Running Title: The Maize PeptideAtlas; a new community resource

For correspondence:

Klaas J. van Wijk - kv35@cornell.edu

One sentence summary (200 characters): A new community web resource with mass spectrometry-based maize proteome information and associated spectral, technical and biological metadata that can also aid in maize genome annotation.

The *Zea mays* PeptideAtlas – a new maize community resource

Klaas J. van Wijk[a*], Tami Leppert[b], Zhi Sun[b], Isabell Guzchenko[a], Erica Debley[a], Georgia Sauermann[a], Pratyush Routray[a], Luis Mendoza[b], Qi Sun[c] and Eric W. Deutsch[b*]

[a] Section of Plant Biology, School of Integrative Plant Sciences (SIPS), Cornell University, Ithaca, NY 14853, USA; [b]Institute for Systems Biology (ISB), Seattle, Washington 98109, USA; [c] Computational Biology Service Unit, Cornell University, Ithaca, NY 14853

ORCID ID: 0000-0001-9536-0487 (K.J.v.W); 0000-0001-8732-0928 (E.W.D.); 0000-0001-6140-2204 (Q.S.).; 0000-0002-7893-8619 (T.L.); 0000-0003-3324-6851 (Z.S.); 0000-0003-0128-8643 (L.M.); 0000-0003-1189-5973 (P.R.)

*correspondence: Klaas J. van Wijk, kv35@cornell.edu; Eric W. Deutsch: edeutsch@systemsbiology.org

ABSTRACT We developed the Maize PeptideAtlas resource (www.peptideatlas.org/builds/maize) to help solve questions about the maize proteome. Publicly available raw tandem mass spectrometry (MS/MS) data for maize were collected from ProteomeXchange and reanalyzed through a uniform processing and metadata annotation pipeline. These data are from a wide range of genetic backgrounds, including the inbred lines B73 and W22, many hybrids and their respective parents. Samples were collected from field trials, controlled environmental conditions, a range of (a)biotic conditions and different tissues, cell types and subcellular fractions. The protein search space included different maize genome annotations for the B73 inbred line from MaizeGDB, UniProtKB, NCBI RefSeq and for the W22 inbred line. 445 million MS/MS spectra were searched, of which 120 million were matched to 0.37 million distinct peptides. Peptides were matched to 66.2% of the proteins (one isoform per protein coding gene) in the most recent B73 nuclear genome annotation (v5). Furthermore, most conserved plastid- and mitochondrial-encoded proteins (NCBI RefSeq annotations) were identified. Peptides and proteins identified in the other searched B73 genome annotations will aid to improve maize genome annotation. We also illustrate high confidence detection of unique SNPs in the W22 proteome as compared to the B73 proteome. N-terminal acetylation, phosphorylation, ubiquitination, and three lysine acylations (K-acetyl, K-malonyl, K-hydroxybutyryl) were identified and can be inspected through a PTM viewer in PeptideAtlas. All matched MS/MS-derived peptide data are linked to spectral, technical and biological metadata. This new PeptideAtlas will be integrated with community resources including MaizeGDB at https://www.maizegdb.org/ and its JBrowse tracks.

Keywords: maize; proteomics; mass spectrometry; ProteomeXchange; B73; JBrowse; resource; post-translational modifications; PeptideAtlas

INTRODUCTION Maize is a crop of major agricultural importance and also serves as an important model system for the study of grasses and C4 photosynthesis (Strable and Scanlon, 2009). Different cultivars and hybrids are grown for agriculture but the B73 inbred was initially chosen for maize genome sequencing and it has become the reference maize line with the most genomic resources for the research community (Hoopes et al., 2019; Portwood et al., 2019; Schnable et al., 2009; Shamimuzzaman et al., 2020). After the first maize B73 genome assembly in 2009 (Schnable et al., 2009), several subsequent B73 assemblies followed in rapid order with RefGen_v5 being the most recent (Hufford et al., 2021; Jiao et al., 2017). In the last few years, additional maize genomes have been sequenced and assembled for more than 25 inbred lines, including W22 (Springer et al., 2018), RP125 (Nie et al., 2021), A188 (Lin et al., 2021), and 25 maize nested association mapping (NAM) founder inbred lines (Hufford et al., 2021).

These different annotated maize genomes each predict well over 40.000 protein coding genes across the 10 maize nuclear chromosomes, and these genes are represented by a larger number of transcripts. Many aspects of the cellular proteome cannot be predicted but must be experimentally determined at the protein level, such as cell-type specific and subcellular protein accumulation, protein post-translational modifications (PTMs), half-life, and protein interactions. Furthermore, even the best annotated genomes cannot easily predict which mRNA splice forms result in proteins, and MS workflows face technical challenges to identify all possible peptide covering splice junctions, see *e.g.* (Agosto et al., 2019; Wang et al., 2018). Therefore, the impact of alternative splicing for the cellular proteome in eukaryotes is still under debate (Blencowe, 2017; Chaudhary et al., 2019; Tress et al., 2017). In addition, plant genomes may also contain an unknown number of various types of small Open Reading Frames (sORFs) encoding for small proteins or peptides, that could be detected by MS (Feng et al., 2023; Hazarika et al., 2017; van Wijk et al., 2021). The use of proteomics data for plant genome annotation is summarized under the term ‘proteogenomics’ (Song et al., 2023), and has been applied for the genomes of Arabidopsis (*Arabidopsis thaliana*) (Zhang et al., 2019; Zhu et al., 2017), rice (*Oryza sativa*) (Chen et al., 2020; Ren et al., 2019), maize (Castellana et al., 2014), grape (*Vitis vinifera*) (Chapman and Bellgard, 2017), and sweet potato (*Ipomoea batatas*) (Al-Mohanna et al., 2019) and the perennial fruit cherry (Xanthopoulou et al., 2021).

Recently, we published two studies that describes a new community proteomics resource for *Arabidopsis thaliana* entitled Arabidopsis PeptideAtlas, http://www.peptideatlas.org/builds/arabidopsis/. (van Wijk et al., 2021; van Wijk et al., 2023). The purpose of this new resource is to help solve central questions about the Arabidopsis proteome, such as the significance of protein splice forms, post-translational modifications (PTMs), and obtain reliable information about specific proteins. PeptideAtlas is based on published mass spectrometry (MS) datasets collected through ProteomeXchange (http://www.proteomexchange.org/) and reanalyzed with the Trans-Proteomic Pipeline (TPP) (Deutsch et al., 2015; Deutsch et al., 2023; Keller et al., 2005)and metadata annotation pipeline. Arabidopsis PeptideAtlas is integrated with community resources including TAIR, JBrowse, PPDB and UniProtKB.

In addition to ProteomeXchange dataset (PXD) submissions for Arabidopsis, ProteomeXchange contains nearly one hundred PXDs for maize, providing an excellent opportunity to build a community proteomics resource also for maize. The current report describes the generation of the first Maize PeptideAtlas based on available maize MS/MS proteomics datasets in ProteomeXchange (cutoff date June 16 2023). In this report, we will first provide an analysis of what questions the maize community has addressed in these submitted MS-based proteome experiments and what genotypes and methodologies were pursued. Overall statistics on identified peptides, proteins and MS/MS match rates against the most recent B73 genome annotation by the maize community (https://www.maizegdb.org/genome/assembly/Zm-B73-REFERENCE-NAM-5.0) as well as previous annotations provides insight about proteome coverage and allows for comparison of B73 genome annotations (protein coding genes) across the different B73 versions in MaizeGDB and NCBI RefSeq. Finally, maize is an ancient polyploid with two subgenomes; a number of proteins are still represented by a gene on both subgenomes (Cheng et al., 2018; Li et al., 2016; Schnable et al., 2011). This new Maize PeptideAtlas could be used to determine to what extent both duplicates are detected at the proteome level and if there is any subgenome bias.

This freely available Maize PeptideAtlas provides the global community with high quality, fully reprocessed MS-based proteome information together with its metadata. PeptideAtlas differs from other databases such as PPDB and Plant PTM Viewer in that the raw MS data from laboratories around the world are reprocessed. All identified peptides, PTMs, and MS/MS (tandem MS) spectra in PeptideAtlas are linked to the metadata collected from the PXDs, publications, and additional information from the submitting labs. PeptideAtlas will be integrated with the maize community resource MaizeGDB (Portwood et al., 2019).

RESULTS AND DISCUSSION

Features of maize ProteomeXchange submissions We downloaded all available raw files from PXDs referring to maize, corn or *Zea mays* from ProteomeXchange with a cutoff date of June 16, 2023 (Supplemental Table 1). Upon inspection, some of the PXDs were not further processed because i) they only contained recombinant maize proteins, ii) were misannotated as containing maize samples, iii) the data were only peptide mass printing (PMF) acquired on a MALDI-TOF instrument, or iv) raw files were missing or corrupt. We excluded PXD00943 because it contained only data independent acquisition (DIA) as these have no MS/MS scans directly associated with MS precursor ions; all other PXDs used data dependent acquisition (DDA). While DIA datasets often have fewer missed ions per run by avoiding the stochastic precursor selection problems of DDA, FDR control is more challenging and uncertain due to the multiplexing of fragment ions (Rosenberger et al., 2017). A large ensemble of DDA runs, especially when complex peptide mixtures are pre-fractionated (*e.g*. SDS-PAGE or HPLC), are more likely to achieve high coverage with low FDR than DIA. Nonetheless, there are many efforts to improve the processing of DIA (Yu et al., 2023) and we are starting to develop a mechanism to integrate DIA data into the PeptideAtlas build process. Of the 99 PXDs that we evaluated, 74 PXDs passed our inspection criteria, their raw MS/MS data were searched, and results incorporated into the first maize PeptideAtlas build. Table 1 provides a summary of these PXDs, whereas Supplemental Table 1A provides more detailed information for these PXDs and Supplemental Table 1B provides information on the rejected PXDs not used in the build. The first maize PXDs in ProteomeXchange were released in December 2012, however, there are older submissions to PRIDE (https://www.ebi.ac.uk/pride/), one of the oldest partners in the ProteomeXchange consortium, but these were not transferred to ProteomeXchange for technical reasons. We resubmitted three of these older, published datasets from the van Wijk lab (Majeran et al., 2010; Majeran et al., 2012; Majeran et al., 2008) to ProteomeXchange such that these could be included in the Maize PeptideAtlas. Figure 1A shows the timeline of public availability of the maize PXDs in ProteomeXchange, most of which were submitted through PRIDE (70%), followed by iProX (22%), MassIVE (7%) and jPOST (1%). For most of these PXDs (90%), the MS data were acquired using an Orbitrap type instrument from the vendor Thermo (Thermo Fisher Scientific) (Figure 1B). Initially these Orbitrap instruments were mostly the early generation of LTQ Orbitrap models (Velos/XL/Elite), followed by many PXDs using one of the different versions of the Q Exactive instrument, as well as lower number of PXDs with more recent Orbitrap models (Lumos, Fusion, Exploris). The remainder of the PXDs were acquired on a variety of other instruments, *i.e.* TripleTOF 5600 and MaXis QTOF, LTQ and LTQ Velos (Figure 1B).

As indicated in Supplemental Table 1, submissions came from many different countries and continents, *i.e.* China (39 PXDs), USA (19 PXDs), Europe (12 PXDs), Japan (1 PXD) and South America (3 PXDs). The 74 PXDs address a wide range of topics, ranging from the response to various abiotic stresses (*e.g.* drought, waterlogging, salt) and biotic stresses (virus, bacteria, fungal infections), as well as biological phenomena such as de-etiolation, autophagy, development and differentiation (Table 1; Supplemental Data Set 1). Moreover, proteomes were extracted from a range of specific plant organs (*e.g.* root, leaf, pollen, silk, stem), or from specific subcellular locations and complexes (*e.g.* chloroplasts, apoplast, mitochondria, microsomes, plastid nucleoids) (Table 1; Figure 1C,D). 15 PXDs used multiplex labeling techniques (iTRAQ or TMT) for quantitative comparative proteomics (Table 1). Furthermore, 13 PXDs contained samples that were affinity-enriched to determine specific PTMs, including phosphorylation, ubiquitination, lysine acetylation, lysine 2-hydroxyisobutyrylation and lysine malonylation (Table 1). Furthermore, the PXDs cover a wide variety of maize cultivars and hybrids: i) 24 PXDs are exclusively from proteomes extracted from inbred line B73, ii) 7 PXDs are from proteins from the W22 inbred line, iii) many other PXD include samples from hybrids and their respective parents or commercial cultivars (*e.g.* Early Golden Bantam) (Table 1).

To build the PeptideAtlas, most PXDs were divided up in experimental sets of raw files in order to accommodate specific search parameters, treatments, cultivars, and sample types. This division of PXDs into experiments also allows the user to better explore the matched spectra, peptides and identified proteins within PeptideAtlas and its metadata annotation. A total of 213 experiments are assigned for these PXDs in PeptideAtlas, as described and summarized in Supplemental Table 2.

Defining the protein search space The maize community has made large investments in the B73 inbred line, including genome sequencing and annotations. Indeed, since the assembly and annotation of the first B73 genome (Schnable et al., 2009; Wei et al., 2009) there have been several successive B73 genome assemblies and annotations. The last three genome annotations by members of the maize community were RefGen_v3, RefGen_v4 (also named AGPv4) (Jiao et al., 2017) and RefGen_v5 (Zm-B73-REFERENCE-NAM-5.0) (Hufford et al., 2021) which each have their own set of (non-overlapping) gene identifiers (see Table 2 and footnotes for identifier formats). Furthermore, NCBI RefSeq also annotated these B73 physical assemblies. The informatics workflow for gene annotation of genome assemblies in RefSeq is different from those used for the maize community databases, and therefore result in only partially identical sets of protein sequences (Table 2).

Table 2. The assembly of maize protein sequences from different sources used as the protein search space, and the respective number of total, distinct, and unique sequences in each source, as well as the sequence-identical intersection among sources.

protein source	# protein identifier (all isoforms)	# Unique protein sequences	Unique in source	# identical protein sequences across maize sources							
				B73 v4	B73 v3	W22	UniProtKB	RefSeq v5	RefSeq v4	mito	chloro
B73 v5 [a]	72539	62559	3	24248	23721	5519	62552	25737	24239	6	9
B73 v4 [b]	143679	110536	11814		22495	8135	24405	22193	98560	96	36
B73 v3 [c]	63241	57882	27160			5296	23888	21548	22335	98	38
W22 [d]	51716	42082	32994				5551	5384	8133	1	0
UniProtKB Zm [e]	63541	63232	350					26037	24279	148	83
RefSeq v5 [f]	57578	45919	14572						22063	149	92
RefSeq v4 [g]	131270	98578	18							2	2
mito RefSeq [h]	163	149	0								2
chloro RefSeq [i]	111	92	0								
contamination (CONTAM) [j]	502	500	500								

[a] MaizeGDB proteome annotation from the Zea mays B73 version 5 genome assembly (https://download.maizegdb.org/Zm-B73-REFERENCE-NAM-5.0/). The naming convention of the protein identifiers is Zm00001ebxxxxxx_Pxxx.

[b] MaizeGDB proteome annotation from the Zea mays B73 version 4 genome assembly (https://download.maizegdb.org/Zm-B73-REFERENCE-GRAMENE-4.0/). The naming convention of the protein identifiers is Zm00001dxxxxxx_Pxxx; GRMZM5Gxxxxxx_Pxx (N=178).

[c] MaizeGDB proteome annotation from the Zea mays B73 version 3 genome assembly. Most identifiers have the format GRMZMxGxxxxx_Pxx, but a small set has other formats (https://download.maizegdb.org/B73_RefGen_v3/). The naming convention of the protein identifiers is GRMZMxGxxxxxx_Pxx; ACxxxxxx.x_FGPxxx; AFxxxxxx.x_FGPxxx; AGPv3_GRMZM5Gxxxxxx_Pxx; AYxxxxxx.x_FGPxxx; EFxxxxxx.x_FGPxxx.

[d] MaizeGDB proteome annotation from the Zea mays W22 cultivar genome assembly (https://download.maizegdb.org/Zm-W22-REFERENCE-NRGENE-2.0/). The naming convention of the protein identifiers is Zm00004bxxxxxx_Pxxx.

[e] UniProtKB Zea Mays reference proteome downloaded 2023-06-12 (https://www.uniprot.org/proteomes/UP000007305). The naming convention of the protein identifiers is five or 10 characters - letters & numbers; e.g. P04707, Q9XGD5, A0A3L6E0R4.

[f] NCBI RefSeq proteome annotation from the Zea mays B73 version 5 genome assembly and Annotation Release 103 (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Zea_mays/103/). The naming convention of the protein identifiers is five or 10 characters - letters & numbers; e.g. P04707, Q9XGD5, A0A3L6E0R4.

[g] NCBI RefSeq proteome annotation from the Zea mays B73 version 4 genome assembly (https://ftp.ncbi.nlm.nih.gov/genomes/genbank/plant/Zea_mays/latest_assembly_versions/GCA_000005005.6_B73_RefGen_v4/) . The naming convention of the protein identifiers is AQKxxxxx.x; AQLxxxxx.x; ONLxxxxx.x; ONMxxxxx.x.

[h] Zea Mays mitochondrial proteome from NCBI RefSeq Zea mays B73 version 5 (https://www.ncbi.nlm.nih.gov/genome/browse#!/proteins/12/838253%7CZea%20mays%20NB/mitochondrion%20MT/). The naming convention of the protein identifiers is YP_xxxxxx.1.

[i] Zea Mays plastid chloroplast proteome from NCBI RefSeq Zea mays B73 version 5 (https://www.ncbi.nlm.nih.gov/genome/browse#!/proteins/12/838253%7CZea%20mays/chloroplast%20Pltd/). The naming convention of the protein identifiers is NP_xxxxxx.1/2.

[j] PeptideAtlas custom set of 499 common contaminant proteins and additional 3 fluorescent proteins eGFP, eYFP, RFP. The naming convention of the protein identifiers is CONTAM_name.

In our search space we included the NCBI predicted maize proteomes GCA_000005005.6 (B73_RefGen_v4) with 131270 protein identifiers (including all isoforms) and the newer NCBI *Zea mays* Annotation Release 103 (based on Zm-B73-REFERENCE-NAM-5.0 GCF_902167145.1) with 72539 protein identifiers (including all isoforms; total 39,756 protein coding genes) (Table 2). The search space also included the maize UniProtKB annotation with 63236 protein identifiers; this annotation is based on the annotation from EnsemblPlants and is continually updated through curation by UniProt. The inbred line W22 is popular in maize research, especially because of the availability of mutant collections with Mutator (Mu) and Dissociation (Ds) transposable element insertions for reverse and forward genetics studies (Springer et al., 2018). Indeed, eight PXDs used protein samples from this W22 background (Table 1) and we therefore also included the MaizeGDB W22 predicted protein sequences (Springer et al., 2018) in our search space (Table 2). In addition to the 10 nuclear chromosomes, maize (like all plants) has a small plastid (chloroplast) and mitochondrial genome, each with a much smaller set of protein coding genes (Clifton et al., 2004; Maier et al., 1995a). Because several of the maize genome annotations do not include these organellar genomes, we included the predicted plastid- and mitochondrial-encoded protein sequences from NCBI RefSeq NC_001666.2 (plastid) and NC_007982.1 (mitochondria) and listed these as individual sources (Table 2; Supplemental Tables 3 and 4). Finally, we also include a set of 500 protein sequences that represent common contaminants in proteome experiments, including keratins (from human skin and hair), trypsin (from autodigestion), proteins added to media such as BSA, various affinity tags, and fluorescent proteins (*e.g.* GFP, YFP, RFP and their variants).

Table 2 shows the number of total protein identifiers (including different isoforms or gene models) and unique protein sequences for each genome annotation within our protein search space, as well as shared protein sequences between the different annotations. For instance, MaizeGDB B73 v5 has 72539 protein identifiers (including all splice forms/isoforms) of which 62559 are unique protein sequences within the v5 annotation. Just three protein sequences are only found in this source but not in any of the other sources, indicative of the overlap between sources. The total search space included 583,838 protein identifiers (including all possible isoforms) representing 226,062 unique maize amino acid sequences and the 500 unique contaminant sequences. In conclusion, we created a protein sequence search space that included the last three maize community annotations (MaizeGDB B73 v3-5), the last two NCBI-RefSeq maize annotations (NCBI RefSeq v4 and v5 *Zea mays* B73), UniProtKB *Zea mays* and the predicted maize plastid- and mitochondrial-encoded proteins (Table 2). This inclusive search space is a unique feature of PeptideAtlas (difficult to do by individual labs due the need for large computing resources) will allow comparison of different B73 genome annotations and discovery of protein coding genes.

Search results for the first Maize PeptideAtlas Unless one uses *de novo* sequencing, MS/MS data can only lead to identification of peptides and proteins by searching these MS/MS data against an assembly of predicted, putative proteins. Proteins or peptides not represented in this protein search space cannot be identified. *De novo* sequencing is in principle possible and various software have been published (reviewed in (Vitorino et al., 2020)), however it is hard to judge the quality of such searches and mapping back to proteins is challenging; searching different maize genome annotations is more efficient. Therefore, we assembled a comprehensive set of maize sequences from a variety of key sources as we described above (Table 2). Following downloading of PXD raw MS files, file conversions and sample annotations, the MS data were searched against this total protein space (see METHODS). We searched in several iterations to allow for the correct search parameters for each PXD and experiment (mostly variable and fixed PTMs). Especially PXDs that involved the use of tandem mass tag (TMT) or isobaric tags for relative and absolute quantitation (iTRAQ) labeling used for multiplexing and comparative proteomics (Table 1) required careful attention and verification of metadata.

The finalized searches and post-search processing for control of false discovery rates resulted in the matching of 120.4 million out of 444.8 million submitted MS/MS spectra, leading to the identification of 0.37 million distinct peptides (1.36 million different peptidoforms when counting different mass modifications) (Table 3). The overall match rate of MS/MS spectra to peptides (peptide spectral matches or PSMs) was 27%, but this match rate varied dramatically across PXDs from a few % to 80% (average and medium match rate is 34% and 33%) (Table 1). For those PXDs where we obtained a low match rate, we re-evaluated the search parameters to ensure that we did not overlook specific sample treatments that could affect the optimal search parameters (*e.g.* labeling techniques). The two PXDs with the lowest match rate were for peptidomics studies (leaf or apoplast) in which extracted proteins and peptides were not digested with trypsin or other enzymes (PXD006751 and PXD01780); this reflects the challenge to identify endogenous plant peptides.

To better understand the underlying data for this maize PeptideAtlas build, we calculated the frequency distributions of peptide charge state and peptide length for the PSMs (Figure 2A,B). The vast majority of matched spectra had a charge state of 2+ (63%), 3+ (29%) or 4+ (2%) and minor amounts of 1+ (5.6%) and 5+ (0.3%) (Figure 2A). We observed a wide range of matched peptide lengths, with seven amino acids being the shortest sequence allowed (Figure 2B). 99% of all matched peptides were between 7 and 35 aa in length with the most frequent peptide length of 15 aa. Figure 2C shows the number of identified distinct (non-redundant) peptides (irrespective of PTMs) as a function of peptide length (aa). The ratio between observed distinct peptides without missed cleavages and distinct peptides with missed cleavages and/or semi-tryptic peptides increased with peptide length (Figure 2C). This increasing ratio with peptide length illustrates that allowing for missed cleavages and semi-tryptic peptides increases average peptide length; it also improves protein sequence coverage and possible discovery of splice forms.

Table 3. Summary statistics on the first Maize PeptideAtlas build (Maize 2023-09)

# ProteomeXchange datasets (PXDs)	74
# Experiments	211
# MS Runs	14,925
PSM FDR threshold	0.00008
# Searched MS/MS Spectra	444,752,369
# Identified MS/MS Spectra (PSMs)	120,422,937
Match rate of MS/MS Spectra	0.3
Distinct Modified Peptides	1,355,302
Distinct Stripped Peptides	372,811
Protein Presence Levels	
Canonical (core proteome)	16,178
Canonical (non-core proteome)	2,882
Indistinguishable Representative	2,453
Insufficient Evidence	106
Marginally Distinguished	6,219

Weak	2,171
FDR identified distinct peptides	0.15%
FDR canonical proteins	<0.006%

Figure 3 shows the number of distinct (non-redundant) peptides (irrespective of PTMs) (Figure 3A) and distinct identified canonical (identified at the highest confidence level) proteins (Figure 3B) as function of the cumulative number of matched MS/MS spectra ordered by PXD identifier (from low to high or old to new) for this first Maize PeptideAtlas. Figure 3A shows that the cumulative number of distinct peptides rapidly increases with the first ~2 million matched MS/MS spectra, followed by a gradual increase of cumulative peptides up to ~90 million matched spectra. The matched spectra between 2 and 90 million (see arrow Figure 3A) all came from PXD002853 which sampled >20 tissue types using 10,417 MS runs on a lower resolution LTQ Velos (Table 1, Supplemental Table 1 (Walley et al., 2016). The yield of additional new cumulative peptide per matched spectrum then rapidly increased because these came from PXDs of very diverse proteome sample sets and using higher resolution instruments. Figure 3B shows that after a rapid increase in newly identified canonical proteins based on the first ~2 million MS/MS spectra, accumulation of additional newly identified canonical proteins gradually increased without showing any obvious saturation. Both Figure 3A and 3B indicate that future incorporation of newer PXDs into maize PeptideAtlas will likely lead to significant increase in newly identified distinct peptides and proteins. This contrasts the current status for the 2^nd^ release of the Arabidopsis PeptideAtlas https://peptideatlas.org/builds/arabidopsis/, where incorporation of additional and high quality PXD data sets resulted in only small increases in distinct peptides and canonical proteins (van Wijk et al., 2023). Details about these maize PXDs, including how many new distinct peptides and proteins they identified, can be found in Table 1 and Supplemental Table 1.

Mapping peptides to the protein search space Table 4 summarizes the proteome coverage (redundant mapping) for the different maize protein sequence sources. 73.7% of all unique protein sequences in B73 MaizeGDB v5 had peptides mapping to them (46134 out of 62559). Similar but slightly lower percentages were observed for B73 MaizeGDB v3 and v4 (68.0% and 72.3%), but similar or somewhat higher for UniprotKB (73.7%), Refseq v4 (78.1%) and RefSeq v5 (74.6%). 68% of protein sequences in W22 were identified. To simplify the protein search space, we created a ‘core proteome’ consisting of one protein isoform per protein-coding gene from MaizeGDB v5 plus the predicted plastid- and mitochondrial-encoded proteome. This predicted core proteome has 40030 different protein identifiers representing 38937 unique protein sequences. We observed peptides for 66.8% (26019) of these predicted proteins, *i.e.* for 66.8% of the protein coding genes we observed some MS-based evidence for their accumulation.

Table 4. Proteome coverage (redundant mapping) for the different protein sources in the search space.

Database	# protein ids including isoforms	# unique proteins [b]	# Observed proteins	observed unique proteins (% of total)	# unobserved proteins
Core (B73 v5 _P001, mito, chloro) [a]	40,030	38,937	26,019	67%	12,918
B73 v5	72,539	62,559	46,134	74%	16,425
mito RefSeq	163	149	39	26%	110
chloro RefSeq	111	92	70	76%	22
B73 v4	143,501	110,376	79,797	72%	30,579
B73 v3	63,419	57,882	39,368	68%	18,514
W22	51,716	42,082	28,439	68%	13,643
UniprotKB Zm	63,236	63,232	46,619	74%	16,613
RefSeq v5	57,578	45,919	34,249	75%	11,670
RefSeq v4	131,270	98,578	77,025	78%	21,553
CONTAM	502	500	315	63%	185

[a] Core proteome consists of one isoform per gene from the MaizeGDB proteome annotation from the *Zea mays* B73 version 5 genome assembly plus the mitochondrial and plastid proteomes from Refseq

[b] The number of protein with unique protein amino acid sequences

To better understand these peptide matches and to create a single Maize PeptideAtlas build, we assigned the matched MS/MS spectra in a hierarchical order with B73 MaizeGDB v5 having the first priority since we assume that the latest B73 genome annotation by the maize community is the most accurate version of the B73 genome annotation. MS/MS spectra not matched to v5, were then assigned to v4, and subsequent sources, following the hierarchy as shown in Table 5. Many genes are represented by gene models (annotated as _P00n) due to possible alternative start and stop sites, different intron/exon boundaries. These gene models are represented by different protein identifiers (xxxx_P00n) but these protein identifiers might represent identical or divergent protein sequences. Table 5 summarizes the hierarchical assignment of PSMs to the listed sources and identified peptides and proteins at different confidence categories. We will first discuss the identification of proteins in the non-core proteome (Figure 5), followed by a more in-depth discussion of the core proteome identification and its PTMs. A listing and associated information for all 2877 non-core protein identified at the highest confidence level (canonical) is provided in Supplemental Table 5. In addition, this Supplemental Table also provide a listing of all 12265 non-core canonical proteins identified at any of the confidence levels.

An example of identification of an alternative protein isoform for MaizeGDB v5 is demonstrated for Zm00001eb057170 encoding for a chloroplast-localized subunit of the thylakoid NDH complex (NDF5) involved in cyclic electron flow (Figure 5A). Sequence comparison of Zm00001eb057170_P001 (364 aa) and its second isoform Zm00001eb057170_P002 (411 aa) shows that the two isoforms are the same except for the C-terminal region, where they use different exons, and with isoform P002 in total 47 aa longer. There is no uniquely-mapping evidence in PeptideAtlas for the P001 isoform, but the P002 isoform is well supported by several uniquely-mapping peptides.

An example of a uniquely identified protein from the UniProtKB annotation is Q9XGD6 (caffeoyl-CoA O-methyltransferase) (Figure 5B). The closest homology in MaizeGDB v5 is Zm00001eb271480_P001. Compared to this v5 homolog, Q9XGD6 shows four small regions with divergence, *i.e.* two small insertions (TKTT, AG) near the N-terminus confirmed with specific identified peptides and a point mutation (N instead of E) near the C-terminus which was also confirmed with a specific identified peptide.

Figure 5C shows an example of a protein (Zm00001d034925_P001; anthocyanin 5-aromatic acyltransferase) that is present in the MaizeGDB v4 proteome annotation but is absent from MaizeGDB v5. It is very well detected with 70% sequence coverage with 28 distinct uniquely mapping peptides, shown as blue rectangles. Darker shades indicate a larger number of PSMs.

Figure 5D shows GRMZM5G895313_P01, an example of a protein (Glycine-rich protein 2b) that is present in the MaizeGDB v3 proteome annotation but is absent from MaizeGDB v4 and v5. It is very well detected with 100% sequence coverage (except for the initiating methionine) with 75 distinct uniquely mapping peptides, shown in blue (and a few short low-complexity peptides such as GGGGGGGR that also map to other proteins, shown in orange).

Figure 5E shows that Zm00004b024498_P001 in the MaizeGDB W22 proteome annotation was identified with many uniquely mapping peptides. It is very well detected with 8014 PSMs and 100% sequence coverage (except for the initiating methionine) with 36 distinct peptides, of which 24 are uniquely mapping to Zm00004b024498_P001. The skipped exon near position 65 is also encoded by a second isoform Zm00001eb209710_P002 from the B73 proteome. However, the single amino acid variants at positions 77, 90, and 102 are unique to the W22 sequence.

These examples show that the maize PeptideAtlas can serve as a resource for maize genome annotation and protein discovery.

Table 5. Peptides assigned to proteins by hierarchy of sources ranging from core to DECOY, with each peptide is assigned only to the highest source possible and then not to any other source.

Hierarchy [a]	Primary Protein Match [d]	# peptides	# PSMs	# Primary Proteins	# peptides (>=3 PSMs)	# PSMs (>=3 PSMs)	# Primary Proteins (>=3 PSMs)	# Primary Proteins (>=2 Distinct Peptides with >=3 PSMs)	Suppl Table 5 [e]
1	**Core (B73 v5 _P001, mito, chloro)**	338111	113683979	23302	252772	113569110	20854	17096	
2	**B73 v5 (_P002 and higher)**	16550	4150816	4547	12183	4144971	3798	2186	x
3	**UniProtKB Zm**	629	308998	121	449	308746	102	65	x
4	**B73 v4**	8360	1494636	2445	5756	1491178	1796	912	x
5	**B73 v3**	1017	232734	454	644	232245	287	101	x
6	**W22**	7460	508685	3529	5374	505821	2709	1161	x
7	**NCBI RefSeq (v4 & v5)**	712	44197	379	420	43807	236	80	x
8	**CONTAM [b]**	3280	1086243	266	1830	1084289	210	153	
9	**DECOY [c]**	528	37173	522	337	36939	333	4	

[a] Hierarchy refers to the order to which peptides are assigned to sources.

[b] Contaminants often found in samples, *e.g.* BSA, Keratin, trypsin, etc

[c] Decoys are all shuffled protein sequences in the search space; this enables accurate calculation of FDR.

[d] sources are defined in table 2

[e] Details for the non-core proteins identified at the highest confidence level (canonical) are provided in Supplemental Table 5

Identification of the core proteome Table 6 shows the proteins identified in the core proteome for each of the three confidence categories (canonical, uncertain, redundant – see METHODS for explanations) by nuclear chromosome (1-10), a small set of 'scaffolds' (*i.e.* predicted protein that could not be assigned to a specific chromosome location) as well as the mitochondrial and plastid chromosomes. In total, the PeptideAtlas build identified 66.2% of the predicted core proteins (as mentioned earlier, for genes with more than one predicted isoform, P001 was used by default).

The identification rate across the 10 nuclear chromosomes was very similar and ranged from 65% to 68.4%. A lower percentage was identified for the unmapped scaffolds (44.2%) likely reflecting the lower quality or confidence of these predicted protein sequences. The number of core proteins identified at the highest confidence level (canonical; FDR is <0.006%) was 16,178 which was 40% of the predicted proteome.

Table 6. Identification status of the maize core proteome in the first maize PeptideAtlas build

Chromosome	Entries (putative protein coding genes)	Canonical		Uncertain		Redundant		Not Observed	
Mitochondrial (including pseudogenes)	163	16	9.8%	9	5.5%	18	11.0%	120	73.6%
updated mitochondrial (protein coding)	42	12	28.6%	11	26.2%	12	28.6%	7	16.7%
Plastidial (including pseudogenes)	111	41	36.9%	14	12.6%	21	18.9%	35	31.5%
updated_plastidial (protein coding)	82	41	50.0%	14	17.0%	21	22.0%	6	7.3%
Scaffolds (a)	721	10	1.4%	22	3.1%	294	40.8%	395	54.8%
1	5,892	2,541	43.1%	598	10.1%	813	13.8%	1,940	32.9%
2	4,751	1,936	40.7%	491	10.3%	741	15.6%	1,583	33.3%
3	4,103	1,779	43.4%	448	10.9%	556	13.6%	1,320	32.2%
4	4,093	1,645	40.2%	390	9.5%	651	15.9%	1,407	34.4%
5	4,485	1,910	42.6%	470	10.5%	681	15.2%	1,424	31.8%
6	3,412	1,329	39.0%	381	11.2%	507	14.9%	1,195	35.0%
7	3,070	1,316	42.9%	296	9.6%	456	14.9%	1,002	32.6%
8	3,536	1,458	41.2%	359	10.2%	600	17.0%	1,119	31.6%
9	2,988	1,139	38.1%	328	11.0%	477	16.0%	1,044	34.9%
10	2,705	1,058	39.1%	263	9.7%	429	15.9%	955	35.3%
Total (including pseudogenes)	40,030	16,178	40.4%	4,069	10.2%	6,244	15.6%	13,539	33.8%

(a) remaining contigs that were not assembled into any of the 10 chromosomes

The predicted sets of plastid and mitochondrial encoded proteins are lacking in the MaizeGDB v5 since this annotation concerns only the 10 nuclear chromosomes (see Table 2). We do note that plant nuclear genomes include fragments of plastid and mitochondrial genomes, most of which are considered pseudogenes but do provide a driving force in evolution (Noutsos et al., 2007). Some of these organellar genome remnants are annotated as protein coding in the nuclear chromosomes. We therefore used the RefSeq annotations that are derived from (Clifton et al., 2004; Maier et al., 1995a), which list 163 and 111 protein identifiers (one per gene) that encode for 159 and 92 unique sequences for mitochondrial and plastids, respectively (Table 2). However, it is quite well established that the number of mitochondrial-encoded proteins in maize and higher plants typically includes around 33 conserved proteins (a dozen of which can be found duplicated) and the two optional ORFs T-urf13 and Orf355/Orf77 associated with cytoplasmic sterility (Allen et al., 2007; Clifton et al., 2004). We verified and updated the NCBI RefSeq annotations (see Supplemental Table 3) and list 42 annotated protein identifiers that include 5 pairs of identical sequences, and 3 pairs of closely related sequences. This also includes a pair of identifiers closely related to the protein with unknown function GRMZM5G892769_P01 in MaizeGDB v3. PSMs are reported for all but seven of these 42 proteins; the seven proteins are Nad4l, three of the cytc biogenesis proteins (Cmb, Ccmc, Ccmfc), Matr, Rpl6 and the ORF TatC/MttB. It is not surprising that no PSMs were obtained for these 7 proteins, since none of them have suitable predicted tryptic peptides that can be detected by MS (Supplemental Table 4). 121 of the protein identifiers are annotated as pseudogenes, based on the information from (Allen et al., 2007). Indeed, for 114 of these annotated pseudogenes, not a single MS/MS spectrum was found. For three pseudogenes, we found one small peptide but this peptide was also matched to more confidently identified proteins. Interestingly, protein products for four annotated pseudogenes were identified at the canonical level (YP_588312.1, YP_588286.1, YP_588271.1, YP_588349.1) and closer inspection suggested that several of them appear *bona fide* identifications. These four proteins are glutharedoxin-like, two putative DNA polymerase B (PolB) and one putative RNA polymerase; for all four an identical protein was annotated in B73 v3 (See Supplemental Table 3). Previous reports stated that these polymerases genes are located on plasmids and that they are degenerate sequences (Allen et al., 2007). However, the data in PeptideAtlas indicate that these proteins clearly do accumulate; the significance of these proteins remains to be determined. Altogether, the information in PeptideAtlas for the mitochondrial-encoded proteins can be used to help generate a better predicted set of expressed mitochondrial-encoded proteins and could be combined with carefully assembled set of mRNA sequences *e.g.* by RNAseq and RiboSeq of purified mitochondria.

The 111 plastid identifiers include 29 pseudogenes (annotated as hypothetical proteins in RefSeq) and the remaining 82 identifiers representing established chloroplast proteins

(Supplemental Table 4). Within these 82 identifiers are five sets of duplicated proteins because of duplication from the inverted repeats. PeptideAtlas identified all but six of these established proteins. These unidentified proteins are the small and hydrophobic integral thylakoid membrane proteins (between 29-40 aa length; gravy index between 0.68 and 1.46) PetG, PetL, PetN of the cytb6f complex and PsbI and PsbJ of Photosystem II, as well as NdhG/Ndh6 (176 aa length, five predicted transmembrane domains, 1.04 gravy index). These are likely not identified because they lack favorable peptides for detection by MS. The number of PSMs ranged from 0.6 million (RBCL) to just one (PsbM). Eleven plastid proteins had identical proteins in the MaizeGDB v5 annotation and two proteins (RbcL and ClpP1) had highly similar protein in the MaizeGDB v5 annotation, even if it is well-established that these plastid proteins are encoded by the plastid genome. These nuclear protein identifiers are likely pseudogenes from plastid genome fragments that are part of the nuclear genome. Finally, we also note that maize lacks plastid-encoded Ycf1, Ycf2 and accD observed in Arabidopsis plastids (Maier et al., 1995b) and reported in the Arabidopsis PeptideAtlas (van Wijk, 2023).

Identification of physiological PTMs in the core proteome Plant proteins can undergo a number of post-translational modifications (PTMs) *in vivo* which can be irreversible (*e.g*. protein processing) or reversable (*e.g*. phosphorylation) (for reviews see (Chen et al., 2021; Friso and van Wijk, 2015; Millar et al., 2019; Willems et al., 2019)). Many of these PTMs can be measured by mass spectrometry but often require specific affinity enrichment because they are present in sub-stoichiometric amounts or because they reduce the detectability of the peptide. A subset of the maize PXDs included such affinity enriched PTM datasets for phosphorylation (10 PXDs), ubiquitination (PXD007880), and three different lysine acylations, *i.e*. acetylation (PXD0146033), malonylation (PXD027417) and hydroxybutyrylation (PXD030131). The lysine ubiquitination and acetylation PTM analysis were both done in the context of the same course of maize leaf de-etiolation (B73) (Wang et al., 2019; Yan et al., 2020). Supplemental Tables 6-11 provide a summarizing information for the maize canonical core proteins for which the PeptideAtlas pipeline identified phosphorylation (S,T,Y) (Supplemental Table 6), N-terminal acetylation (Supplemental Table 7), lysine acetylation (Supplemental Table 8), lysine ubiquitination (Supplemental Table 9), hydroxybutyrylation (Supplemental Table 10), malonylation (Supplemental Table 11). PTM sites and number of PSMs per site within different p-value confidence intervals are also provided in these tables. The PTM viewer in PeptideAtlas allows for further investigation of all PTM-site assignments (including at lower confidence intervals) and associated MS/MS spectra as well as metadata.

For our analysis described here (Figure 5), we only considered those PTMs for which the site assignments were of high confidence, *i.e*. in the probability intervals 0.95 < p ≤ 0.99, 0.99 < p ≤ 1.00 and no choice (there was only one possible site for that PTM in the peptide). In total we observed 7419 canonical core proteins with one or more of these PTMs (46% of all 16178 observed canonical core proteins) and the overlap between proteins with phosphorylation, N-terminal acetylation or lysine PTMs (any of the four) is shown in the Venn diagrams (Figure 5A). This shows that the number of observed phosphoproteins was by far the highest at 6053, followed by 1845 proteins for which their N-terminus was acetylated (NTA), and 1484 proteins with one or more PTMs on the ε-amino of the lysine sidechains.

Figure 5B illustrates shows the number of phosphosites at the three highest probability intervals for specific aa site assignment of the phosphate group for the canonical core proteome. The number of PSMs at the two highest probability intervals was about the same. There was total number of 36,451 phospho-sites (on average 6 p-sites per phosphoprotein) observed by 2.59 million PSMs. The ratio between pS, pT and pY sites was 86:13:1 which is consistent with other publications for meta-phosphoproteomics in Arabidopsis (Mergner et al., 2020; van Wijk et al., 2014; van Wijk et al., 2023).

The lysine side-chain amine (ε-amine) can undergo a range of PTMs, including (poly)ubiquitination (Vere et al., 2020) and a dozen types of different acylations in particular in histones and a smaller number of types in non-histones (Barnes et al., 2019; Sabari et al., 2017; Xu et al., 2022). Malonylation, acetylation and 2-hydroxyisobutyrylation can be enzymatically catalyzed by lysine acyltransferases (KATs) and removed by lysine deacylases (KDACs). The functional significance for these acyl modifications is still very poorly understood. In contrast protein ubiquitination is typically a signal for degradation by the 26S proteasome, in particular in case of polyubiquitination and K48 linkages between the ubiquitin moieties. However, in most proteomics workflows, including in the PXD included here, only a di-glycyl (GG) remnants remains and hence information on whether the lysine was mono-ubiquitination or polyubiquitinated (and its linkage type) is lost. Figure 5B shows a Venn diagram for the proteins containing these different lysine modifications and ubiquitination. There was a significant overlap in proteins with more than one of these lysine PTMs, with 83 proteins have all four PTM. It should be pointed out that we did not apply a minimum number of PSMs for any of these PTMs – hence closer inspection and consideration how the frequency of observation will be important if readers or users of the maize PeptideAtlas are interested to investigate specific proteins or PTMs. The Supplemental Tables 5-11 and of course the Maize PeptideAtlas itself provide such information.

Mass modifications typically due to sample preparation In addition to biological PTMs (which require specific affinity enrichment for detection, except for N-terminal acetylation), the MS searches also include additional mass modifications that are induced during sample processing (see Methods). These modifications have generally very little biological relevance. The frequencies of these modifications can greatly vary between PXDs and experiments within PXDs depending on *e.g.* the use of organic solvents, urea, oxidizing conditions, temperature, alkylating reagent (alkylation of other residues than the intended cysteines), pH and use of SDS-PAGE gels. These mass modifications are included in the search parameters since many of these modified peptides would otherwise not be identified or lead to false assignments. The frequencies of these mass modifications (calculated as PSMs with the mass modification normalized to the total number of PSMs) are summarized in Figure 5D. This shows that the oxidation of methionine is by far the most frequent (4.5% of all PSMs), followed by deamidation of asparagine (0.9%), pyro-glutamate from N-terminal glutamate (0.7%) and deamidation of glutamine (0.2%), and very low levels (<0.1%) of tryptophan and histidine oxidation, formylation of threonine and serine, and threonine (0.7%), oxidation of proline (0.7%), pyro-glutamate from N-terminal glutamate and pyro-carbamidomethylation of cysteine. Carbamidomethylation of cysteine searched as fixed modification due to standard treatments with alkylating agents was 4.9%. These mass modifications are also visible in the PeptideAtlas web interface with viewable spectra and their interpretations.

The first maize PeptideAtlas build and insights This is the first effort to collect and reprocess all publicly available maize protein mass spectrometry data to provide a resource for the maize community. This resource will allow users to investigate proteins for their detection across many sample types, their protein sequence coverage which also might provide insight in post-translational processing (*e.g*. removal of the N-terminal signal peptide), accumulation of specific splice forms and possible PTMs. Moreover, because different B73 genome assemblies and annotations were searched, the Maize PeptideAtlas can be used to evaluate the quality and relevance of these different annotations, since it is unlikely that these annotations of the most recent B73 genome assembly are the final and perfect annotations. Mapping of spectra against the W22 genome annotation can be used to evaluate the quality of its predicted proteome, especially when considering samples from W22 plants.

Comparison to the Arabidopsis PeptideAtlas Compared to our most recent *Arabidopsis thaliana* PeptideAtlas (2023-10 release) (van Wijk et al., 2023) with an identification rate of 78.6%

of the predicted proteome (counting one isoform per protein coding gene) across all confidence tiers, the percentage of identified maize proteins (counting one isoform per protein coding gene) in the latest B73 annotation was significantly lower (66.2%). This is likely a reflection of the lower amount of MS/MS data acquired on high resolution mass spectrometry instruments. The number of PXDs and associated publications for Arabidopsis is at least 5-fold higher than for maize and the latest Arabidopsis PeptideAtlas build includes 115 PXDs whereas this first maize build is based on 74 PXDs. Figure 3B shows that the cumulative number of detected proteins in the build continues to increase substantially even for the final dataset, whereas the analogous Figure 2B in (van Wijk et al., 2023) for the 2023-10 Arabidopsis build shows a clear saturation in the number of detected proteins. It is also quite likely that the higher complexity of the maize genome compared to Arabidopsis might further affect the identification rate in PeptideAtlas. However, as the number of high-quality maize PXD submissions will undoubtedly increase in the coming years, a future maize PeptideAtlas build will cover a higher percentage of the predicted proteome.

METHODS

Selection and downloads of ProteomeXchange submissions We downloaded all available raw files with information referring to maize or corn (*Zea mays*) from ProteomeXchange with a cutoff date of June 16, 2023. Upon inspection, some of the PXDs were not further processed because i) they only contained recombinant maize proteins, ii) were miss-annotated as containing maize samples, iii) the data were only peptide mass printing (PMF) acquired on a MALDI-TOF instrument, or iv) raw files were missing. Supplemental Table 1 provides the list of all PXDs, number of raw files and MS/MS spectra (searched and matched), genotypes of the samples, identified proteins and peptides, submitting lab and associated publication, as well as several informative key words.

Extraction and annotation of metadata For each selected PXD, we obtained information associated with the submission, as well as the publication if available. This information was used to determine search parameters and provide meaningful tags that describe the samples in some detail. These tags are visible for the relevant proteins in the PeptideAtlas. If needed, we contacted the submitters for more information about the raw files. To facilitate the metadata assignments and association to specific raw files, we developed a metadata annotation system to provide detailed information to each matched spectrum for the users of the PeptideAtlas, and these metadata can be viewed in the Maize PeptideAtlas.

Assembly of protein search space We assembled a comprehensive protein search space (Table 2) comprising the predicted *Zea mays* protein sequences from https://www.maizegdb.org/ for the maize cultivar B73 (Zm-B73-REFERENCE-NAM-5.0 (aka B73 RefGen_v5 or MaizeGDB B73 v5), Zm-B73-REFERENCE-GRAMENE-4.0 (aka B73 RefGen_v4 or MaizeGDB B73 v4), B73 RefGen_v3 or MaizeGDB B73 v3 , and maize cultivar W22 (Zm-W22-REFERENCE-NRGENE-2.0), ii) UniProtKB Zm (UniProt, 2021), iii) NCBI RefSeq v4 and v5 (https://www.ncbi.nlm.nih.gov/refseq) (Li et al., 2021), iv) the mitochondrial genome (B37N genotype - NC_007982.1 - original paper (Clifton et al., 2004)), v) the plastid genome (genotype might be in the German PhD thesis of Fritsche 1988 NC_001666.2 – original paper (Maier et al., 1995a), and vi) 500 contaminant protein sequences (*e.g.* keratins, trypsin, BSA) and GFP, RFP and YFP protein sequences commonly used as reporters and affinity enrichments frequently observed in proteome samples. The footnotes in Table 2 provide the www links and information about the format of the protein identifier for each source.

The Trans-Proteomic Pipeline (TPP) data processing pipeline For all selected datasets, the vendor-format raw files were downloaded from the hosting ProteomeXchange repository, converted to mzML files (Martens et al., 2011) using ThermoRawFileParser (Hulstaert et al., 2020) for Thermo Fisher Scientific instruments or the msconvert tool from the ProteoWizard toolkit (Chambers et al., 2012) for SCIEX wiff files, and then analyzed with the TPP (Deutsch et al., 2015; Keller et al., 2005). The TPP analysis consisted of sequence database searching with either Comet (Eng and Deutsch, 2020) for LTQ-based fragmentation spectra or MSFragger (Kong et al., 2017) for higher resolution fragmentation spectra and post-search validation with several additional TPP tools as follows: PeptideProphet (Keller et al., 2002) was run to assign probabilities of being correct for each peptide-spectrum match (PSM) using semi-parametric modeling of the search engine expect scores with z-score accurate mass modeling of precursor m/z deltas. These probabilities were further refined via corroboration with other PSMs, such as multiple PSMs to the same peptide sequence but different peptidoforms or charge states, using the iProphet tool (Shteynberg et al., 2011). For most datasets in which trypsin was used as the protease to cleave proteins into peptides, two parallel searches were performed, one with full tryptic specificity and one with semi-tryptic specificity (except for PXD002853, PXD002877, and PXD006751 from LTQ instruments with in total ~90 million MS/MS spectra, which were searched only tryptic due to computational resource constraints). The semi-tryptic searches were carried out by default with the following possible variable modifications (5 max per peptide): oxidation of Met, Trp (+15.9949), peptide N-terminal Gln to pyro-Glu (-17.0265), peptide N-terminal Glu to pyro-Glu (-18.0106), deamidation of Asn or Gln (+0.9840), protein N-terminal acetylation (+42.0106), and if peptides were specifically affinity enriched for phosphopeptides, also phosphorylation of Ser, Thr or Tyr (+79.9663). For the fully tryptic searches, we also added oxidation of His (+15.9949) and formylation of peptide N-termini, Ser, or Thr (+27.9949) - we deliberately restricted these mass modifications to only fully tryptic (rather than also allowing semi-tryptic) to reduce the search space and computational needs. Formylation is a very common chemical modification that occurs in extracted proteins/peptides during sample processing, whereas His oxidation is observed less frequently, but nevertheless at significant levels (Hawkins and Davies, 2019; Verrastro et al., 2015). In both semi-tryptic and fully tryptic searches, fixed modifications for carbamidomethylation of Cys (+57.0215) if treated with reductant and iodoacetamide. Isobaric tag modifications (TMT, iTRAQ) were applied as appropriate. Four missed cleavages were allowed (RP or KP do not count as a missed cleavage). Several datasets were generated with a combination of LysC and trypsin, but no other proteases were used. Some of the datasets (PXD006751, PXD017080, PXD017081) contain the analysis of extracted peptidomes in which no protease treatment was used and these datasets were searched with 'no enzyme'. Several experiments required special additional mass modifications due to the sample preparation. We have tabulated the exact mass modifications used for each experiment in detail in Supplemental Dataset 2 via a set of single-character keys (ABCD…) that denote each mass modification pattern; each dataset has a combination of keys to provide the complete set of parameters.

PeptideAtlas Assembly In order to create the combined PeptideAtlas build of all experiments, all datasets were thresholded at a probability that yields an iProphet model-based FDR of 0.001 at the peptide level. The exact probability varies from experiment to experiment depending on how well the modeling can separate correct from incorrect. This probability threshold is typically greater than 0.99. As more and more experiments are combined, the total FDR increases unless the threshold is made more stringent (Deutsch et al., 2016). The final iProphet model-based peptide sequence level FDR across all experiments is 0.001, corresponding to a PSM-level false discover rate (FDR) of 0.0001. Throughout the procedure, decoy identifications are retained and then used to compute final decoy-based FDRs. The decoy-based PSM-level FDR is 0.00008, FDR for identified distinct peptides is 0.15%, and the final protein-level FDR is below 0.006%. Because of the tiered system, quality MS/MS spectra that are matched to a peptide are never lost, even if a single matched peptide by itself cannot confidently identify a protein.

Protein identification confidence levels and classification For the PeptideAtlas projects, proteins are identified at different confidence levels using standardized assignments to different confidence levels based on various attributes and relationships to other proteins using a relatively complex but precise ten-tier system developed over many years for the human proteome PeptideAtlas (Farrah et al., 2011) and recently also applied to the Arabidopsis PeptideAtlas (van Wijk et al., 2021). Details are provided in Table 2 in (van Wijk et al., 2021). We also simplified this ten-tier system to a simpler four category system which is more accessible to non-experts, and use this to summarize most of our findings for the maize PeptideAtlas, similar as we did for Arabidopsis (van Wijk et al., 2021; van Wijk et al., 2023). In the simpler four-category system, proteins that have no uniquely mapping peptides but do not qualify as canonical (same as Tier 1) are categorized as 'uncertain', corresponding to the sum of tiers 2-6. Proteins are categorized as 'redundant' if they have only shared peptides that can be assigned to other entries and thus these proteins are not needed to explain the observed peptide evidence (tiers 7-9). Finally, all other proteins that completely lack any peptides observed at our minimum PSM significance threshold are categorized as 'not observed' (tier 10). For all protein identifications and categorizations, all peptides must first meet the stringent PSM threshold already described above. For both systems, the highest confidence level category is the "canonical" category (Tier 1), which requires at least two uniquely-mapping non-nested (one not inside the other) peptides of at least nine amino acids in length with a total coverage of at least 18 amino acids, as required by the HPP guidelines (Deutsch et al., 2019).

Handling of gene models and splice forms. The protein coding genes in maize Ref_gen-v5 are represented by gene models (transcript isoforms), which are identified by the term 'P00n' after an underscore at the end of Zm identifier (*e.g.* Zm00001eb370310_P001). Protein identifier formats for the other protein sources are all different and listed in the footnotes of Table 2. We refer to the translations of these gene models as protein isoforms. Most protein isoforms are very similar (differing only a few amino acid residues often at the N- or C-terminus) or even identical at the protein level. It is often hard to distinguish between different protein isoforms due to the incomplete sequence coverage inherent to most MS proteomics workflows. For the assignment of canonical proteins (at least two uniquely mapping peptides identified), we selected by default only one of the protein isoforms as the canonical protein; this was the 'P001' isoform. However, if other protein isoforms did have detected peptides that are unique from the canonical protein isoform (*e.g.* perhaps due to a different exon), then they can be given tier 1 or less confident tier status depending on the nature of the additional uniquely mapping peptides (length and numbers). If the other protein isoforms do not have any uniquely mapping peptides amongst all protein isoforms (for that gene), then they are classified as redundant (tiers 7-9 in the more complex system).

Integration of PeptideAtlas results in other web-based resources PeptideAtlas is accessible through its web interface at https://peptideatlas.org. Furthermore, direct links for each protein identifier (B73 v5) are provided between PeptideAtlas and PPDB (http://ppdb.tc.cornell.edu/) and will be provided in UniProtKB (https://www.uniprot.org/) (Jan 2024). Links to matched peptide entries in PeptideAtlas will soon be available in the B73 v5 maize annotated genome through tracks in JBrowse at MaizeGDB (Portwood et al., 2019).

DATA AVAILABILITY All data are available via the free-of-charge maize PeptideAtlas interface.

FUNDING This project was primarily supported by a grant from the National Science Foundation IOS #1922871 to K.J.V.W., E.W.D., and Q.S.

AUTHOR CONTRIBUTIONS T.L. carried out all MS searches to create the PeptideAtlas build. Q.S. supported various steps in selecting and annotation of the maize sequence sources and established the links with the PPDB. Z.S. developed PeptideAtlas interface enhancements and assisted with the PeptideAtlas build process. L.M. developed PeptideAtlas interface enhancements and the dataset annotation tool. I.G., E.D., G.S., P.R. helped with incorporation of the metadata from the PXDs and associated publications into the PeptideAtlas internal annotation system. E.D. supervised the PeptideAtlas building process. K.J.V.W. contributed to the selection of PXDs and all aspects of specific plant biology-related issues. E.W.D. and K.J.V.W. developed this project, raised the funding, and wrote the paper.

ACKNOWLEDGEMENT We thank the MaizeGDB team for their discussions and for incorporating the maize PeptideAtlas peptides into the maize B73 JBrowser.

DECLARATION OF INTERESTS No conflicts of interest.

TABLES

Table 1. Summarizing information of the 4 selected PXD datasets for the maize PeptideAtlas build. This includes PXD #, publication, number of matched MS/MS spectra and % match rate, the number of identified proteins (canonical and groups of proteins), the number of matched

distinct MS/MS peptides, the MS instrument, information about the sample (plant part, subcellular fraction, enrichment for PTMs. An extended table with additional information is provided as Supplemental Table 1.

Table 2. The assembly of maize protein sequences from different sources used as the protein search space, and the respective number of total, distinct, and unique sequences in each source, as well as the sequence-identical intersection among sources.

Table 3. Summary statistics on the first maize PeptideAtlas build (Maize 2023-09).

Table 4. Proteome coverage for the different protein sources in the search space.

Table 5. Peptides assigned to proteins by hierarchy of sources ranging from core to DECOY, with each peptide is assigned only to the highest source possible and then not to any other source.

Table 6. Identification status of the maize core proteome in the first maize PeptideAtlas build.

FIGURE LEGENDS

Figure 1. Features of maize PDXs used for the first maize PeptideAtlas build.

(**A**) Cumulative PXD for maize in ProteomeXchange passing our selection criteria across the different years for maize.

(**B**) Mass spectrometry instruments used to acquire data in the maize PXDs used in the maize PeptideAtlas grouped in different categories by vendor and model.

(**C**) Distribution of PSMs of samples from different plant parts in the maize PeptideAtlas

(D) Distribution of PSMs of samples for specific flower parts and seeds.

Figure 2. Key statistics of matched MS/MS data (PSMs) for this maize PeptideAtlas build.

(**A**) Frequency distribution of PSMs for peptide charge state (z).

(**B**) Frequency distribution of PSMs for peptide length (aa). Note that when R or K is followed by P, trypsin does not cleave and hence these are not counted towards missed cleavages.

(**C**) Number of distinct peptides as a function of peptide length (aa).

Figure 3. Number of distinct (non-redundant) peptides (left panel) and identified canonical proteins (right panel) as a function of the cumulative number of PSMs (peptide-spectrum matches) for the first maize PeptideAtlas. The cumulative count is ordered by PXD identifier (from low to high or old to new). The build is based on 211 experiments across the 74 selected PXDs, where each PXD may be decomposed into several experiments/samples when such information can be determined). The PSM FDR is 8.10^{-5}.

(A) Number of distinct (non-redundant) peptides as function of the cumulative number of MS/MS spectra matched. 372,811 distinct peptides are identified at a peptide-level FDR of 0.15%. Blue rectangles represent the cumulative number of distinct peptides as experiments are added to the build, whereas orange rectangles represent the number of distinct peptides in each experiment. The arrow from 2 million to 90.7 million MS/MS spectra are from PXD002853 acquired on a lower resolution LTQ-Velos instrument.

(B) Number of distinct (non-redundant) canonical proteins as function of the cumulative number of MS/MS spectra matched at a canonical protein-level FDR of <0.006%. Blue rectangles represent the cumulative number of canonical proteins as experiments are added to the build, whereas red rectangles represent the number of canonical proteins in each experiment.

Figure 4. Examples of proteins that are well detected in the Maize PeptideAtlas that are not part of the MaizeGDB v5 core proteome.

(A) Sequence comparison of Zm00001eb057170_P001 (NDH-dependent cyclic electron flow 5) and its second isoform Zm00001eb057170_P002 (both from MaizeGDB v5). The two isoforms are the same except for the C-terminal region, where they use different exons. There is no support in PeptideAtlas for the P001 isoform, but the P002 isoform is well supported by several peptides.

(B) Sequence comparison of UniProt Q9XGD6 (Caffeoyl-CoA O-methyltransferase 1), UniProt Q9XGD6 (Caffeoyl-CoA O-methyltransferase 2) and Zm00001eb271480_P001, the closest entry in MaizeGDB v5. Although Q9XGD6 and Zm00001eb271480_P001 are highly similar (except for a V – I difference at position 69), there is no analog for Q9XGD6 in MaizeGDB v5.

(C) Example of a protein (Zm00001d034925_P001) that is present in the MaizeGDB v4 proteome annotation but is absent from MaizeGDB v5. It is very well detected with 70% sequences coverage with 28 distinct uniquely mapping peptides, shown as blue rectangles. Darker shades indicate a larger number of PSMs.

(D) Example of a protein (GRMZM5G895313_P01) that is present in the MaizeGDB v3 proteome annotation but is absent from MaizeGDB v4 and v5. It is very well detected with 100% sequence coverage (except for the initiating methionine) with 75 distinct uniquely mapping peptides, shown in blue (and a few short low-complexity peptides such as GGGGGGGR that also map to other proteins, shown in orange).

(E) Example of a protein (Zm00004b024498_P001) in the MaizeGDB W22 proteome annotation that has many uniquely mapping peptides. It is very well detected with 8014 PSMs and 100% sequence coverage (except for the initiating methionine) with 36 distinct peptides, of which 24 are uniquely mapping to Zm00004b024498_P001. The skipped exon near position 65 is also encoded by a second isoform Zm00001eb209710_P002 from the B73 proteome. However, the single amino acid variants at positions 77, 90, and 102 are unique to the W22 sequence.

Figure 5. Canonical maize core proteins with one or more PTMs and frequency of PSMs with other mass modifications.

(**A**) Overlap between proteins that carry one or more phosphorylations (S, T or Y), N-terminal protein acetylation (NTA), or one or more lysine modifications (K-ubiquitination, K-acetylation, K-malonylation, K-hydroxybutyrylation). There in total there are 6053 proteins with one or more phosphorylations, 2017 proteins with N-terminal acetylation and 1484 proteins with one or more lysine side chain modifications (acylation or ubiquitination).

(B) Number of serine, threonine and tyrosine phosphorylation sites (P-sites) at different p-value intervals for the core proteome.

(**C**) Overlap between proteins with one or more lysine modifications. Colors: Acetylation - purple, Ubiquitination - yellow, Hydroxyisobutyrylation - red, Malonylation – blue.

(**D**) Percentage of PSMs (of total) with mass modifications mostly due to sample preparations. Numbers are computed as the total number of PSMs that include at least one instance of the listed mass modification. Some PSMs contain more than one mass modification of the same type (not multiple counted) or different type (multiple counted).

SUPPLEMENTARY INFORMATION

Supplemental Table 1. Summarizing table all maize PXDs in ProteomeXchange (cutoff date 16 June 2023) with information about the MS instrument, sample (*e.g.* subcellular proteome, plant organ), number of raw files and MS/MS spectra (searched and matched), identified proteins and peptides, submitting lab and associated publication, as well as several informative key words. A separate worksheet provides information on the PXDs that did not pass our criteria for incorporation into the maize PeptideAtlas.

Supplemental Table 2. Experiment list and search key for PTMs and other mass modifications.

Supplemental Table 3. Mitochondrial-encoded protein identifiers from NCBI-RefSeq and their annotation, features and observation in PeptideAtlas.

Supplemental Table 4. Plastid-encoded protein identifiers from NCBI-RefSeq and their annotation, features and observation in PeptideAtlas.

Supplemental Table 5. Listing and associated information for all non-core protein identified at all tiers or at the highest confidence level (canonical).

Supplemental Table 6. Identification of phosphorylation (S,T,Y) sites in canonical core proteins in PeptideAtlas. Sites listed have at least 1 PSM for the PTM across the three highest probability PTM site scoring intervals (0.95 < p ≤ 0.99, 0.99 < p ≤ 1.00 and no choice).

Supplemental Table 7. Identification of N-terminal acetylation (NTA) sites in canonical core proteins in PeptideAtlas. Sites listed have at least 1 PSM for the PTM across the three highest probability PTM site scoring intervals (0.95 < p ≤ 0.99, 0.99 < p ≤ 1.00 and no choice).

Supplemental Table 8. Identification of lysine acetylation (Kac) sites in canonical core proteins in PeptideAtlas. Sites listed have at least 1 PSM for the PTM across the three highest probability PTM site scoring intervals (0.95 < p ≤ 0.99, 0.99 < p ≤ 1.00 and no choice).

Supplemental Table 9. Identification of ubiquitination sites in canonical core proteins in PeptideAtlas. Sites listed have at least 1 PSM for the PTM across the three highest probability PTM site scoring intervals (0.95 < p ≤ 0.99, 0.99 < p ≤ 1.00 and no choice).

Supplemental Table 10. Identification of lysine hydroxisobutyrylation sites in canonical core proteins in PeptideAtlas. Sites listed have at least 1 PSM for the PTM across the three highest probability PTM site scoring intervals (0.95 < p ≤ 0.99, 0.99 < p ≤ 1.00 and no choice).

Supplemental Table 11. Identification of lysine malonylation sites in canonical core proteins in PeptideAtlas. Sites listed have at least 1 PSM for the PTM across the three highest probability PTM site scoring intervals (0.95 < p ≤ 0.99, 0.99 < p ≤ 1.00 and no choice).

REFERENCES

Agosto, L.M., Gazzara, M.R., Radens, C.M., Sidoli, S., Baeza, J., Garcia, B.A., and Lynch, K.W. (2019). Deep profiling and custom databases improve detection of proteoforms generated by alternative splicing. Genome Res 29:2046-2055.

Al-Mohanna, T., Ahsan, N., Bokros, N.T., Dimlioglu, G., Reddy, K.R., Shankle, M., Popescu, G.V., and Popescu, S.C. (2019). Proteomics and Proteogenomics Analysis of Sweetpotato (Ipomoea batatas) Leaf and Root. J Proteome Res 18:2719-2734.

Allen, J.O., Fauron, C.M., Minx, P., Roark, L., Oddiraju, S., Lin, G.N., Meyer, L., Sun, H., Kim, K., Wang, C., et al. (2007). Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. Genetics 177:1173-1192.

Barnes, C.E., English, D.M., and Cowley, S.M. (2019). Acetylation & Co: an expanding repertoire of histone acylations regulates chromatin and transcription. Essays Biochem 63:97-107.

Blencowe, B.J. (2017). The Relationship between Alternative Splicing and Proteomic Complexity. Trends in Biochemical Sciences 42:407-408.

Castellana, N.E., Shen, Z., He, Y., Walley, J.W., Cassidy, C.J., Briggs, S.P., and Bafna, V. (2014). An automated proteogenomic method uses mass spectrometry to reveal novel genes in Zea mays. Mol Cell Proteomics 13:157-167.

Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., et al. (2012). A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 30:918-920.

Chapman, B., and Bellgard, M. (2017). Plant Proteogenomics: Improvements to the Grapevine Genome Annotation. Proteomics 17.

Chaudhary, S., Jabre, I., Reddy, A.S.N., Staiger, D., and Syed, N.H. (2019). Perspective on Alternative Splicing and Proteome Complexity in Plants. Trends Plant Sci 24:496-506.

Chen, M.X., Zhu, F.Y., Gao, B., Ma, K.L., Zhang, Y., Fernie, A.R., Chen, X., Dai, L., Ye, N.H., Zhang, X., et al. (2020). Full-Length Transcript-Based Proteogenomics of Rice Improves Its Genome and Proteome Annotation. Plant Physiol 182:1510-1526.

Chen, Y., Wang, Y., Yang, J., Zhou, W., and Dai, S. (2021). Exploring the diversity of plant proteome. J Integr Plant Biol 63:1197-1210.

Cheng, F., Wu, J., Cai, X., Liang, J., Freeling, M., and Wang, X. (2018). Gene retention, fractionation and subgenome differences in polyploid plants. Nat Plants 4:258-268.

Clifton, S.W., Minx, P., Fauron, C.M., Gibson, M., Allen, J.O., Sun, H., Thompson, M., Barbazuk, W.B., Kanuganti, S., Tayloe, C., et al. (2004). Sequence and comparative analysis of the maize NB mitochondrial genome. Plant Physiol 136:3486-3503.

Deutsch, E.W., Lane, L., Overall, C.M., Bandeira, N., Baker, M.S., Pineau, C., Moritz, R.L., Corrales, F., Orchard, S., Van Eyk, J.E., et al. (2019). Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0. J Proteome Res 18:4108-4116.

Deutsch, E.W., Mendoza, L., Shteynberg, D., Slagel, J., Sun, Z., and Moritz, R.L. (2015). Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. Proteomics Clin Appl 9:745-754.

Deutsch, E.W., Mendoza, L., Shteynberg, D.D., Hoopmann, M.R., Sun, Z., Eng, J.K., and Moritz, R.L. (2023). Trans-Proteomic Pipeline: Robust Mass Spectrometry-Based Proteomics Data Analysis Suite. J Proteome Res doi: 10.1021/acs.jproteome.2c00748. Online ahead of print.

Deutsch, E.W., Overall, C.M., Van Eyk, J.E., Baker, M.S., Paik, Y.K., Weintraub, S.T., Lane, L., Martens, L., Vandenbrouck, Y., Kusebauch, U., et al. (2016). Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. J Proteome Res 15:3961-3970.

Eng, J.K., and Deutsch, E.W. (2020). Extending Comet for Global Amino Acid Variant and Post-Translational Modification Analysis Using the PSI Extended FASTA Format. Proteomics 20:e1900362.

Farrah, T., Deutsch, E.W., Omenn, G.S., Campbell, D.S., Sun, Z., Bletz, J.A., Mallick, P., Katz, J.E., Malmstrom, J., Ossola, R., et al. (2011). A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. Mol Cell Proteomics 10:M110 006353.

Feng, Y., Jiang, M., Yu, W., and Zhou, J. (2023). Identification of short open reading frames in plant genomes. Front Plant Sci 14:1094715.

Friso, G., and van Wijk, K.J. (2015). Posttranslational Protein Modifications in Plant Metabolism. Plant Physiol 169:1469-1487.

Hawkins, C.L., and Davies, M.J. (2019). Detection, identification, and quantification of oxidative protein modifications. J Biol Chem 294:19683-19708.

Hazarika, R.R., De Coninck, B., Yamamoto, L.R., Martin, L.R., Cammue, B.P., and van Noort, V. (2017). ARA-PEPs: a repository of putative sORF-encoded peptides in Arabidopsis thaliana. BMC Bioinformatics 18:37.

Hoopes, G.M., Hamilton, J.P., Wood, J.C., Esteban, E., Pasha, A., Vaillancourt, B., Provart, N.J., and Buell, C.R. (2019). An updated gene atlas for maize reveals organ-specific and stress-induced genes. Plant J 97:1154-1167.

Hufford, M.B., Seetharam, A.S., Woodhouse, M.R., Chougule, K.M., Ou, S., Liu, J., Ricci, W.A., Guo, T., Olson, A., Qiu, Y., et al. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science 373:655-662.

Hulstaert, N., Shofstahl, J., Sachsenberg, T., Walzer, M., Barsnes, H., Martens, L., and Perez-Riverol, Y. (2020). ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. J Proteome Res 19:537-542.

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., Chin, C.S., et al. (2017). Improved maize reference genome with single-molecule technologies. Nature 546:524-527.

Keller, A., Eng, J., Zhang, N., Li, X.J., and Aebersold, R. (2005). A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol 1:2005 0017.

Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74:5383-5392.

Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D., and Nesvizhskii, A.I. (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods 14:513-520.

Li, L., Briskine, R., Schaefer, R., Schnable, P.S., Myers, C.L., Flagel, L.E., Springer, N.M., and Muehlbauer, G.J. (2016). Co-expression network analysis of duplicate genes in maize (Zea mays L.) reveals no subgenome bias. BMC Genomics 17:875.

Li, W., O'Neill, K.R., Haft, D.H., DiCuccio, M., Chetvernin, V., Badretdin, A., Coulouris, G., Chitsaz, F., Derbyshire, M.K., Durkin, A.S., et al. (2021). RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. Nucleic Acids Res 49:1020-1028.

Lin, G., He, C., Zheng, J., Koo, D.H., Le, H., Zheng, H., Tamang, T.M., Lin, J., Liu, Y., Zhao, M., et al. (2021). Chromosome-level genome assembly of a regenerable maize inbred line A188. Genome Biol 22:175.

Maier, R.M., Neckermann, K., Igloi, G.L., and Kossel, H. (1995a). Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. J Mol Biol 251:614-628.

Maier, U.G., Rensing, S.A., Igloi, G.L., and Maerz, M. (1995b). Twintrons are not unique to the Euglena chloroplast genome: structure and evolution of a plastome cpn60 gene from a cryptomonad. Mol Gen Genet 246:128-131.

Majeran, M., Friso, G., Ponnala, L., Connolly, B., Huang, M., Reidel, E., Zhang, C., Asakura, Y., Bhuiyan, N.H., Sun, Q., et al. (2010). Structural and metabolic transitions of C4 leaf development and differentiation defined by microscopy and quantitative proteomics. The Plant Cell 22:3509-3542.

Majeran, W., Friso, G., Asakura, Y., Qu, X., Huang, M., Ponnala, L., Watkins, K.P., Barkan, A., and van Wijk, K.J. (2012). Nucleoid-enriched proteomes in developing plastids and chloroplasts from maize leaves: a new conceptual framework for nucleoid functions. Plant Physiol 158:156-189.

Majeran, W., Zybailov, B., Ytterberg, A.J., Dunsmore, J., Sun, Q., and van Wijk, K.J. (2008). Consequences of C4 differentiation for chloroplast membrane proteomes in maize mesophyll and bundle sheath cells. Mol Cell Proteomics 7:1609-1638.

Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W.H., Rompp, A., Neumann, S., Pizarro, A.D., et al. (2011). mzML--a community standard for mass spectrometry data. Mol Cell Proteomics 10:R110 000133.

Mergner, J., Frejno, M., List, M., Papacek, M., Chen, X., Chaudhary, A., Samaras, P., Richter, S., Shikata, H., Messerer, M., et al. (2020). Mass-spectrometry-based draft of the Arabidopsis proteome. Nature 579:409-414.

Millar, A.H., Heazlewood, J.L., Giglione, C., Holdsworth, M.J., Bachmair, A., and Schulze, W.X. (2019). The Scope, Functions, and Dynamics of Posttranslational Protein Modifications. Annu Rev Plant Biol prepublication online:119-151.

Nie, S., Wang, B., Ding, H., Lin, H., Zhang, L., Li, Q., Wang, Y., Zhang, B., Liang, A., Zheng, Q., et al. (2021). Genome assembly of the Chinese maize elite inbred line RP125 and its EMS mutant collection provide new resources for maize genetics research and crop improvement. Plant J 108:40-54.

Noutsos, C., Kleine, T., Armbruster, U., DalCorso, G., and Leister, D. (2007). Nuclear insertions of organellar DNA can create novel patches of functional exon sequences. Trends Genet 23:597-601.

Portwood, J.L., 2nd, Woodhouse, M.R., Cannon, E.K., Gardiner, J.M., Harper, L.C., Schaeffer, M.L., Walsh, J.R., Sen, T.Z., Cho, K.T., Schott, D.A., et al. (2019). MaizeGDB 2018: the maize multi-genome genetics and genomics database. Nucleic Acids Res 47:D1146-D1154.

Ren, Z., Qi, D., Pugh, N., Li, K., Wen, B., Zhou, R., Xu, S., Liu, S., and Jones, A.R. (2019). Improvements to the Rice Genome Annotation Through Large-Scale Analysis of RNA-Seq and Proteomics Data Sets. Mol Cell Proteomics 18:86-98.

Rosenberger, G., Bludau, I., Schmitt, U., Heusel, M., Hunter, C.L., Liu, Y., MacCoss, M.J., MacLean, B.X., Nesvizhskii, A.I., Pedrioli, P.G.A., et al. (2017). Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. Nat Methods 14:921-927.

Sabari, B.R., Zhang, D., Allis, C.D., and Zhao, Y. (2017). Metabolic regulation of gene expression through histone acylations. Nat Rev Mol Cell Biol 18:90-101.

Schnable, J.C., Springer, N.M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci U S A 108:4069-4074.

Schnable, P.S., and Ware, D., and Fulton, R.S., and Stein, J.C., and Wei, F., and Pasternak, S., and Liang, C., and Zhang, J., and Fulton, L., and Graves, T.A., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112-1115.

Shamimuzzaman, M., Gardiner, J.M., Walsh, A.T., Triant, D.A., Le Tourneau, J.J., Tayal, A., Unni, D.R., Nguyen, H.N., Portwood, J.L., 2nd, Cannon, E.K.S., et al. (2020). MaizeMine: A Data Mining Warehouse for the Maize Genetics and Genomics Database. Front Plant Sci 11:592730.

Shteynberg, D., Deutsch, E.W., Lam, H., Eng, J.K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R.L., Aebersold, R., and Nesvizhskii, A.I. (2011). iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics 10:M111 007690.

Song, Y.C., Das, D., Zhang, Y., Chen, M.X., Fernie, A.R., Zhu, F.Y., and Han, J. (2023). Proteogenomics-based functional genome research: approaches, applications, and perspectives in plants. Trends Biotechnol.

Springer, N.M., Anderson, S.N., Andorf, C.M., Ahern, K.R., Bai, F., Barad, O., Barbazuk, W.B., Bass, H.W., Baruch, K., Ben-Zvi, G., et al. (2018). The maize W22 genome provides a foundation for functional genomics and transposon biology. Nat Genet 50:1282-1288.

Strable, J., and Scanlon, M.J. (2009). Maize (Zea mays): a model organism for basic and applied research in plant biology. Cold Spring Harb Protoc 2009:pdb emo132.

Tress, M.L., Abascal, F., and Valencia, A. (2017). Alternative Splicing May Not Be the Key to Proteome Complexity. Trends Biochem Sci 42:98-110.

UniProt, C. (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49:D480-D489.

van Wijk, K.J., Bentolila, S., Leppert, T., Sun, Q., Sun, Z., Mendoza, L., Li, M., Deutsch, E.W. (2023). Detection and editing of the updated plastid- and mitochondrial-encoded proteomes for Arabidopsis with PeptideAtlas. Plant Physiology accepted

van Wijk, K.J., Friso, G., Walther, D., and Schulze, W.X. (2014). Meta-Analysis of Arabidopsis thaliana Phospho-Proteomics Data Reveals Compartmentalization of Phosphorylation Motifs. Plant Cell 26:2367-2389.

van Wijk, K.J., Leppert, T., Sun, Q., Boguraev, S.S., Sun, Z., Mendoza, L., and Deutsch, E.W. (2021). The Arabidopsis PeptideAtlas: Harnessing worldwide proteomics data to create a comprehensive community proteomics resource. Plant Cell 33:3421-3453.

van Wijk, K.J., Leppert, T., Sun, Z., Kearly, A., Li, M., Mendoza, L., Guzchenko, I., Debley, E., Sauermann, G., Routray, P., et al. (2023). Mapping the Arabidopsis thaliana proteome in

PeptideAtlas and the nature of the unobserved (dark) proteome; strategies towards a complete proteome. biorxiv.

Vere, G., Kealy, R., Kessler, B.M., and Pinto-Fernandez, A. (2020). Ubiquitomics: An Overview and Future. Biomolecules 10.

Verrastro, I., Pasha, S., Jensen, K.T., Pitt, A.R., and Spickett, C.M. (2015). Mass spectrometry-based methods for identifying oxidized proteins in disease: advances and challenges. Biomolecules 5:378-411.

Vitorino, R., Guedes, S., Trindade, F., Correia, I., Moura, G., Carvalho, P., Santos, M.A.S., and Amado, F. (2020). De novo sequencing of proteins by mass spectrometry. Expert Rev Proteomics 17:595-607.

Walley, J.W., Sartor, R.C., Shen, Z., Schmitz, R.J., Wu, K.J., Urich, M.A., Nery, J.R., Smith, L.G., Schnable, J.C., Ecker, J.R., et al. (2016). Integration of omic networks in a developmental atlas of maize. Science 353:814-818.

Wang, X., Codreanu, S.G., Wen, B., Li, K., Chambers, M.C., Liebler, D.C., and Zhang, B. (2018). Detection of Proteome Diversity Resulted from Alternative Splicing is Limited by Trypsin Cleavage Specificity. Mol Cell Proteomics 17:422-430.

Wang, Y.F., Chao, Q., Li, Z., Lu, T.C., Zheng, H.Y., Zhao, C.F., Shen, Z., Li, X.H., and Wang, B.C. (2019). Large-scale Identification and Time-course Quantification of Ubiquitylation Events During Maize Seedling De-etiolation. Genomics Proteomics Bioinformatics 17:603-622.

Wei, F., Zhang, J., Zhou, S., He, R., Schaeffer, M., Collura, K., Kudrna, D., Faga, B.P., Wissotski, M., Golser, W., et al. (2009). The physical and genetic framework of the maize B73 genome. PLoS Genet 5:e1000715.

Willems, P., Horne, A., Van Parys, T., Goormachtig, S., De Smet, I., Botzki, A., Van Breusegem, F., and Gevaert, K. (2019). The Plant PTM Viewer, a central resource for exploring plant protein modifications. Plant J 99:752-762.

Xanthopoulou, A., Moysiadis, T., Bazakos, C., Karagiannis, E., Karamichali, I., Stamatakis, G., Samiotaki, M., Manioudaki, M., Michailidis, M., Madesis, P., et al. (2021). The perennial fruit tree proteogenomics atlas: a spatial map of the sweet cherry proteome and transcriptome. Plant J.

Xu, Y., Shi, Z., and Bao, L. (2022). An Expanding Repertoire of Protein Acylations. Mol Cell Proteomics 21:100193.

Yan, Z., Shen, Z., Gao, Z.F., Chao, Q., Qian, C.R., Zheng, H., and Wang, B.C. (2020). A comprehensive analysis of the lysine acetylome reveals diverse functions of acetylated proteins during de-etiolation in Zea mays. J Plant Physiol 248:153158.

Yu, F., Teo, G.C., Kong, A.T., Frohlich, K., Li, G.X., Demichev, V., and Nesvizhskii, A.I. (2023). Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. Nat Commun 14:4154.

Zhang, H., Liu, P., Guo, T., Zhao, H., Bensaddek, D., Aebersold, R., and Xiong, L. (2019). Arabidopsis proteome and the mass spectral assay library. Sci Data 6:278.

Zhu, F.Y., Chen, M.X., Ye, N.H., Shi, L., Ma, K.L., Yang, J.F., Cao, Y.Y., Zhang, Y., Yoshida, T., Fernie, A.R., et al. (2017). Proteogenomic analysis reveals alternative splicing and translation as part of the abscisic acid response in Arabidopsis seedlings. Plant J 91:518-533.

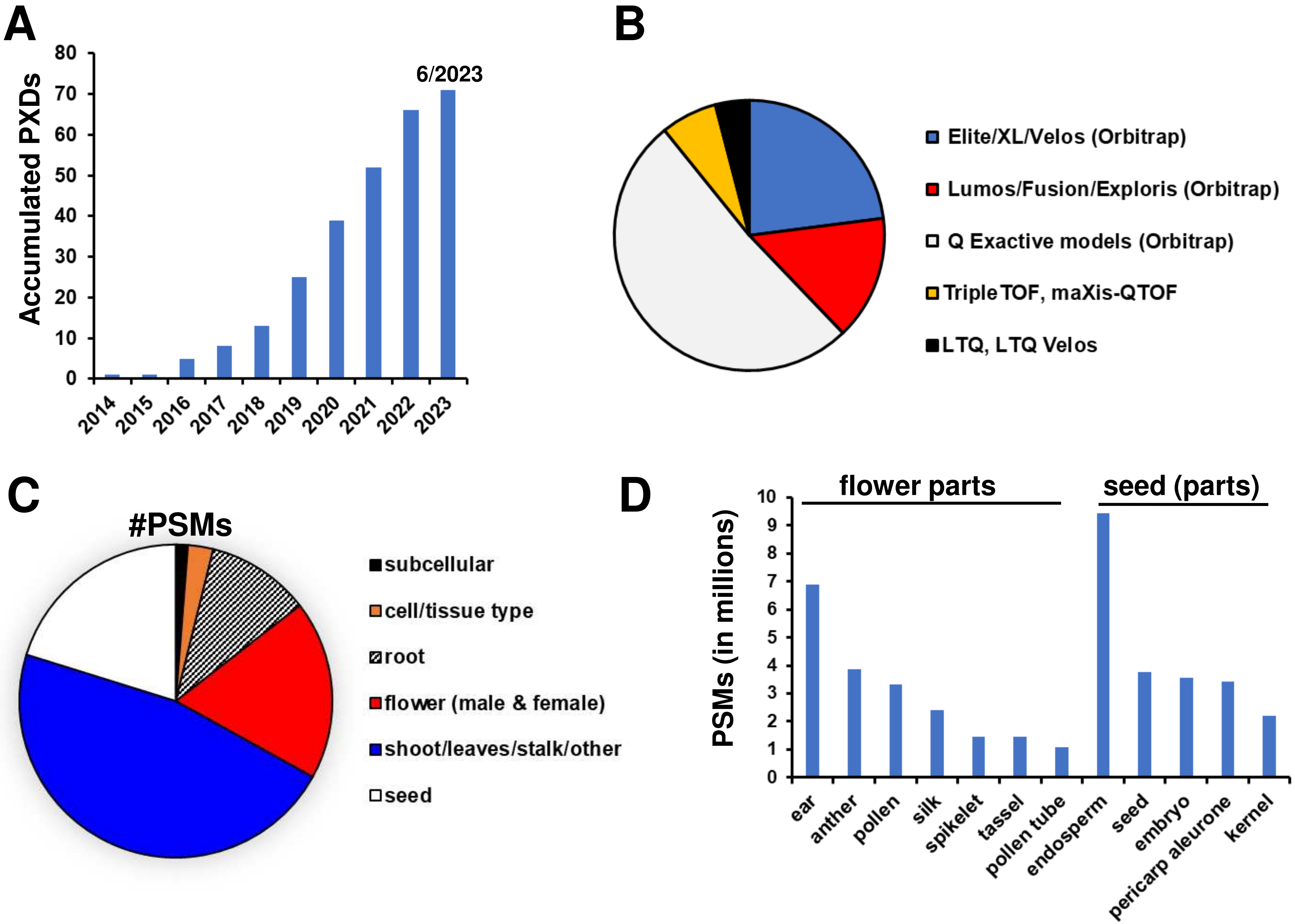


Figure 1.

Figure 1. Features of maize PDXs used for the first maize PeptideAtlas build.
(**A**) Cumulative PXD for maize in ProteomeXchange passing our selection criteria across the different years for maize.
(**B**) Mass spectrometry instruments used to acquire data in the maize PXDs used in the maize PeptideAtlas grouped in different categories by vendor and model.
(**C**) Distribution of PSMs of samples from different plant parts in the maize PeptideAtlas
(D) Distribution of PSMs of samples for specific flower parts and seeds.

Figure 2

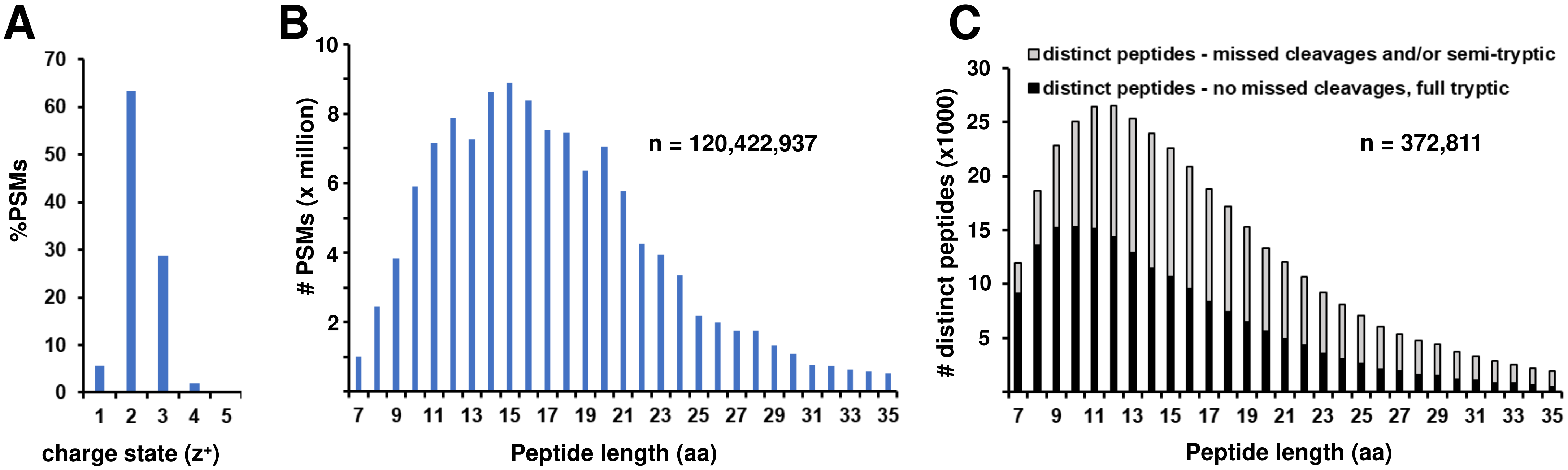


Figure 2. Key statistics of matched MS/MS data (PSMs) for this maize PeptideAtlas build.
(**A**) Frequency distribution of PSMs for peptide charge state (z).
(**B**) Frequency distribution of PSMs for peptide length (aa). Note that when R or K is followed by P, trypsin does not cleave and hence these are not counted towards missed cleavages.
(**C**) Number of distinct peptides as a function of peptide length (aa).

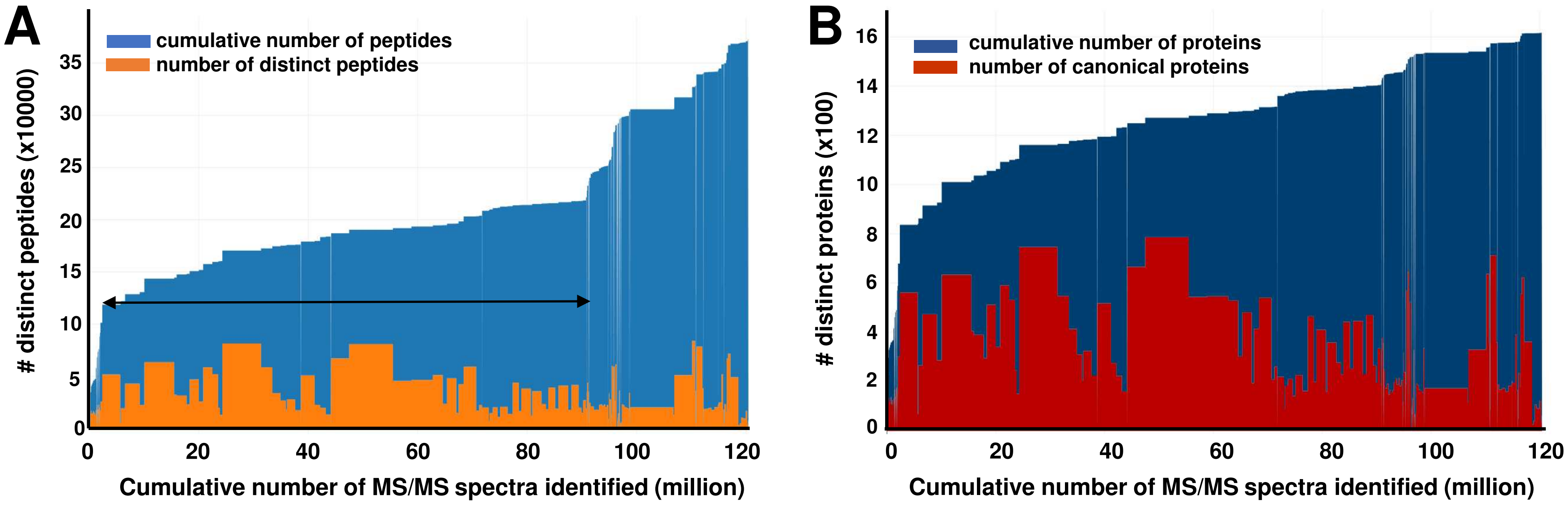
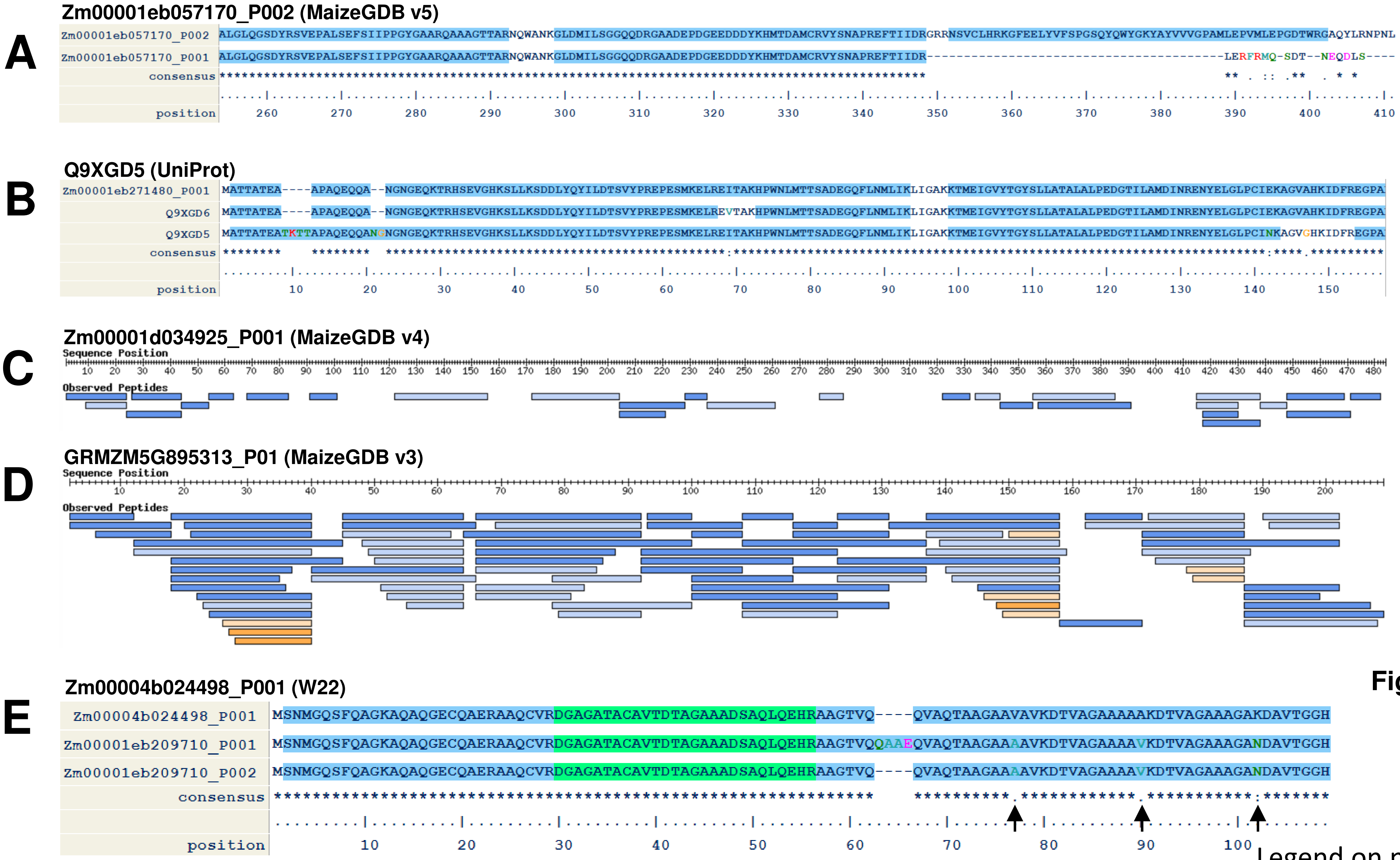


Figure 3

Figure 3. Number of distinct (non-redundant) peptides (left panel) and identified canonical proteins (right panel) as a function of the cumulative number of PSMs (peptide-spectrum matches) for the first maize PeptideAtlas. The cumulative count is ordered by PXD identifier (from low to high or old to new). The build is based on 211 experiments across the 74 selected PXDs, where each PXD may be decomposed into several experiments/samples when such information can be determined). The PSM FDR is 8.10^{-5}.

(A) Number of distinct (non-redundant) peptides as function of the cumulative number of MS/MS spectra matched. 372,811 distinct peptides are identified at a peptide-level FDR of 0.15%. Blue rectangles represent the cumulative number of distinct peptides as experiments are added to the build, whereas orange rectangles represent the number of distinct peptides in each experiment. The arrow from 2 million to 90.7 million MS/MS spectra are from PXD002853 acquired on a lower resolution LTQ-Velos instrument.

(B) Number of distinct (non-redundant) canonical proteins as function of the cumulative number of MS/MS spectra matched at a canonical protein-level FDR of <0.006%. Blue rectangles represent the cumulative number of canonical proteins as experiments are added to the build, whereas red rectangles represent the number of canonical proteins in each experiment.



```
Zm00001eb057170_P002 ALGLQGSDYRSVEPALSEFSIIPPGYGAARQAAAGTTARNQWANKGLDMILSGGQQDRGAADEPDGEEDDDYKHMTDAMCRVYSNAPREFTIIDRGRRNSVCLHRKGFEELYVFSPGSQYQWYGKYAYVVVGPAMLEPVMLEPGDTWRGAQYLRNPNL
Zm00001eb057170_P001 ALGLQGSDYRSVEPALSEFSIIPPGYGAARQAAAGTTARNQWANKGLDMILSGGQQDRGAADEPDGEEDDDYKHMTDAMCRVYSNAPREFTIIDR----------------------------------------LERFRMQ-SDT--NEQDLS----
           consensus ***********************************************************************************************                                        ** . :: .**  . * *
                     ......|.........|.........|.........|.........|.........|.........|.........|.........|.........|.........|.........|.........|.........|.........|.........|.
            position       260       270       280       290       300       310       320       330       340       350       360       370       380       390       400       410
```

```
Zm00001eb271480_P001 MATTATEA----APAQEQQA--NGNGEQKTRHSEVGHKSLLKSDDLYQYILDTSVYPREPESMKELREITAKHPWNLMTTSADEGQFLNMLIKLIGAKKTMEIGVYTGYSLLATALALPEDGTILAMDINRENYELGLPCIEKAGVAHKIDFREGPA
              Q9XGD6 MATTATEA----APAQEQQA--NGNGEQKTRHSEVGHKSLLKSDDLYQYILDTSVYPREPESMKELREVTAKHPWNLMTTSADEGQFLNMLIKLIGAKKTMEIGVYTGYSLLATALALPEDGTILAMDINRENYELGLPCIEKAGVAHKIDFREGPA
              Q9XGD5 MATTATEATKTTAPAQEQQANGNGNGEQKTRHSEVGHKSLLKSDDLYQYILDTSVYPREPESMKELREITAKHPWNLMTTSADEGQFLNMLIKLIGAKKTMEIGVYTGYSLLATALALPEDGTILAMDINRENYELGLPCINKAGVGHKIDFREGPA
           consensus ********    ********  **********************************************:************************************************************************:****.**********
                     .........|.........|.........|.........|.........|.........|.........|.........|.........|.........|.........|.........|.........|.........|.........|.......
            position          10        20        30        40        50        60        70        80        90       100       110       120       130       140       150
```

```
Zm00004b024498_P001 MSNMGQSFQAGKAQAQGECQAERAAQCVRDGAGATACAVTDTAGAAADSAQLQEHRAAGTVQ----QVAQTAAGAAVAVKDTVAGAAAAAKDTVAGAAAGAKDAVTGGH
Zm00001eb209710_P001 MSNMGQSFQAGKAQAQGECQAERAAQCVRDGAGATACAVTDTAGAAADSAQLQEHRAAGTVQQAAEQVAQTAAGAAAAVKDTVAGAAAAVKDTVAGAAAGANDAVTGGH
Zm00001eb209710_P002 MSNMGQSFQAGKAQAQGECQAERAAQCVRDGAGATACAVTDTAGAAADSAQLQEHRAAGTVQ----QVAQTAAGAAAAVKDTVAGAAAAVKDTVAGAAAGANDAVTGGH
           consensus **************************************************************    **********.************.***********:*******
                     .........|.........|.........|.........|.........|.........|.........|.........|.........|.........|.......
            position          10        20        30        40        50        60        70        80        90       100
```

Figure 4.

Legend on next page

Figure 4. Examples of proteins that are well detected in the Maize PeptideAtlas that are not part of the MaizeGDB v5 core proteome.

(A) Sequence comparison of Zm00001eb057170_P001 (NDH-dependent cyclic electron flow 5) and its second isoform Zm00001eb057170_P002 (both from MaizeGDB v5). The two isoforms are the same except for the C-terminal region, where they use different exons. There is no support in PeptideAtlas for the P001 isoform, but the P002 isoform is well supported by several peptides. **(B)** Sequence comparison of UniProt Q9XGD6 (Caffeoyl-CoA O-methyltransferase 1), UniProt Q9XGD6 (Caffeoyl-CoA O-methyltransferase 2) and Zm00001eb271480_P001, the closest entry in MaizeGDB v5. Although Q9XGD6 and Zm00001eb271480_P001 are highly similar (except for a V – I difference at position 69), there is no analog for Q9XGD6 in MaizeGDB v5.

(C) Example of a protein (Zm00001d034925_P001) that is present in the MaizeGDB v4 proteome annotation but is absent from MaizeGDB v5. It is very well detected with 70% sequences coverage with 28 distinct uniquely mapping peptides, shown as blue rectangles. Darker shades indicate a larger number of PSMs.

(D) Example of a protein (GRMZM5G895313_P01) that is present in the MaizeGDB v3 proteome annotation but is absent from MaizeGDB v4 and v5. It is very well detected with 100% sequence coverage (except for the initiating methionine) with 75 distinct uniquely mapping peptides, shown in blue (and a few short low-complexity peptides such as GGGGGGGR that also map to other proteins, shown in orange).

(E) Example of a protein (Zm00004b024498_P001) in the MaizeGDB W22 proteome annotation that has many uniquely mapping peptides. It is very well detected with 8014 PSMs and 100% sequence coverage (except for the initiating methionine) with 36 distinct peptides, of which 24 are uniquely mapping to Zm00004b024498_P001. The skipped exon near position 65 is also encoded by a second isoform Zm00001eb209710_P002 from the B73 proteome. However, the single amino acid variants at positions 77, 90, and 102 are unique to the W22 sequence.

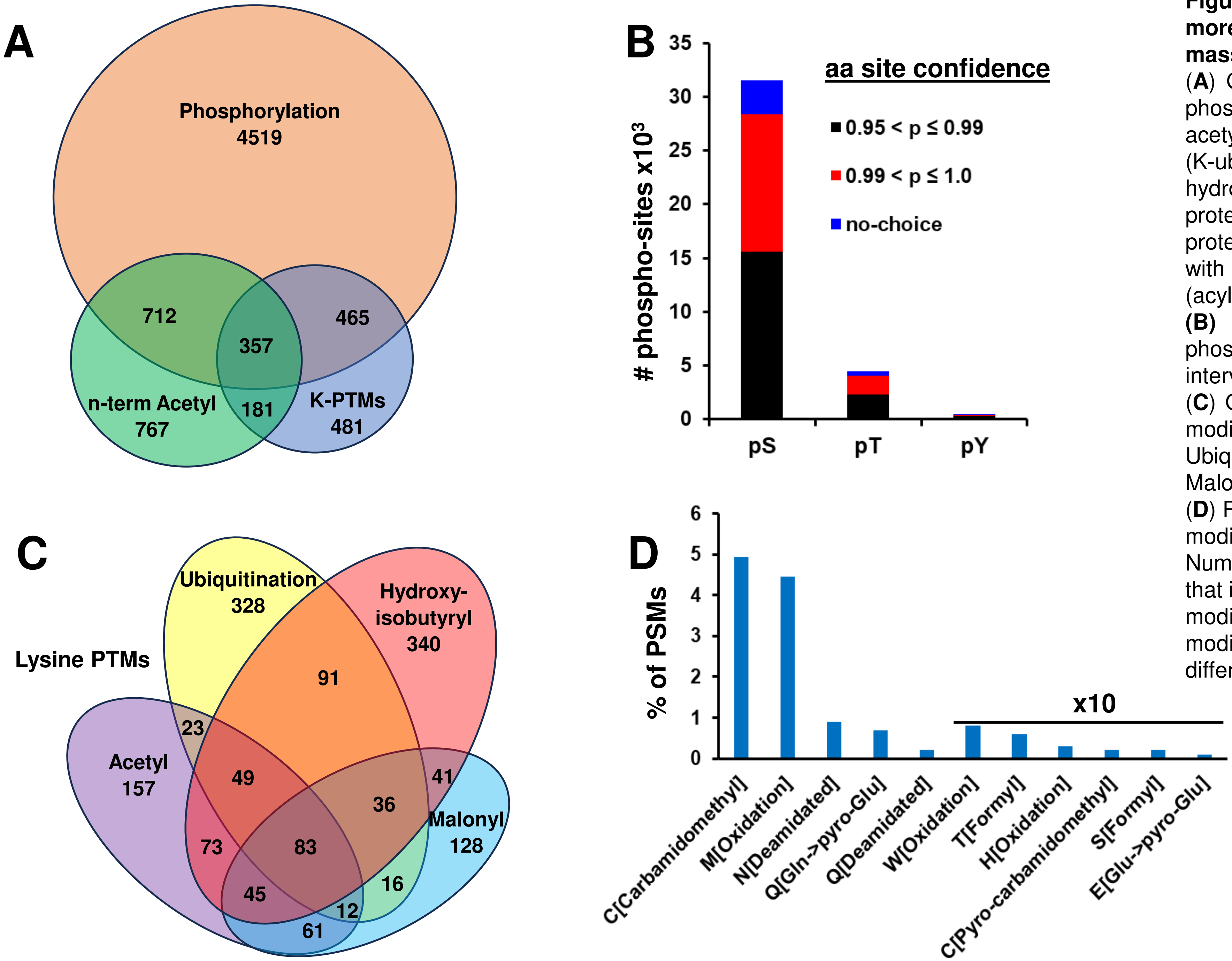


Figure 5. Canonical maize core proteins with one or more PTMs and frequency of PSMs with other mass modifications.

(**A**) Overlap between proteins that carry one or more phosphorylations (S, T or Y), N-terminal protein acetylation (NTA), or one or more lysine modifications (K-ubiquitination, K-acetylation, K-malonylation, K-hydroxybutyrylation). There in total there are 6053 proteins with one or more phosphorylations, 2017 proteins with N-terminal acetylation and 1484 proteins with one or more lysine side chain modifications (acylation or ubiquitination).

(B) Number of serine, threonine and tryrosine phosphorylation sites (P-sites) at different p-value intervals for the core proteome.

(**C**) Overlap between proteins with one or more lysine modifications. Colors: Acetylation - purple, Ubiquitination - yellow, Hydroxyisobutyrylation - red, Malonylation – blue.

(**D**) Percentage of PSMs (of total) with mass modifications mostly due to sample preparations. Numbers are computed as the total number of PSMs that include at least one instance of the listed mass modification. Some PSMs contain more than one mass modification of the same type (not multiple counted) or different type (multiple counted).

Figure 5.