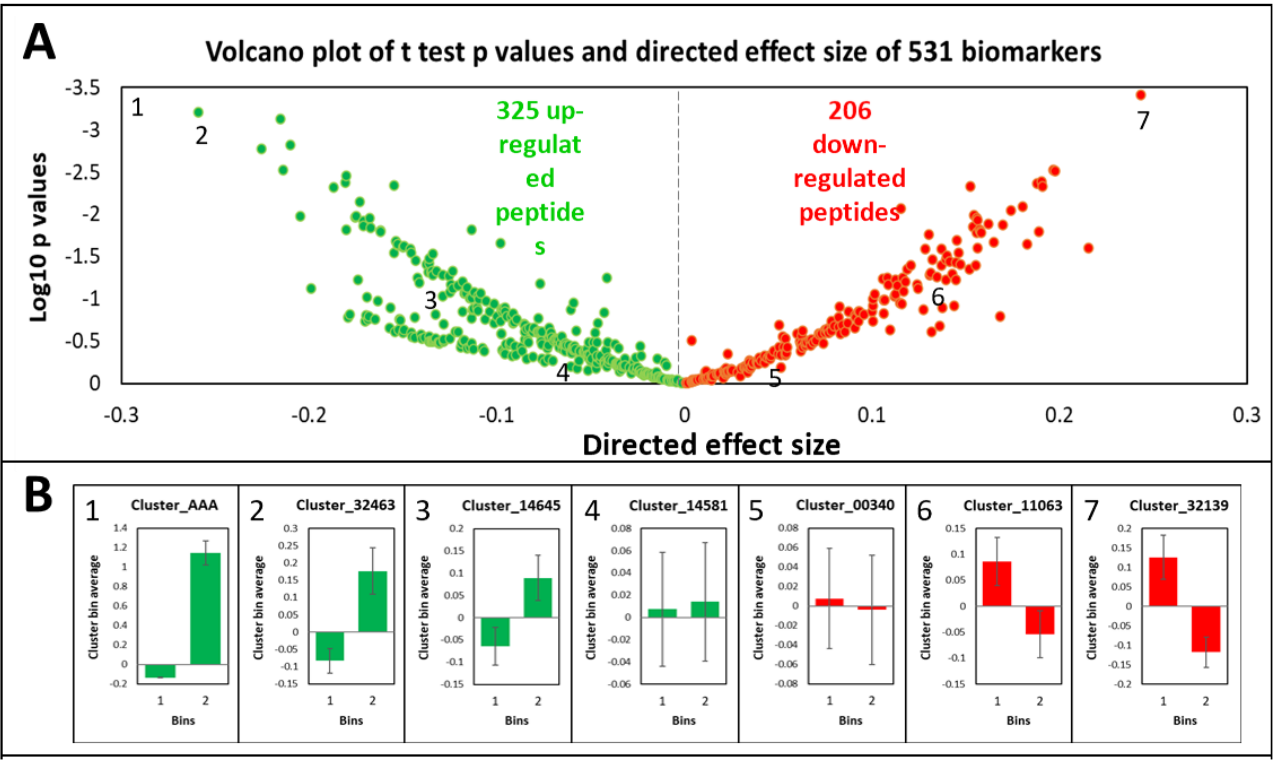
531 LMA biomarkers selected with at least one of the following criteria:

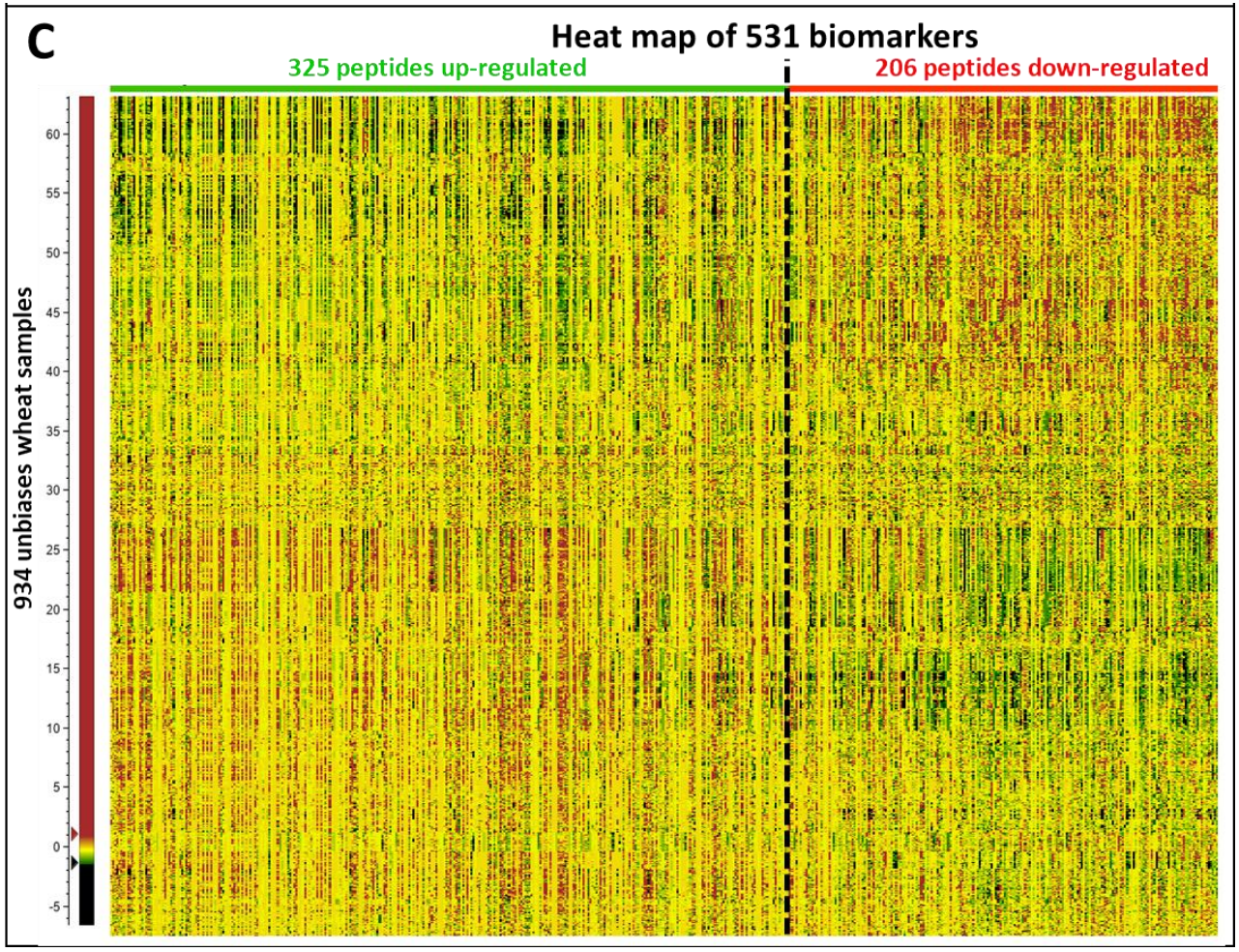
- Linear model p-value < 5%, and/or
- Correlation>15%, and/or
- SOM group (4,3), and/or
- K-means group 14, and/or
- Divisive HCA Group 1915-1947

Cluster	1: Correlation	75: SOM_Cluster Row	75: SOM_Cluster Colu	77: K-Means_k=20	79: Linear Model	111: Cluster Order
AAA'	1	4	3	14	1E-35	1936
Cluster_22401'	0.303912401	4	3	11	0.002108038	2005
Cluster_20947'	0.274328351	4	3	11	0.006105006	2012
Cluster_07104'	0.259491146	4	3	11	0.017609116	2009
Cluster_36671'	0.223005176	4	3	11	2.14577E-06	1530
Cluster_15590'	0.218448997	4	3	11	0.007762192	2004
Cluster_03685'	0.212924361	4	3	11	0.288428545	2011
Cluster_27255'	0.203597844	6	7	3	0.001879931	3007
Cluster_26957'	0.200891793	6	7	3	0.002861142	3008
Cluster_29815'	0.197527409	6	7	3	0.002364874	3006
Cluster_08864'	0.197312891	3	6	11	0.010285019	1929
Cluster_24547'	0.180988193	6	8	3	0.000482082	2997
Cluster_37329'	0.179359794	6	6	14	0.002170265	986
Cluster_36110'	0.1773808	2	6	18	0.001100659	1008
Cluster_03658'	0.172569871	6	7	3	0.039712604	3004
Cluster_15371'	0.172173619	6	6	7	0.002248883	3913
Cluster_02070'	0.17196101	6	4	18	0.000221431	3277
Cluster_16885'	0.171344578	6	6	7	0.005627691	3914
Cluster_10195'	0.170690715	6	6	14	0.002759993	985
Cluster_03425'	0.170167267	6	7	3	0.003591776	2584
Cluster_25361'	0.165119827	1	1	19	0.002921819	1059
Cluster_25705'	0.158247709	3	8	4	0.008291959	161
Cluster_09541'	0.157809198	6	6	14	0.006407202	982
Cluster_09728'	0.157651603	1	1	19	0.002828062	1057
Cluster_24874'	0.15672189	1	1	19	0.00367874	1058
Cluster_30251'	0.154172003	6	6	7	0.004098297	3813
Cluster_23729'	0.152121782	6	6	7	0.021191597	3917
Cluster_10059'	0.150398433	3	6	11	0.026453555	1926
Cluster_33052'	0.150056422	6	6	7	0.002813816	3812
Cluster_34465'	0.134042859	3	6	11	0.008888842	1928
Cluster_39518'	0.129546583	6	8	14	0.000142515	3056
Cluster_33492'	0.127616823	6	8	14	0.000324011	3067
Cluster_13006'	0.122289538	6	1	3	0.000894904	5338
Cluster_23029'	0.118868172	4	3	11	0.160089076	2000
Cluster_03711'	0.118657947	6	8	14	0.000476301	3070
Cluster_35794'	0.118228734	6	4	18	0.000300884	3264
Cluster_43428'	0.11719209	6	8	14	0.002024889	3075
Cluster_17010'	0.116909027	3	5	14	0.000902057	1995
Cluster_39634'	0.115722418	3	6	4	0.08217711	1935
Cluster_28606'	0.109239757	6	6	18	0.004714012	3265
Cluster_28653'	0.108831108	3	4	14	0.000654459	1993
Cluster_40208'	0.102374136	6	8	14	0.011272252	3057
Cluster_05084'	0.101284146	6	6	14	0.07008893	3802
Cluster_14079'	0.10110271	4	3	11	0.593683124	2010
Cluster_22882'	0.099902749	6	8	14	0.001497388	3065
Cluster_15481'	0.099127114	6	8	14	0.019695463	3073
Cluster_42380'	0.098648489	6	8	14	0.026695609	3074
Cluster_06924'	0.098106086	6	8	14	0.033769365	3066

T test, volcano plot, and heat map on 531 LMA biomarkers

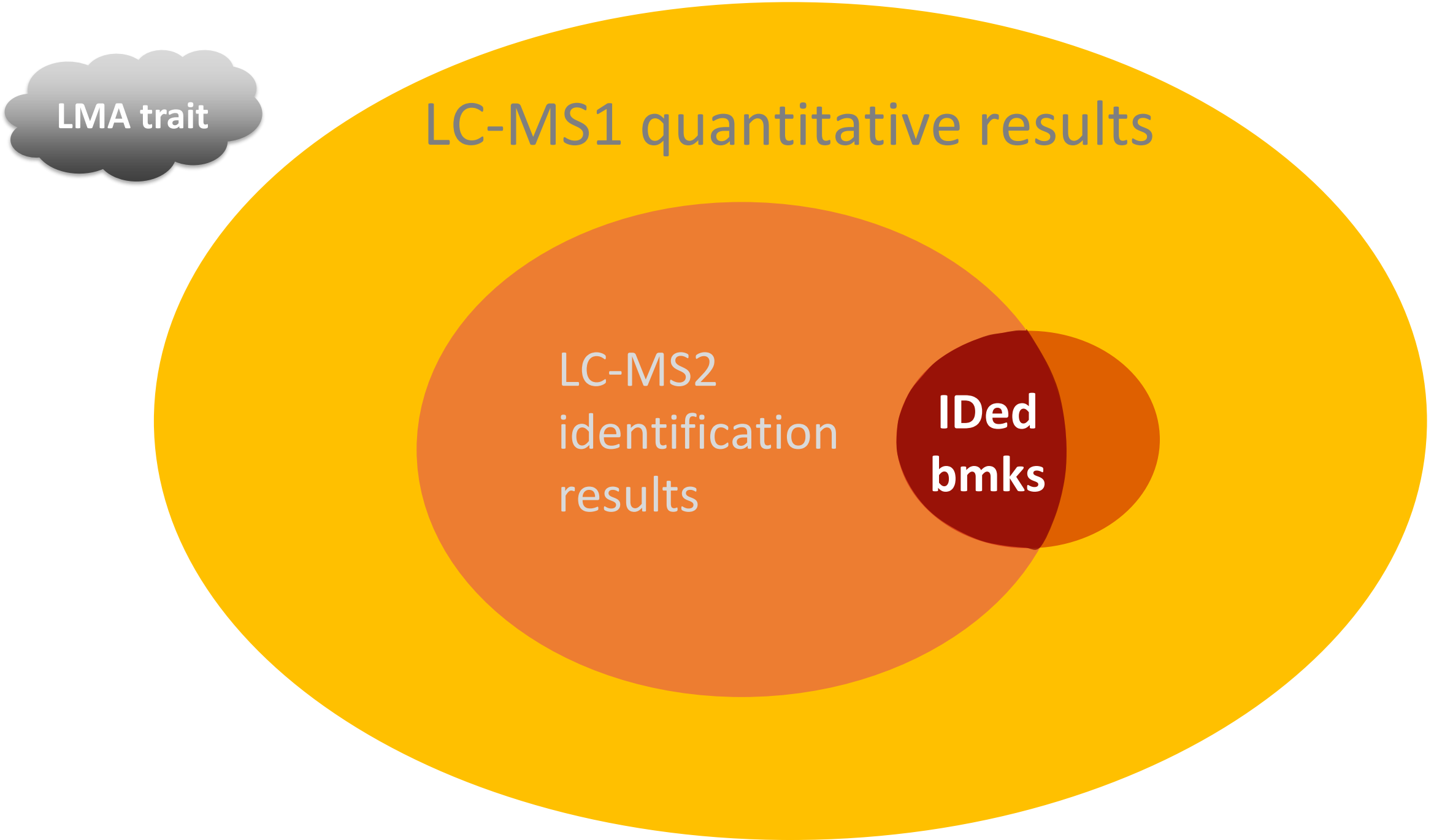
A volcano plot is a type of scatterplot that shows statistical significance (q-value) versus magnitude of change (effect size or fold change). It is often utilised for biomarker discovery in differential expression studies (Li 2012).





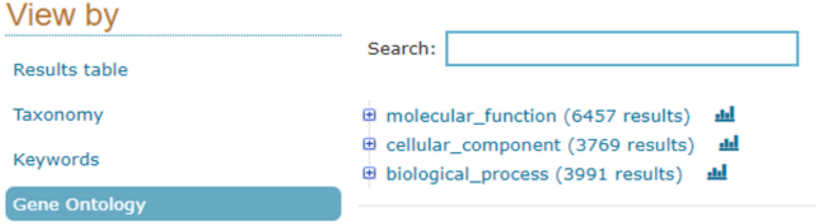
Using the 2-bin strategy on the unbiased dataset allowed to categorise biomarkers as being either up-regulated or down-regulated.

The heat map shows their full expression pattern.


LMA trait
LC-MS1 quantitative results
LC-MS2 identification results
IDed bmks

Identified proteins – qualitative data mining – description and gene ontology (GO) classes

A. 7939 out of 8044 uniprotKB AC/ID identifiers were successfully mapped to 7939 UniProtKB IDs



8,044 identities mapped in UniProtKB.

Described by 1,385 unique protein names.
Glutenins are prominent (word cloud).

They were GO-classified with:

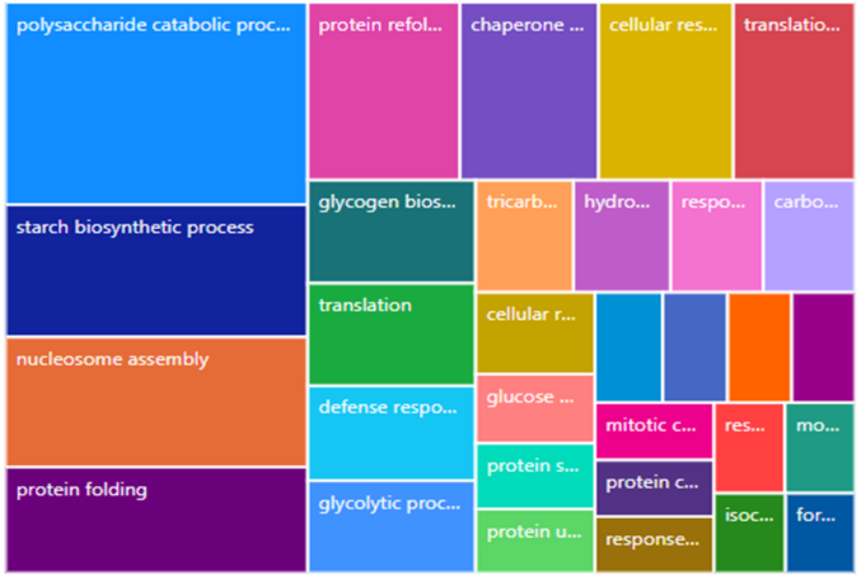
- 6,457 MF terms,
- 3,769 CC terms,
- 3,991 BP terms.

Sugar, starch, cytoplasm and nutrient reservoir dominate the wheat grain proteome (Tree maps).

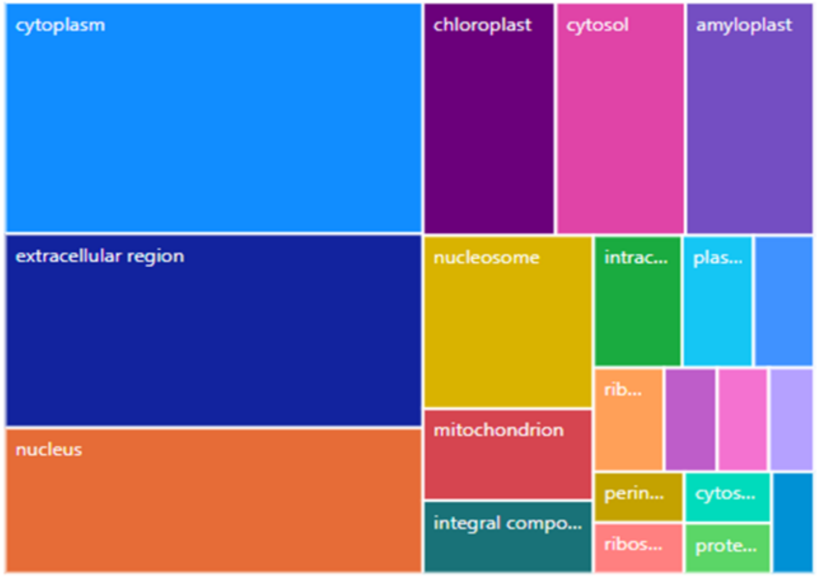
α-amylase was not identified

B. All protein names

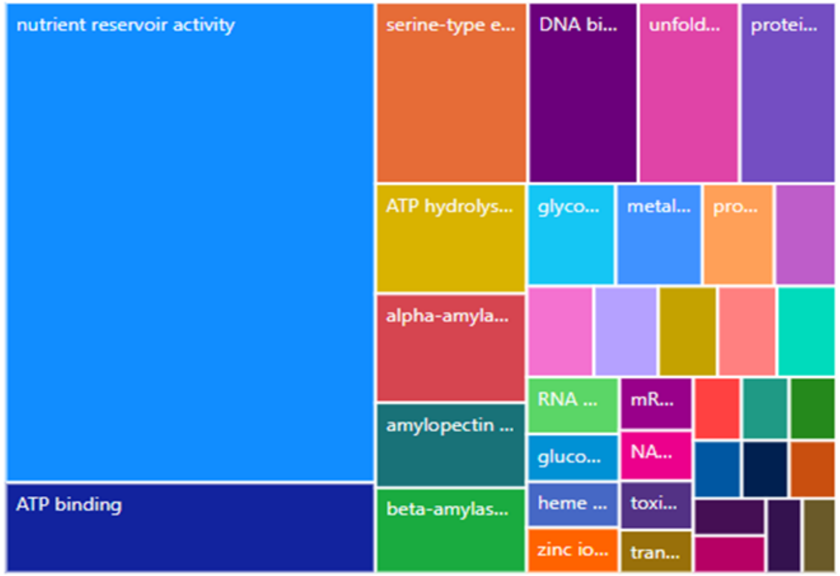
E. GOBP



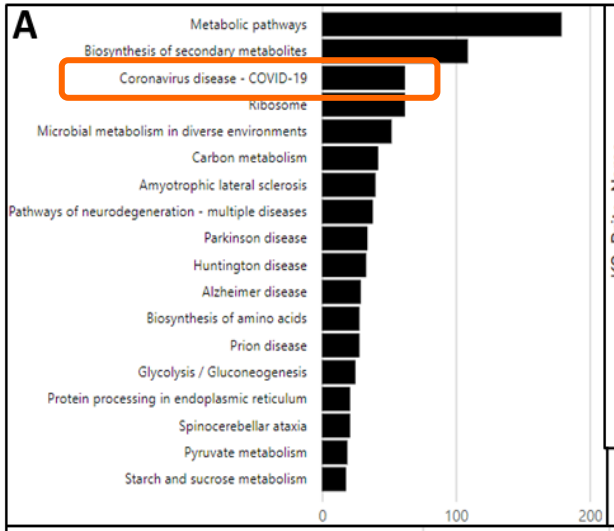
F. GOCC

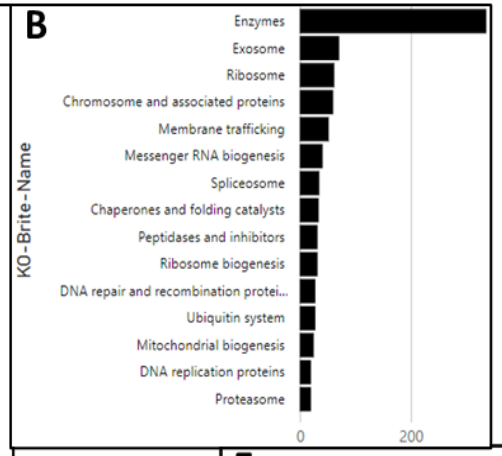


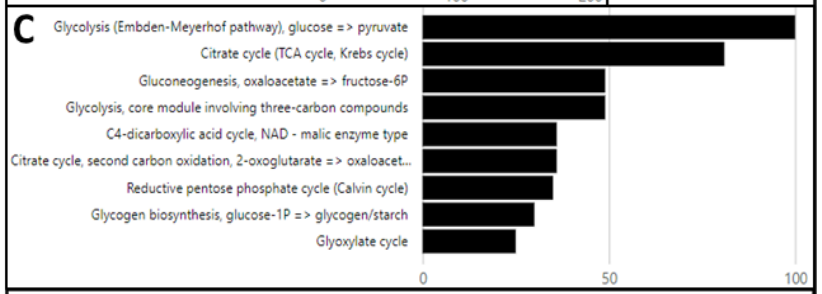
G. GOMF

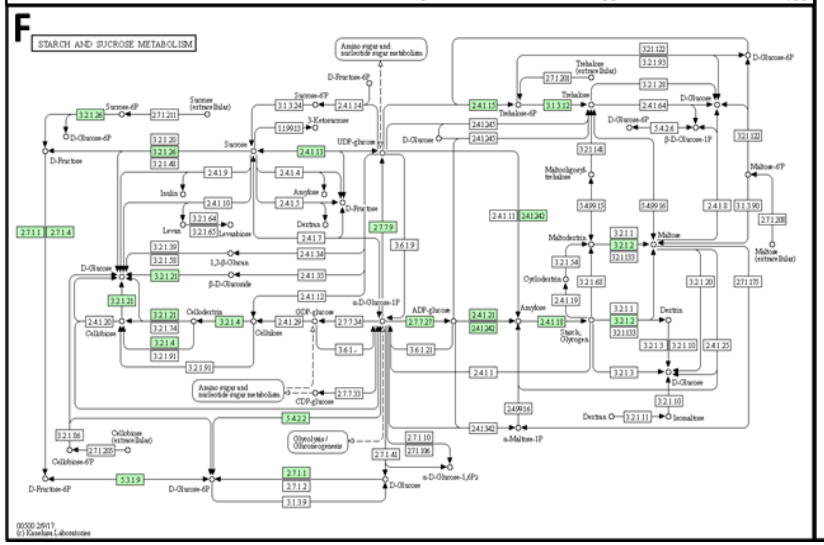


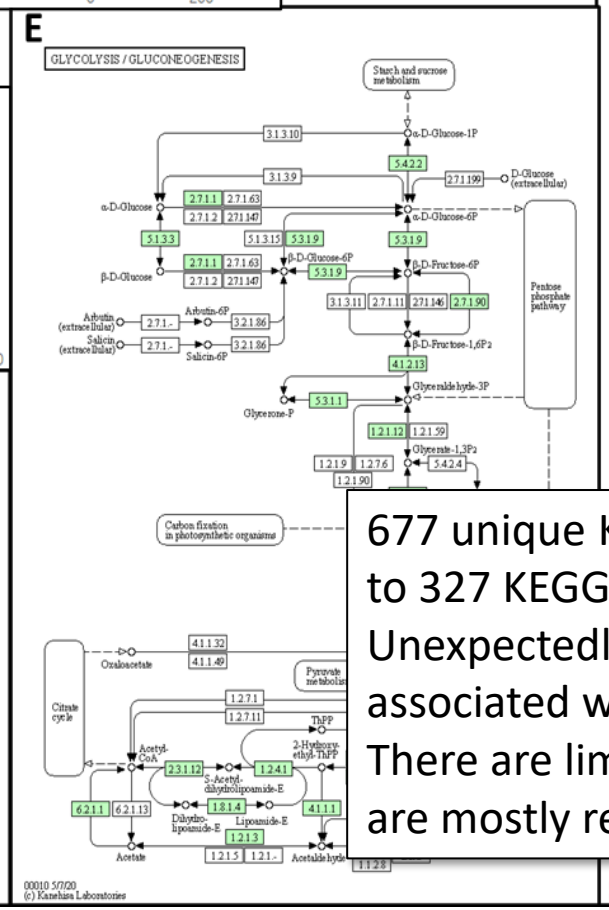
Identified proteins – qualitative data mining – KEGG pathways to visualise metabolisms involved in grain metabolisms

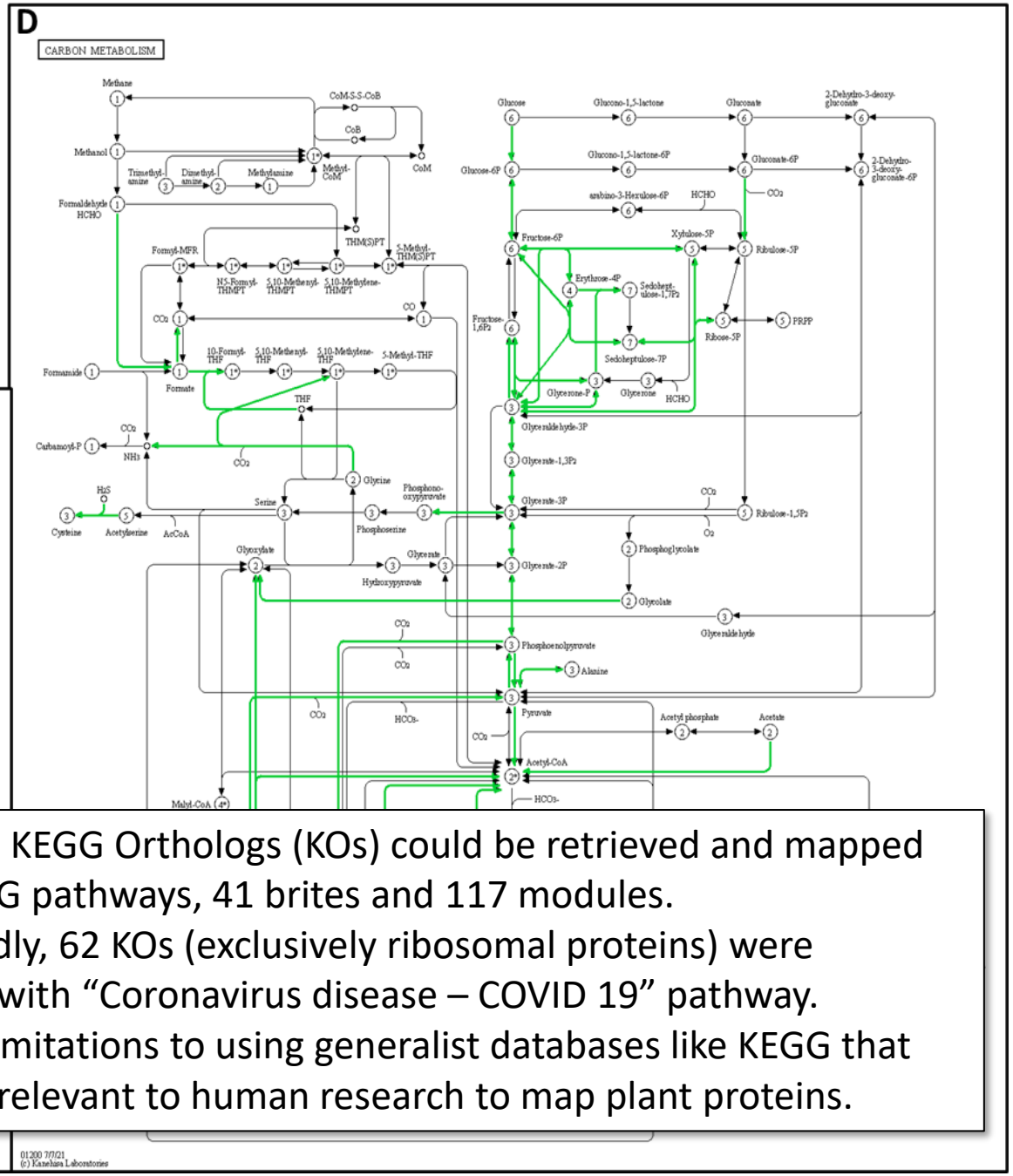






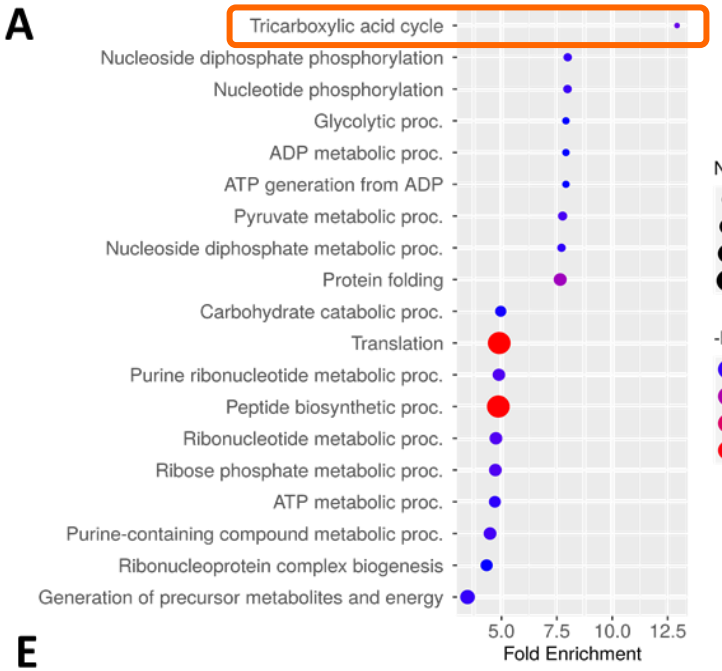


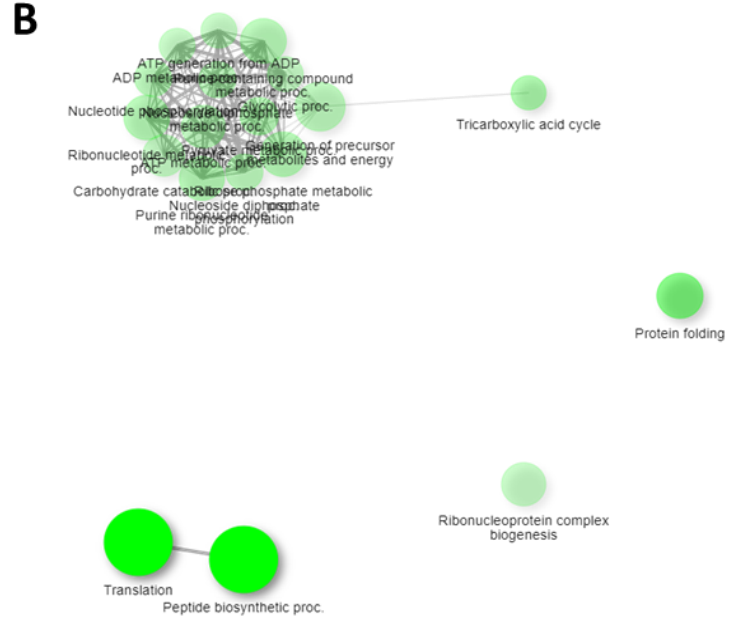


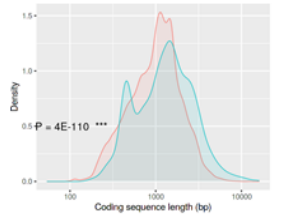


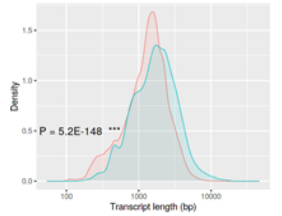
677 unique KEGG Orthologs (KOs) could be retrieved and mapped to 327 KEGG pathways, 41 brites and 117 modules.
Unexpectedly, 62 KOs (exclusively ribosomal proteins) were associated with “Coronavirus disease – COVID 19” pathway.
There are limitations to using generalist databases like KEGG that are mostly relevant to human research to map plant proteins.

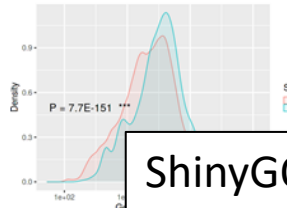
Identified proteins – qualitative data mining – ShinyGO to easily retrieve chromosomal locations of protein-expressed genes

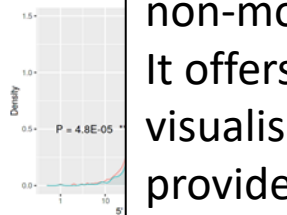


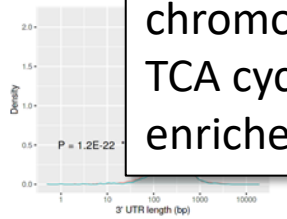


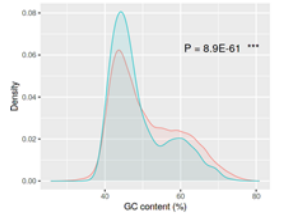


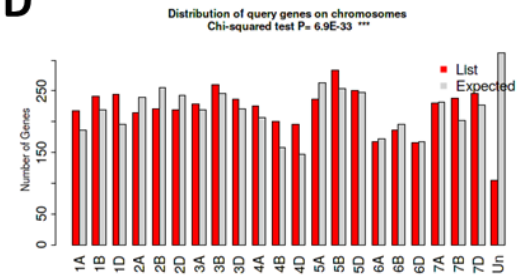


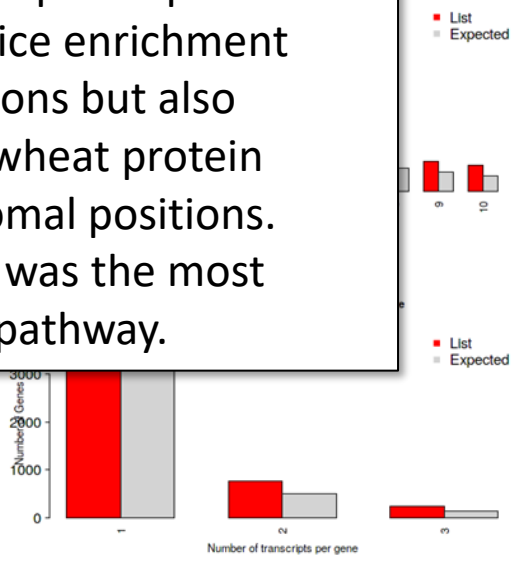


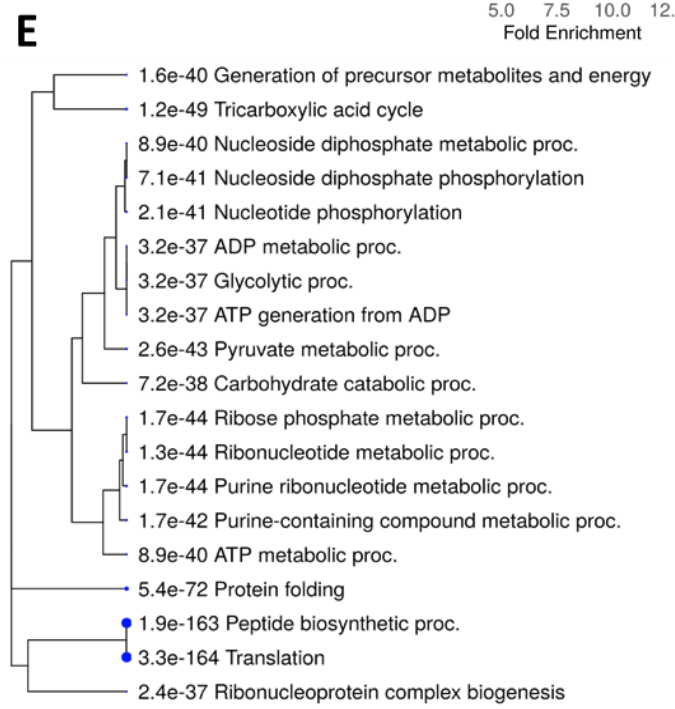


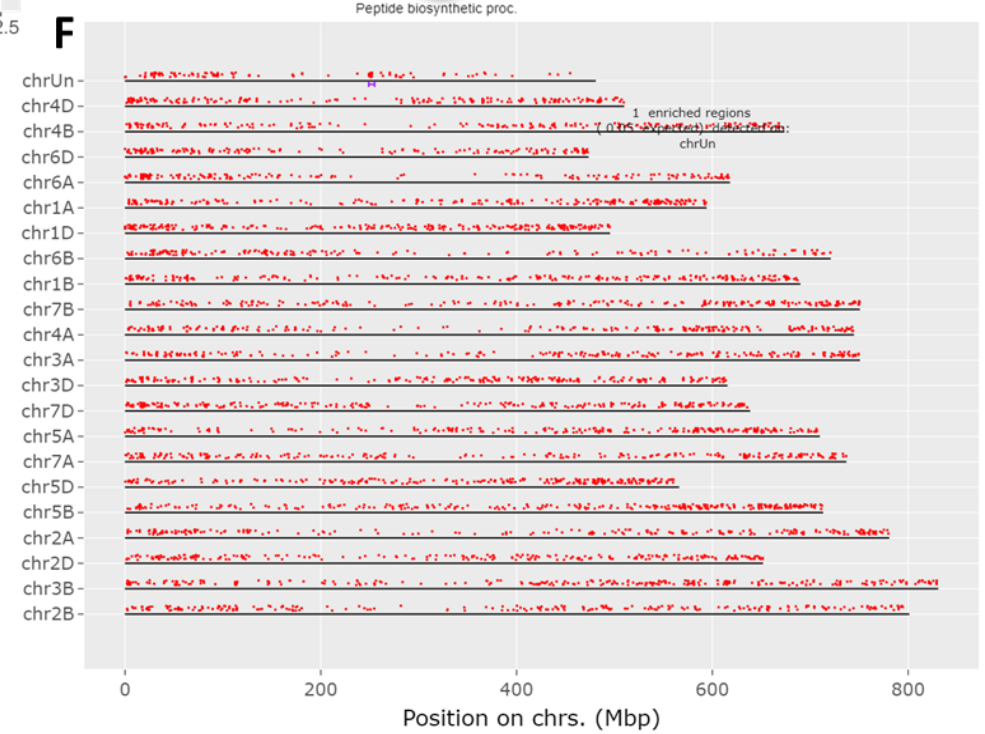








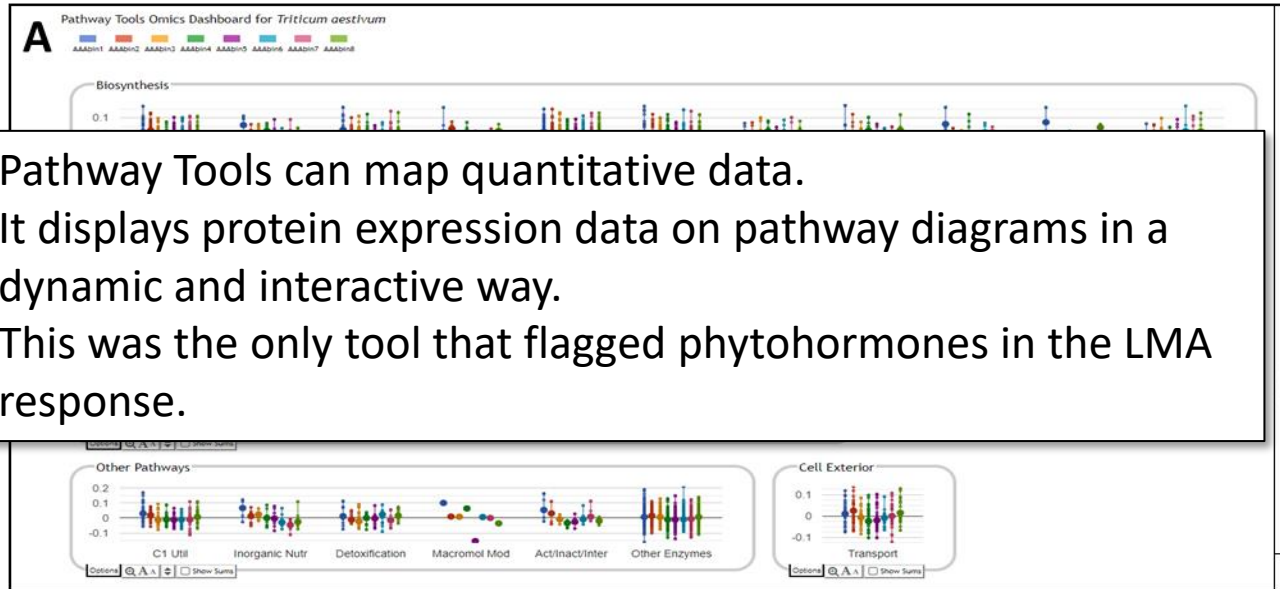


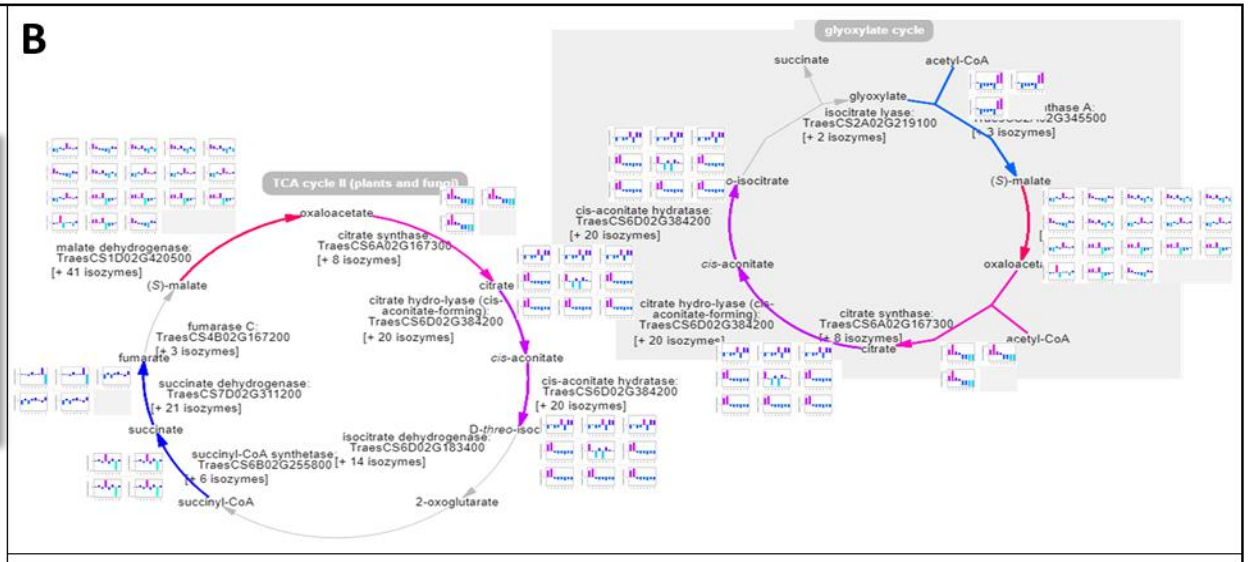


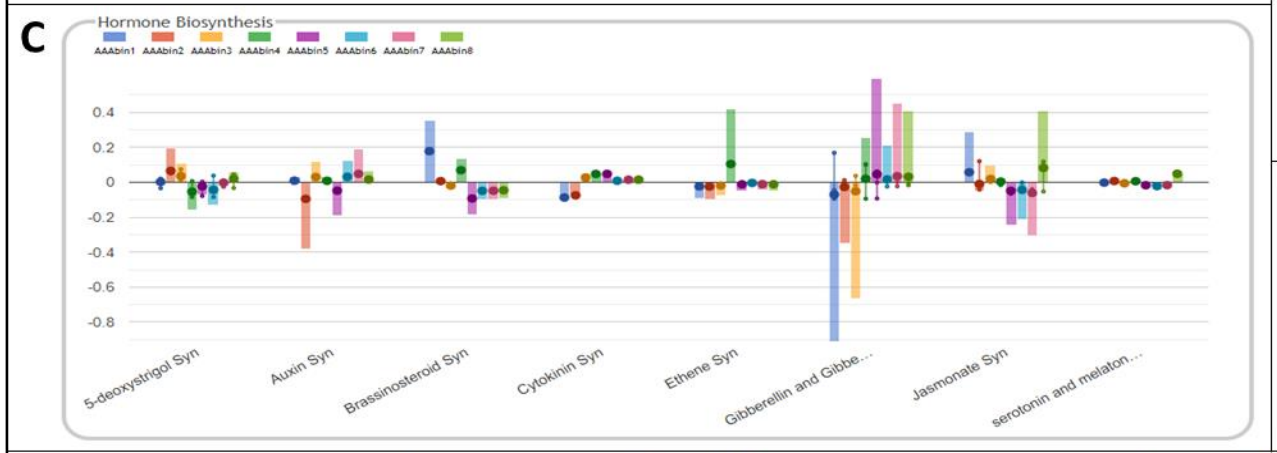
ShinyGO works very well on non-model plant species.
It offers nice enrichment visualisations but also provides wheat protein chromosomal positions.
TCA cycle was the most enriched pathway.

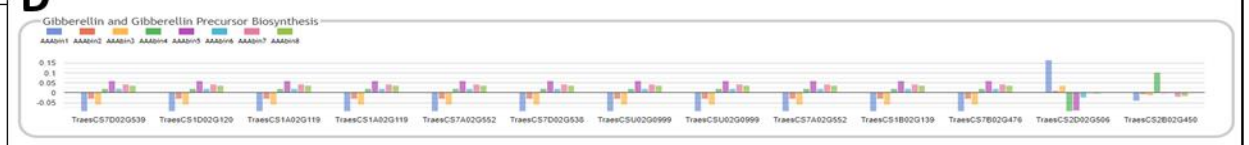
Identified proteins – quantitative data mining – Pathway Tools and 8-bin profiles to highlight perturbed pathways

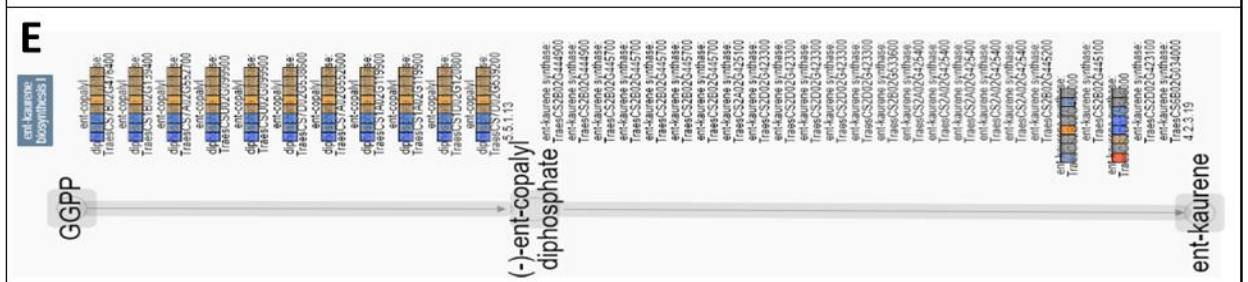
Pathway Tools can map quantitative data.
It displays protein expression data on pathway diagrams in a dynamic and interactive way.
This was the only tool that flagged phytohormones in the LMA response.

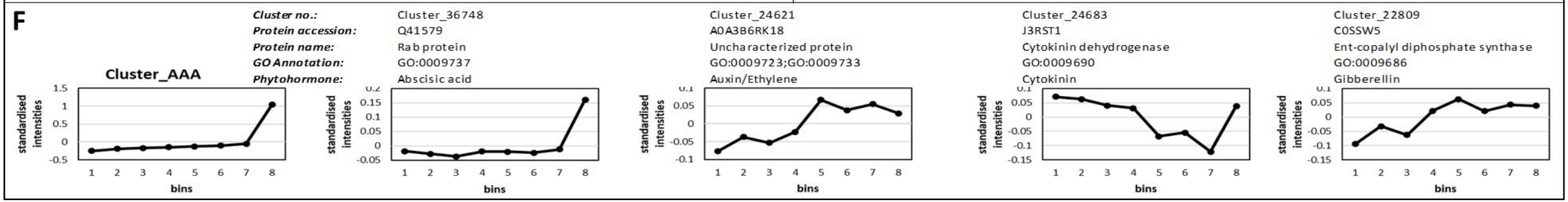


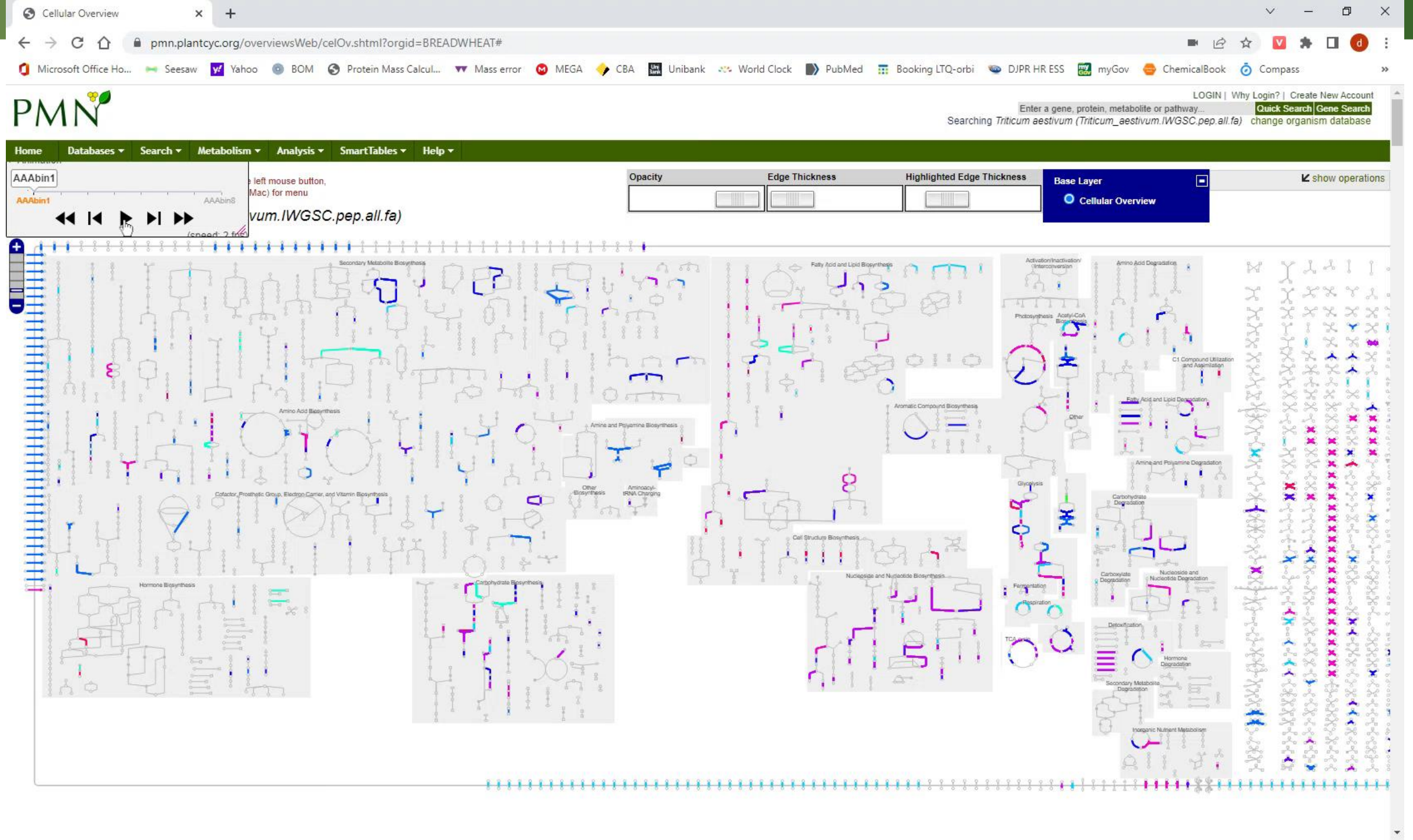






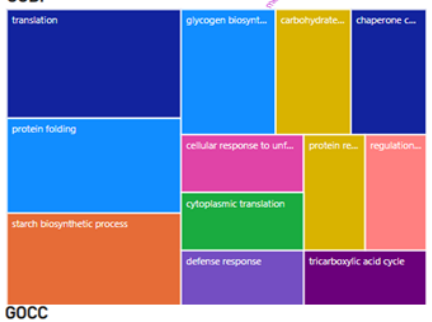


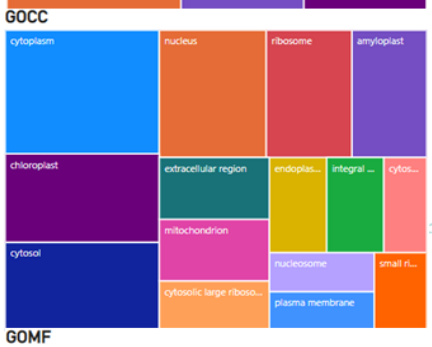


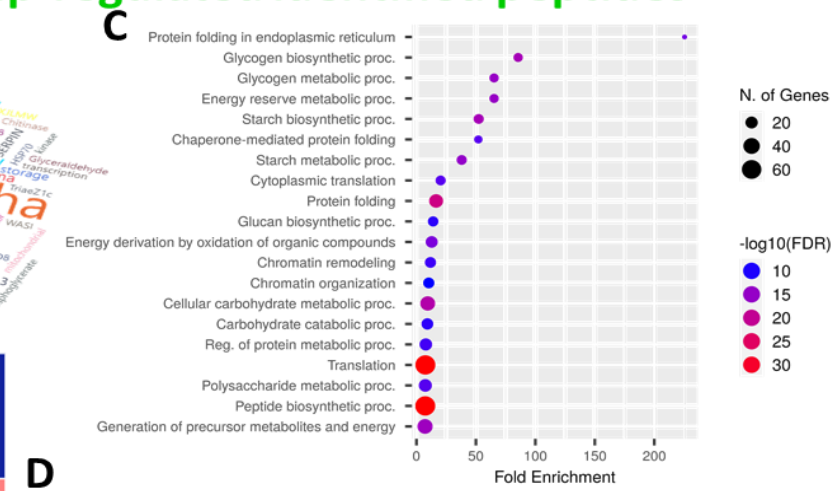

Cellular Overview
pmn.plantcyc.org/overviewsWeb/celOv.shtml?orgid=BREADWHEAT#
Microsoft Office Ho...
Seesaw
Yahoo
BOM
Protein Mass Calcul...
Mass error
MEGA
CBA
Unibank
World Clock
PubMed
Booking LTQ-orbi
DJPR HR ESS
myGov
ChemicalBook
Compass
PMN
LOGIN | Why Login? | Create New Account
Enter a gene, protein, metabolite or pathway...
Quick Search
Gene Search
Searching Triticum aestivum (Triticum_aestivum.IWGSC.pep.all.fa)
change organism database
Home
Databases
Search
Metabolism
Analysis
SmartTables
Help
AAAbin1
AAAbin1
AAAbin8
left mouse button,
Mac) for menu
vum.IWGSC.pep.all.fa)
Opacity
Edge Thickness
Highlighted Edge Thickness
Base Layer
Cellular Overview
show operations
Secondary Metabolite Biosynthesis
Amino Acid Biosynthesis
Amine and Polyamine Biosynthesis
Other Biosynthesis
Aminoacyl-tRNA Charging
Cofactor, Prosthetic Group, Electron Carrier, and Vitamin Biosynthesis
Hormone Biosynthesis
Carbohydrate Biosynthesis
Fatty Acid and Lipid Biosynthesis
Aromatic Compound Biosynthesis
Cell Structure Biosynthesis
Nucleoside and Nucleotide Biosynthesis
Activation/Inactivation/ Interconversion
Photosynthesis
Acetyl-CoA Biosynthesis
Other
Glycolysis
Fermentation
Respiration
TCA cycle
Amino Acid Degradation
C1 Compound Utilization and Assimilation
Fatty Acid and Lipid Degradation
Amine and Polyamine Degradation
Carbohydrate Degradation
Carboxylate Degradation
Nucleoside and Nucleotide Degradation
Detoxification
Hormone Degradation
Secondary Metabolite Degradation
Inorganic Nutrient Metabolism

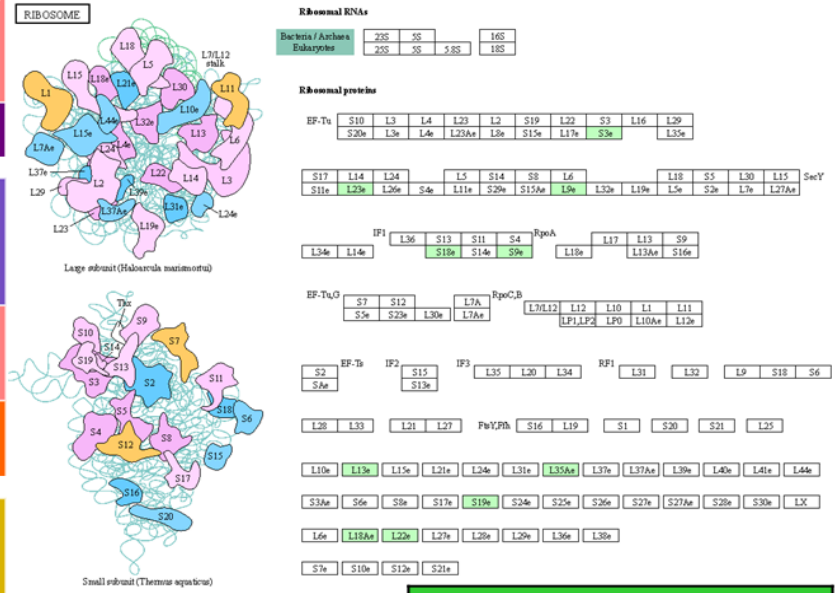
Identified LMA biomarkers – impacted metabolisms

207 up-regulated identified peptides





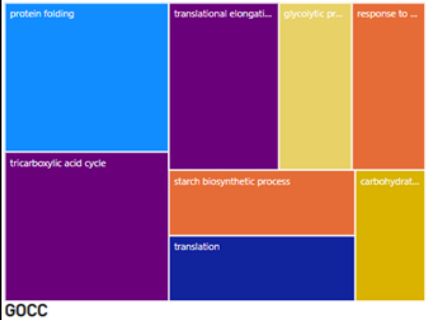


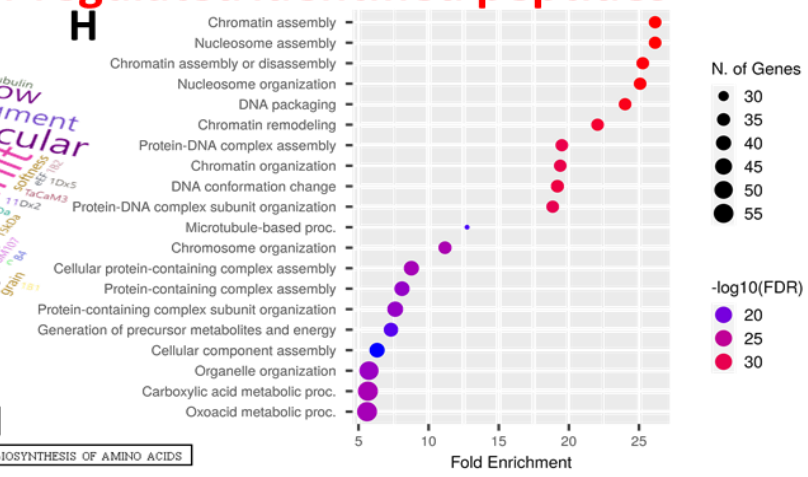


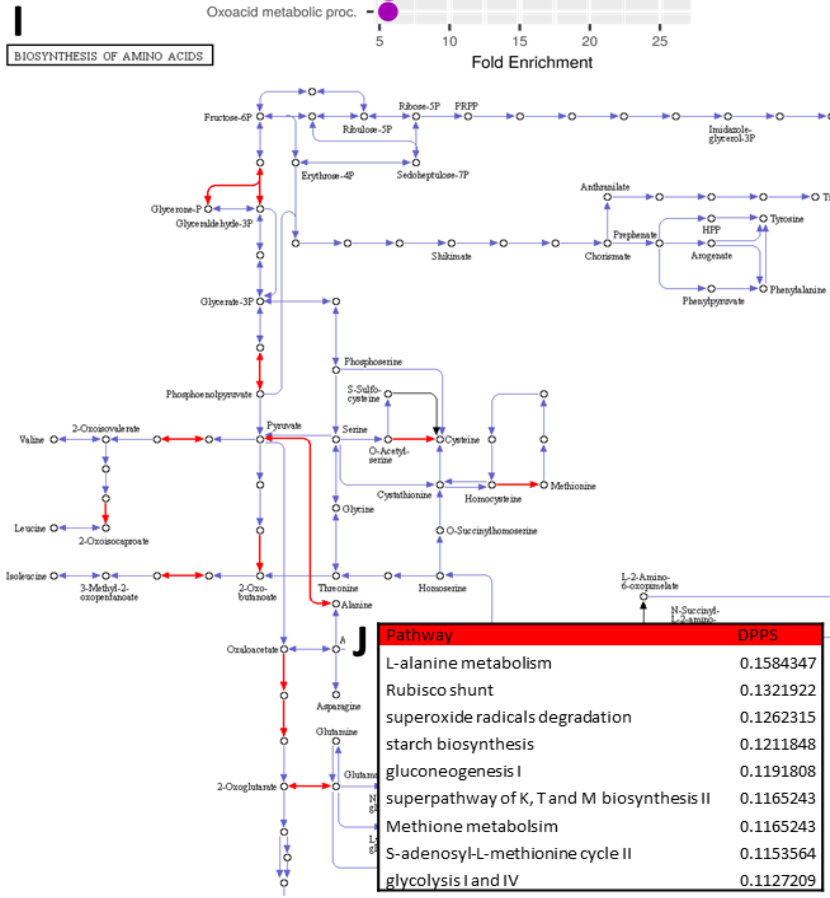
E

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

183 down-regulated identified peptides







J

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

Identified LMA biomarkers – Circos plots

Circos plots (Krzywinski 2009) efficiently capture qualitative and quantitative information.

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.

Plotting their fold changes outlined that most genome areas hosted both up- and down-regulated biomarkers bar a few exceptions on chromosomes 4, 6 and 7 for all 3 genomes A, B, and D.

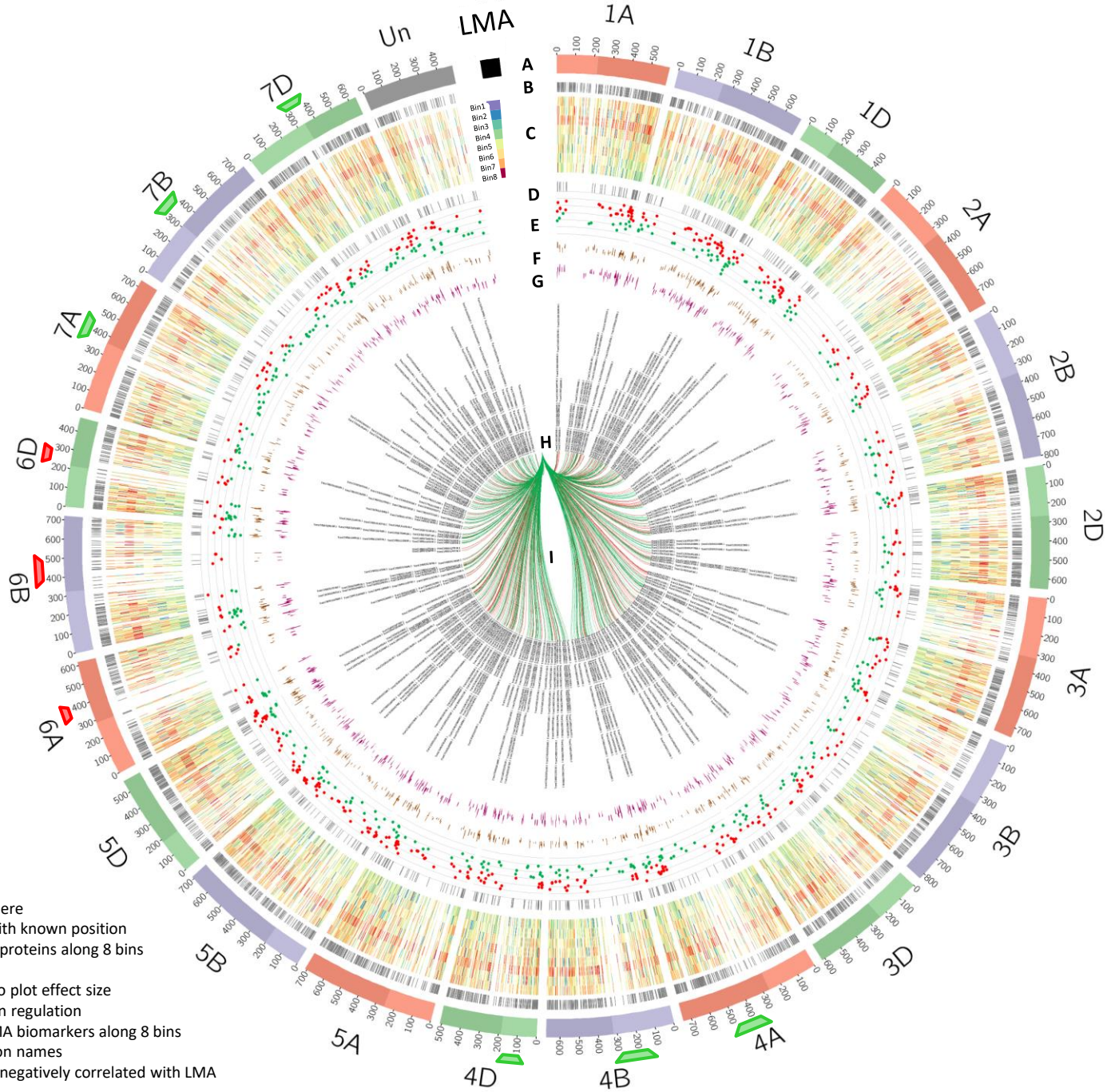
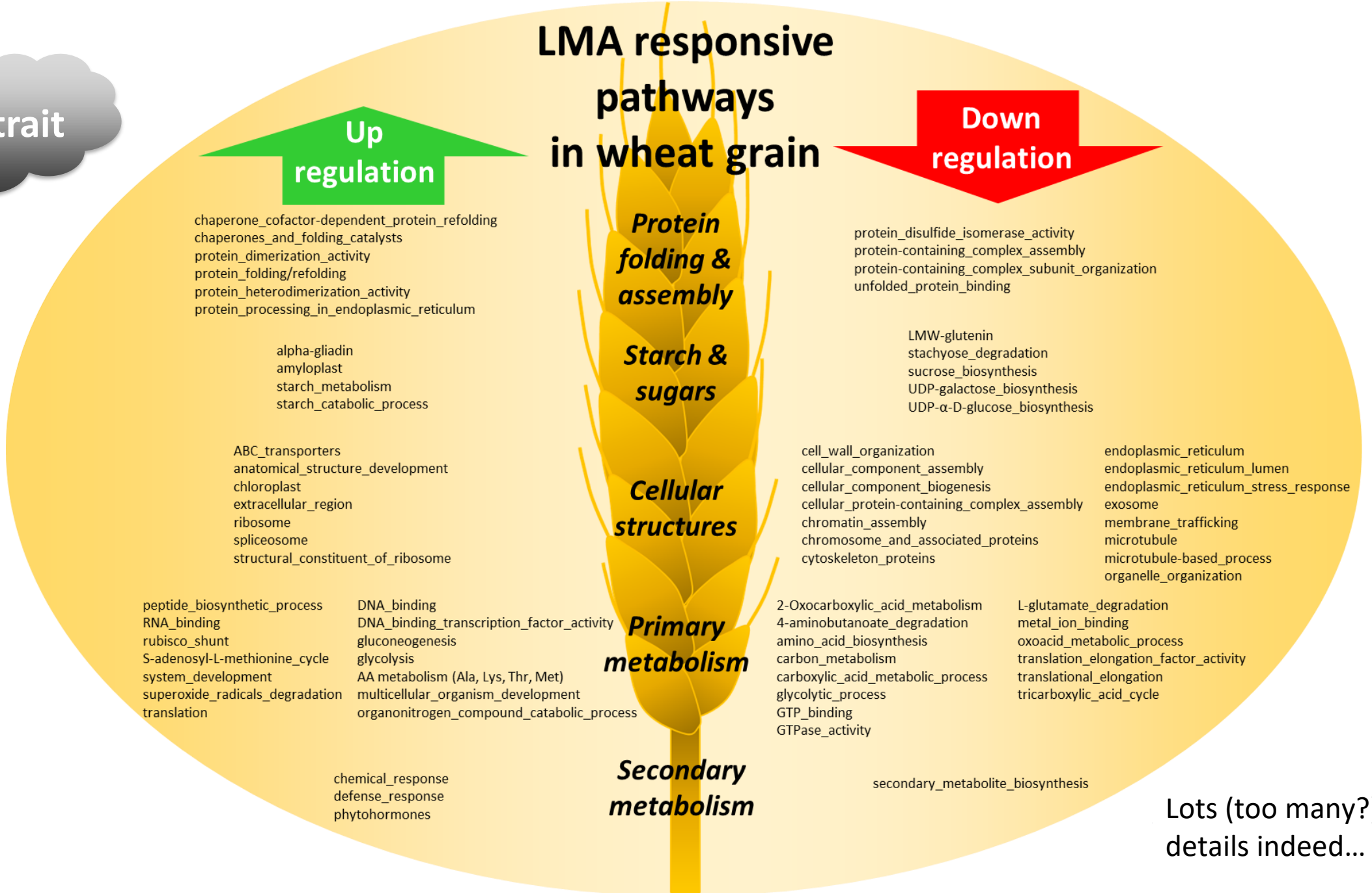


Figure legend:
A karyotype with centromere
B all identified proteins with known position
C expression profile of all proteins along 8 bins
D LMA biomarkers
E LMA biomarkers Volcano plot effect size
F LMA biomarker up/down regulation
G expression profile of LMA biomarkers along 8 bins
H LMA biomarker accession names
I biomarkers positively or negatively correlated with LMA

LMA trait



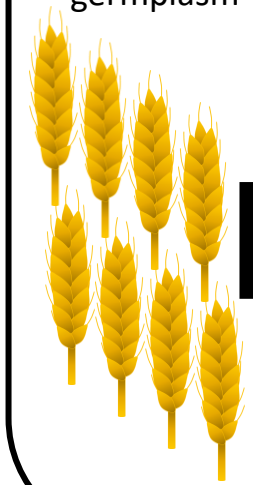
Lots (too many?) of details indeed…

Conclusions

- For the first time, LMA phenotype was explored via proteomics.
- All the differentially regulated biological processes highlighted in this study required various data mining tools.
- Whilst we did not find the LMA needle in the proteome haystack, proteomics deserves to be part of the wheat LMA molecular toolkit and should be adopted by LMA scientists and breeders in the future.
- We have a resource that can be applied to any other trait(s) and any other species.

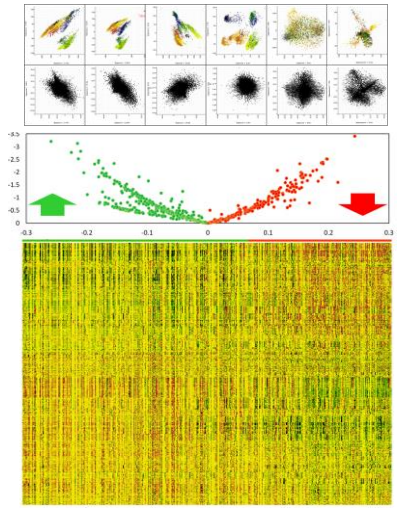
Community resource

Wheat grain germplasm

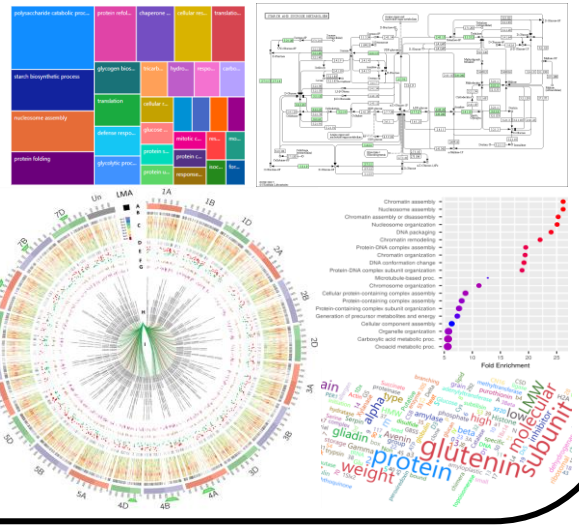


High-throughput MS-proteomics workflow

Big data analysis for biomarker discovery

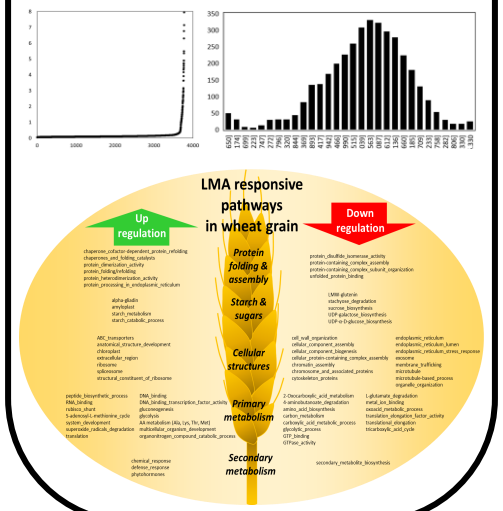


Data mining for biological relevance



Application

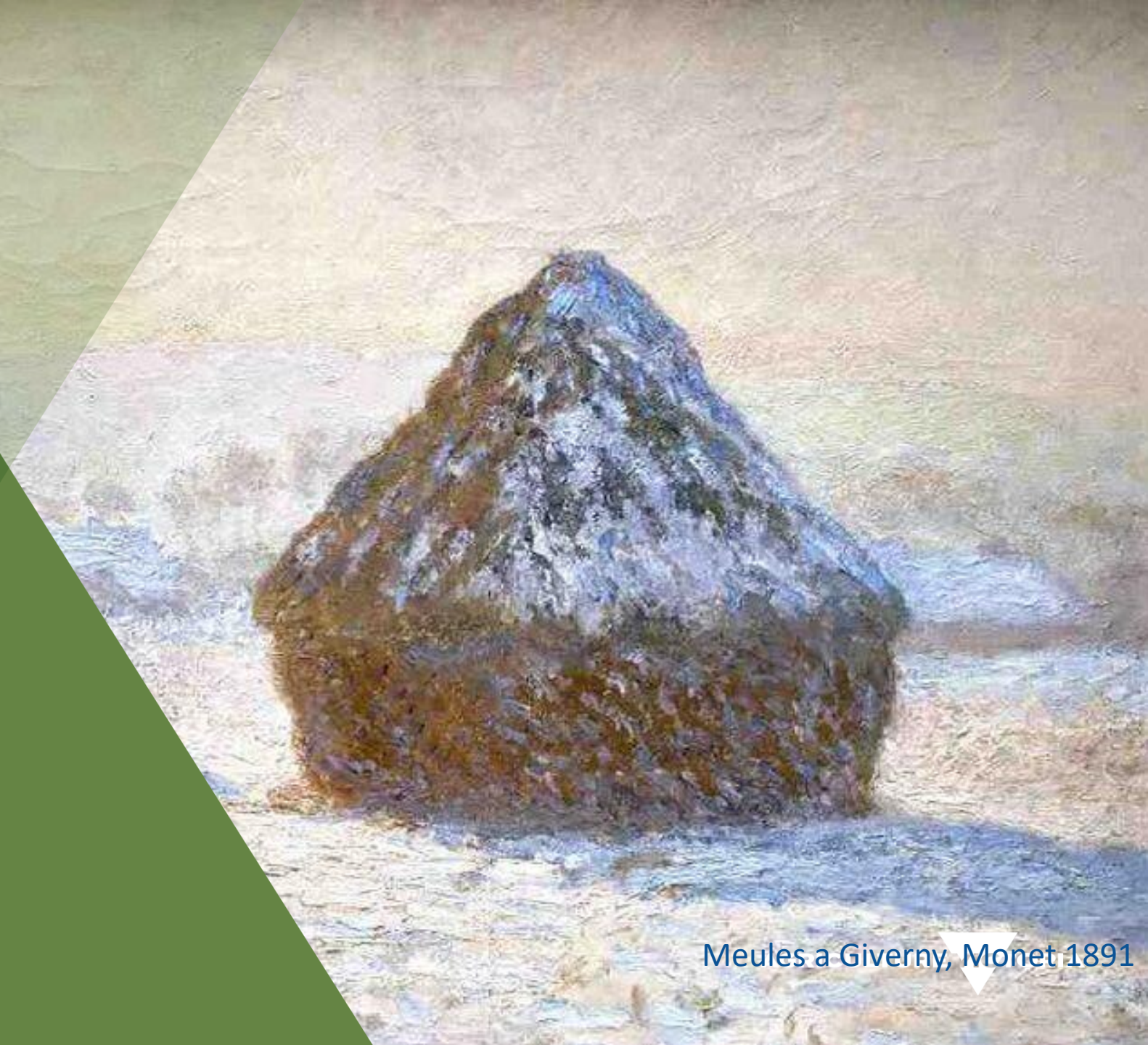
Understanding LMA response



Molecular LMA response - Summary

- In this work, stored LMA-affected grains activated their primary metabolisms such as glycolysis and gluconeogenesis, TCA cycle.
- It also including DNA- and RNA binding mechanisms, as well as 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 flagged notably with the up-regulation of phytohormones, chemical and defense responses.
- LMA further invoked cellular structures among which ribosomes, microtubules, and chromatin.
- Finally, and unsurprisingly, LMA expression greatly impacted grain starch and other carbohydrates with the up-regulation of alpha-gliadins and starch metabolism, while LMW glutenin, stachyose, sucrose, UDP-galactose and UDP-glucose were down-regulated.

Thank you!



Meules a Giverny, Monet 1891