Journal of proteome research

pubs.acs.org/jpr Article

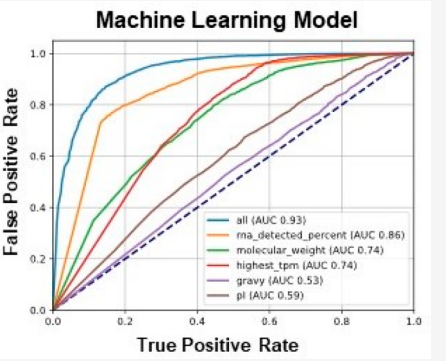
Detection of the *Arabidopsis* Proteome and Its Post-translational Modifications and the Nature of the Unobserved (Dark) Proteome in PeptideAtlas

Klaas J. van Wijk,* Tami Leppert, Zhi Sun, Alyssa Kearly, Margaret Li, Luis Mendoza, Isabell Guzchenko, Erica Debley, Georgia Sauermann, Pratyush Routray, Sagunya Malhotra, Andrew Nelson, Qi Sun, and Eric W. Deutsch*

Cite This: *J. Proteome Res.* 2024, 23, 185−214 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: This study describes a new release of the *Arabidopsis thaliana* PeptideAtlas proteomics resource (build 2023−10) providing protein sequence coverage, matched mass spectrometry (MS) spectra, selected post-translational modifications (PTMs), and metadata. 70 million MS/MS spectra were matched to the Araport11 annotation, identifying ∼0.6 million unique peptides and 18,267 proteins at the highest confidence level and 3396 lower confidence proteins, together representing 78.6% of the predicted proteome. Additional identified proteins not predicted in Araport11 should be considered for the next *Arabidopsis* genome annotation. This release identified 5198 phosphorylated proteins, 668 ubiquitinated proteins, 3050 N-terminally acetylated proteins, and 864 lysine-acetylated proteins and mapped their PTM sites. MS support was lacking for 21.4% (5896 proteins) of the predicted Araport11 proteome: the "dark" proteome. This dark proteome is highly enriched for E3 ligases, transcription factors, and for certain (e.g., CLE, IDA, PSY) but not other (e.g., THIONIN, CAP) signaling peptides families. A machine learning model trained on RNA expression data and protein properties predicts the probability that proteins will be detected. The model aids in discovery of proteins with short half-life (e.g., SIG1,3 and ERF-VII TFs) and for developing strategies to identify the missing proteins. PeptideAtlas is linked to TAIR, tracks in JBrowse, and several other community proteomics resources.



KEYWORDS: *Arabidopsis, PeptideAtlas, post-translational modifications, ProteomeXchange, machine learning, signaling peptides, E3 ligases*

INTRODUCTION

Arabidopsis thaliana (*Arabidopsis*) was established as a universal plant model system in the 1980s as a means of advancing the plant science field.[1,2] The power of *Arabidopsis* as an experimental model system to discover novel gene functions and molecular pathways was first demonstrated using loss-of function mutants in the photorespiratory pathway.[3,4] Since then, the field of plant biology, and specifically plant molecular biology and genetics, has expanded enormously, and produced a wealth of knowledge and understanding of plants.[5,6] A well-organized *Arabidopsis* community with powerful public resources is facilitating and accelerating new discoveries in Plant Biology.[6,7]

Arabidopsis also has been established as a model for analysis of its proteome in particular because mass spectrometry (MS)-based proteomics immensely benefits from having a well-annotated genome with a robust set of predicted proteins.[8] A poorly annotated genome and poorly predicted proteins diminish the ability to carry out quantitative proteome analyses and determine the rich complexity of post-translational modifications (PTMs), including the assignment of PTMs to specific amino acid residues. A range of plant proteome databases by individual laboratories have been developed, mostly for *Arabidopsis* proteins, often focused on a particular aspect of plant proteomics, such as subcellular compartments,[9,10] protein location (SUBA and PPDB),[11,12] or PTMs.[13,14] A comprehensive *Arabidopsis* proteome database (ATHENA) was released to allow mining of a large-scale experimental proteome data set involving multiple tissue types as published in (Mergner et al., 2020). In 2021, we launched the first release of the *Arabidopsis* PeptideAtlas to address central questions about the *Arabidopsis* proteome, such as experimental evidence for accumulation of proteins, their

Received: August 24, 2023
Revised: October 27, 2023
Accepted: November 5, 2023
Published: November 21, 2023

approximate relative abundance, the significance of protein splice forms, and selected PTMs[8] (https://peptideatlas.org/builds/arabidopsis/). Species-specific PeptideAtlas resources have also been developed for nonplant species including human,[15] various animals such pigs,[16] chicken,[17] fish,[18] different yeast species,[19,20] and bacteria.[21−23] Each PeptideAtlas is based on published MSMS data sets collected through the ProteomeXchange Consortium[24] and reanalyzed through a uniform processing pipeline. In the case of the *Arabidopsis* PeptideAtlas, we also annotate the metadata associated with the raw MS data and link all peptide identifications to spectral, technical, and biological metadata. To that end, we developed a metadata annotation system within PeptideAtlas. These metadata are critical to determine cell-type or subcellular specific protein accumulation patterns and help accomplish the long-term goal of the *Arabidopsis* community to develop a detailed *Arabidopsis* Plant Cell Atlas.[25]

The first PeptideAtlas release was based on 52 PXDs containing 6148 raw files with 142.7 million MSMS spectra. The current study describes the second release, with an additional 63 PXDs containing 4330 raw files with 116.7 million MSMS spectra. The new *Arabidopsis* PeptideAtlas release is by far the largest resource for MS-based identified *Arabidopsis* proteins; other far smaller resources for identification of *Arabidopsis* protein by MS are ATHENA based on 57 million searched MSMS spectra (https://athena.proteomics.wzw.tum.de/)[26] and less than 10 million for the Plant Proteome Data Base (http://ppdb.tc.cornell.edu/).[11] The objectives for this second release are to map peptides to proteins that were not identified in the first release, extend sequence coverage of already identified proteins, and increase the number of identified proteins at the highest level of confidence tier (canonical proteins). In addition, the aim of this second release is to provide deeper coverage for protein phosphorylation, N-terminal and lysine acetylation, and now also includes PXDs that included specific enrichment workflows for ubiquitinated proteins.[27,28] To try to increase the detection or sequence coverage of proteins, we employed four criteria for the selection of new PXDs: (i) PXDs of specific cell types or specialized subcellular fractions, (ii) PXDs that concern specific protein complexes or protein affinity enrichments, (iii) PXDs that are enriched for specific post-translational modifications, and (iv) PXDs that appear to have very high dynamic resolution and sensitivity by using the latest technologies in mass spectrometry and/or sample fractionation. The new PeptideAtlas release now maps peptides to 78.6% of the predicted *Arabidopsis* proteome, with each mapped peptide connected to the metadata and spectrum matches. With the ultimate goal of identifying the complete *Arabidopsis* predicted proteome, we investigated why 21% of the predicted *Arabidopsis* proteome was not yet observed in this new build. We named this unobserved proteome the "dark" proteome. This term has also been coined for protein sequences for where there is no experimental structure[29] and for putative proteins generated by noncanonical open reading frames.[30] The term "dark matter" in proteomics was coined for those MSMS spectra that could not be assigned to any protein (in most cases due to chemical-induced mass modifications or PTMs).[31] In our current study, we will use the term "dark proteome" for the set of predicted proteins for which this new PeptideAtlas build found no matching MSMS spectra. A significant portion of these unobserved proteins have physicochemical properties that likely impede detection by MS (e.g., very small and very hydrophobic). Other unobserved proteins likely accumulate under highly specific conditions or cell-types and/or have low cellular abundance due to short protein lifetime. Here we used large scale RNA-seq data sets for *Arabidopsis* to determine mRNA expression patterns for these unobserved proteins sampling across many tissue- and cell types, developmental stages, as well as biotic and abiotic stress conditions. We developed machine learning models based on these mRNA expression features and physicochemical protein properties to calculate the probability for each protein to be detected. GO enrichment analysis showed overrepresentation of specific functions in the dark proteome, e.g., E3 ligases and signaling peptides. The machine learning model outputs will help design optimal and targeted experimental strategies to detect these unobserved proteins. Finally, this second PeptideAtlas release, including its associated metadata and our machine learning output, provides an ideal platform to contribute to a community *Arabidopsis* proteome cell atlas[25,32] and also contribute to the ongoing reannotation of the *Arabidopsis* genome. The new PeptideAtlas release is integrated into TAIR (https://www.arabidopsis.org/) and linked to JBrowse (https://jbrowse.arabidopsis.org), PPDB,[11] SUBA,[33] UniProtKB,[34] ATHENA,[26] PhosPhat,[35] and Plant PTM Viewer.[36]

METHODS

Selection and Downloads of ProteomeXchange Submissions

Raw files for the selected PXDs for this new build 2023−10 were downloaded from ProteomeXchange (http://www.proteomexchange.org/) repositories. Supplemental Table Data Set 1 provides detailed information about the 63 newly selected PXDs, as well as the 52 PXDs that were part of the first build; this includes information about instrument, sample (e.g., subcellular proteome, plant organ), protease(s) used for sample digestion, number of raw files and MSMS spectra (searched and matched), identified proteins and peptides, submitting lab and associated publication, as well as several informative keywords.

Extraction and Annotation of Metadata

For each selected data set, we obtained information associated with the submission, and the publication if available, to determine search parameters and provide meaningful tags that describe the samples in some detail. These tags are visible for the relevant proteins at the PeptideAtlas interface. If needed, we contacted the authors for more information about the raw files. All collected metadata are stored in our annotation system as previously described.[8] These metadata can be viewed for each identified protein in PeptideAtlas.

Assembly of Protein Search Space

We assembled a comprehensive protein search space comprising the predicted *Arabidopsis* protein sequences from (i) Araport11,[37] (ii) TAIR10,[38] (iii) UniProtKB,[39] (iv) RefSeq (https://www.ncbi.nlm.nih.gov/refseq),[40] (v) from the repository ARA-PEPs[41] with 7901 small Open Reading Frames (sORFs), 16,809 low molecular weight peptides and proteins (LWs; between 26 and 250 aa; median 37 aa), as well as 607 novel stress-induced peptides (SIPs) most of which are currently not annotated in TAIR10 or Araport11, vi) from Dr Eve Wurtele (Iowa State University) assembled based on RNA-seq data, vii) GFP, RFP and YFP protein sequences

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

commonly used as reporters and affinity enrichments, viii) 116 contaminant protein sequences frequently observed in proteome samples (e.g., keratins, trypsin, BSA) (https://www.thegpm.org/crap/). This search space is quite similar to that for the first PeptideAtlas release, except that the UniProtKB and RefSeq contributions were updated to the latest version as of 2021−04. Also added was the complete set of predicted protein sequences for the 950 Araport11 pseudogenes (1240 gene models) that we generated through 3-frame translation (the pseudogene sequences have transcription direction, but no frame). The amino acid sequences for this complete search space can be down-loaded at the PeptideAtlas Web site.

We also included an update on the plastid- and mitochondrial-encoded proteins to address redundancies in plastid- and mitochondrial ATGC and ATMG identifiers and inclusion of protein sequences for those plastid- and mitochondrial encoded proteins that are predicted to be affected by RNA editing. For the mitochondrial-encoded proteins, we included 420 editing sites in 29 mitochondrial-encoded proteins and two ORFs, most of which are described in ref 42, whereas we included edited sequences for 17 plastid-encoded proteins that included 31 amino acid changes and generation of one start methionine. These organellar-encoded sequences included unedited sequences and completely edited sequences, and if editing sites were sufficiently close together to appear in a single peptide, we also include all permutations of edits and nonedits. This resulted in the addition of 10,368 sequences for plastid- and mitochondrial encoded variants to the search database. In a separate study,[43] we provide details on the annotation and redundancy of plastid- and mitochondrial encoded proteins, the expression of organellar ORFs, and the impact of RNA editing.

The Trans-Proteomic Pipeline (TPP) Data Processing Pipeline

For all selected data sets, the vendor-format raw files were downloaded from the hosting ProteomeXchange repository, converted to mzML files[44] using ThermoRawFileParser[45] for Thermo Fisher Scientific instruments or the msconvert tool from the ProteoWizard toolkit[46] for SCIEX wiff files, and then analyzed with the TPP[47,48] version 6.2.0.[49] The TPP analysis consisted of sequence database searching with either Comet[50] for LTQ-based fragmentation spectra or MSFragger 3.2[51] for higher resolution fragmentation spectra and postsearch validation with several additional TPP tools as follows: PeptideProphet[52] was run to assign probabilities of being correct for each peptide-spectrum match (PSM) using semiparametric 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.[53]

For data sets 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 semitryptic specificity. The semitryptic searches were carried out with the following possible variable modifications (5 max per peptide for Comet and 3 for MSFragger): oxidation of Met, Trp, Pro (+15.9949 Da), acetylation of Lys (+42.0106 Da), peptide N-terminal Gln to pyro-Glu (−17.0265 Da), peptide N-terminal Glu to pyro-Glu (−18.0106 Da), peptide N-terminal carbamidomethyl-Cys to Pyro-carbamidomethyl-Cys (−17.02650 Da), deamidation of Asn or Gln (+0.9840 Da), protein N-terminal acetylation (+42.0106 Da), and if peptides were specifically affinity enriched for phosphopeptides, also phosphorylation of Ser, Thr or Tyr (+79.9663 Da). For the fully tryptic searches, we also added oxidation of His (+15.9949 Da) and formylation of peptide N-termini, Ser, or Thr (+27.9949 Da); we deliberately restricted these latter PTMs to only full tryptic (rather than also allowing semitryptic) 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.[54,55] In both semitryptic and full tryptic searches, fixed modifications for carbamidomethylation of Cys (+57.0215 Da) if treated with reductant and iodoacetamide (or other alkylating reagents) and isobaric tag modifications (TMT, iTRAQ) were applied as appropriate. Both variable and fixed modifications were applied to dimethyl labeled data sets as appropriate. Four missed cleavages were allowed (RP or KP does not count as a missed cleavage). The results of the fully tryptic and the semitryptic searches were combined using iProphet, which computes revised probabilities for the combined result based on the PeptideProphet probabilities from the individual searches as described in.[53] Several PXDs were generated using other proteases (GluC, ArgC, and Chymotrypsin); these data sets were processed similarly to those generated by trypsin (i.e., same minimal length of matched peptides (7 aa), same number of missed cleavages with relevant exceptions for each protease as defined in the TPP pipeline, and same fixed and variable modifications), with the exception that the relevant enzyme was chosen. Some of the data sets derived from analysis of extracted peptidomes in which "no protease treatment" was used and these data sets were searched with "no enzyme".

PeptideAtlas Assembly

In order to create the combined PeptideAtlas build of all experiments, all data sets were thresholded at an iProphet probability that yields the model-based PSM FDR of 0.0008. The exact probability varied from experiment to experiment depending on how well the modeling can separate correct from incorrect. This probability threshold was typically greater than 0.99. As more and more experiments are combined, the total FDR increases unless the threshold is made more stringent.[56] Throughout the procedure, decoy identifications were retained and then used to compute the final decoy-based FDRs. The decoy count-based PSM-level FDR was 0.0001 (8001 decoy PSMs out of 70 million), peptide sequence-level FDR is 0.001 (728 decoy sequences out of 596,839), and the final canonical protein-level FDR was 0.0005 (10 decoy proteins out of 18,267 with canonical status). Among proteins with lesser status (weak, insufficient evidence, etc.) there are 645 decoys out of 21,854 yielding an FDR of 0.03. Because of the tiered system, high quality MSMS spectra that were matched to a peptide are never lost, even if a single matched peptide cannot confidently identify a protein.

Protein Identification Confidence Levels and Classification

Proteins were identified at different confidence levels using standardized assignments to different confidence levels based on various attributes and relationships to other proteins. The highest level is canonical, and the lowest is "not detected". In between are various levels of uncertain and redundant

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

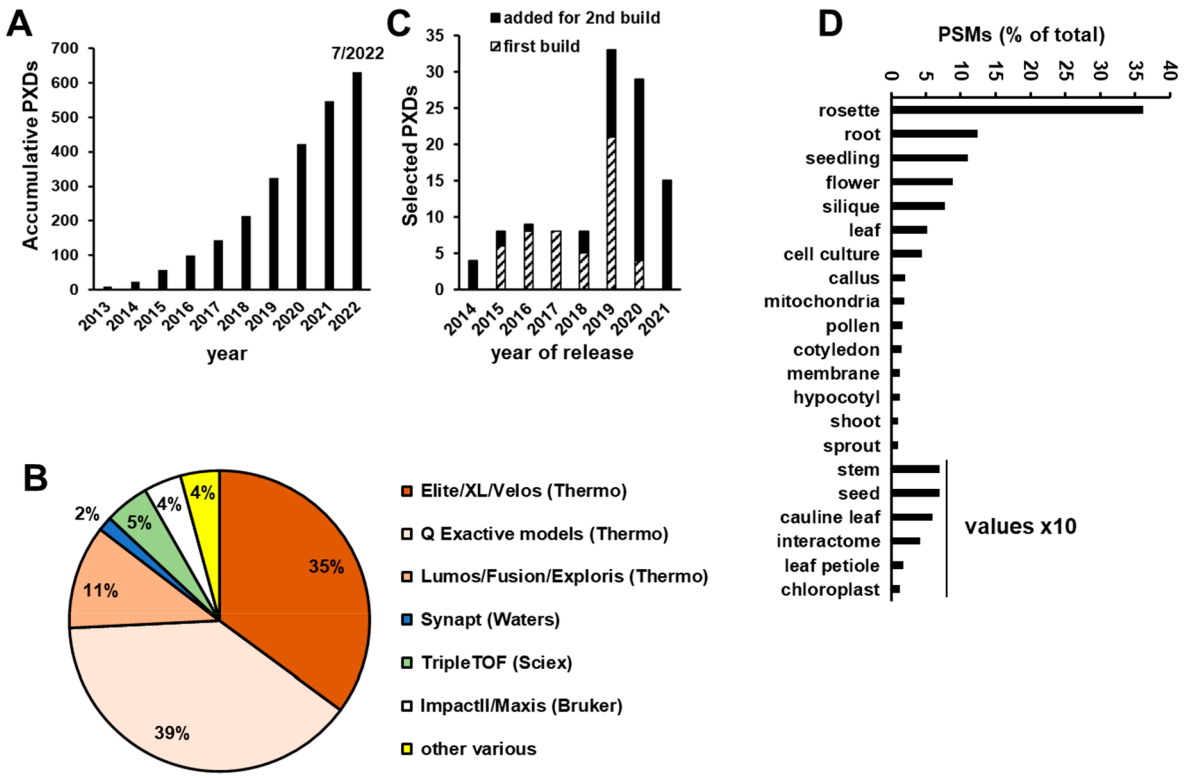


Figure 1. Publicly available PXDs and mass spectrometry instrumentation for *Arabidopsis thaliana* in ProteomeXchange. (A) Cumulative PXD available and selection of PXDs across the different years for *Arabidopsis* PeptideAtlas build 1 and build 2. (B) Mass spectrometry instruments used to acquire data in these PXDs grouped in different categories by vendor and model. The category "other" includes low resolution instruments such as LCQs, LTQs, QStar, as well as MALDI-TOF-TOF. (C) PXDs used in build 1 and PXD added for build 2, grouped per year of PXD release to the public. (D) Distribution of PSMs of samples from different plant parts in build 2.

proteins; this tier system was described in detail in ref 8 and will not be repeated here.

Handling of Gene Models and Splice Forms

The 27,655 protein coding genes in Araport11 are represented by 48,359 gene models (transcript isoforms), which are identified by the digit after the AT identifier (e.g., AT1G10000.1). We refer to the translations of these gene models as protein isoforms. Most protein isoforms are either identical or very similar (differing only a few amino acid residues, often at the N- or C-terminus). 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 "0.1" isoform unless one of the other isoforms had a higher number of matched peptides. 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 canonical (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 among all protein isoforms (for that gene), then they are classified as redundant.

Integration of PeptideAtlas Results in Other Web-Based Resources

PeptideAtlas is accessible through its web interface at https://peptideatlas.org. Furthermore, direct links are provided between PeptideAtlas and PPDB (http://ppdb.tc.cornell.edu/), UniProtKB (https://www.uniprot.org/), TAIR (https://www.arabidopsis.org/), Plant PTM Viewer (https://www.psb.ugent.be/webtools/ptm-viewer/), PhosPhAt (http://phosphat.uni-hohenheim.de/), SUBA5 (https://suba.live/), ATHENA (http://athena.proteomics.wzw.tum.de:5002/master_arabidopsisshiny/), and several more. Links to matched peptide entries in PeptideAtlas are available in the *Arabidopsis* annotated genome through a specific track in JBrowse at https://jbrowse.arabidopsis.org.

Protein Physicochemical Properties and Functions

To characterize the canonical and unobserved proteomes, physicochemical properties were calculated or predicted by using various web-based tools. These include protein length, mass, GRAVY index, isoelectric point (pI), number of transmembrane domains (http://www.cbs.dtu.dk/services/TMHMM), and sorting sequences for the ER, plastids, and mitochondria (http://www.cbs.dtu.dk/services/TargetP-1.0/).

Assembly and Quality Control Filtering of the RNA Data Set

13,673 single and paired end RNA-seq data sets from[57] were run through featureCounts[58] to count reads aligning to each of the 27,655 *Arabidopsis* genes. Lower quality data sets were filtered out based on a minimum total read count (5,000,000), eliminating 7,994 data sets. Transcripts Per Million (TPM) expression values were calculated for the remaining 5,679 data sets. Genes for which expression above zero TPM was not detected in any of the remaining data sets were removed, eliminating 398 genes. The median TPM value for each data

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

Table 1. Summarizing Information for Each PXD in Build 2[a]

Data set identifier	Publication	Also in build 1	Matched # of MS/MS spectra	Matched MS/MS spectra (%)	# Distinct peptides	Instrument	Plant parts (ecotype; default is Col-0)	Subcellular fraction, complex, or interactome	N-Termini; enriched PTMs (S/T/Y-phos, K-ac, K-ubi)	(a) Biotic condition; development; hormone; other
PXD000136	Hesse et al. (2016)	yes	18082	12.96%	3316	LTQ FT	rosette leaves	chloroplast; envelop, thylakoid, stroma		
PXD000521	Svozil et al. (2014)	no	163204	31.06%	14382	LTQ Orbitrap XL	roots		ubiquitination	
PXD000546	Tomizioli et al. (2014)	yes	126969	20.61%	6840	LTQ Orbitrap Velos	rosette leaves	chloroplast; thylakoid domains		
PXD000565	Svozil et al. (2014)	no	174929	37.98%	23291	LTQ Orbitrap XL	rosette leaves		ubiquitination	
PXD000566	Svozil et al. (2014)	no	53695	21.19%	3992	LTQ Orbitrap XL	roots		ubiquitination	
PXD000567	Svozil et al. (2014)	no	907437	47.57%	27003	LTQ Orbitrap XL	roots		ubiquitination	
PXD000568	Svozil et al. (2014)	no	441504	41.84%	22249	LTQ Orbitrap XL	roots		ubiquitination	
PXD000660	Köhler et al. (2015)	yes	8460	9.00%	2442	LTQ Orbitrap Velos	rosette leaves	chloroplast	N-terminome (TAILS)	import mutants
PXD000869	Zhang et al. (2018)	yes	51685	32.81%	3281	LTQ Orbitrap Velos	rosette leaves	chloroplast		clpc1 mutant
PXD000908	Baerenfaller et al. (2015)	yes	359466	16.63%	12343	LTQ Orbitrap XL	rosette leaves			photoperiod
PXD000941	Svozil, Gruissem, and Baerenfaller (2015)	no	206471	27.21%	9626	LTQ Orbitrap XL	rosette leaves	epidermis, mesophyll, vasculature	ubiquitination	
PXD000942	Svozil, Gruissem, and Baerenfaller (2015)	no	66306	6.51%	5573	LTQ Orbitrap XL	rosette leaves	epidermis	ubiquitination	
PXD001207	Köhler et al. (2015)	yes	21063	27.58%	5648	LTQ Orbitrap Velos	rosette leaves	chloroplast; membranes, tic56		
PXD001473	Lin et al. (2015)	yes	10693	8.98%	480	LTQ Orbitrap	cell culture (ler)		phosphorylation	Brassinosteroid
PXD001719	Zhang et al. (2015)	yes	36025	15.93%	10166	LTQ Orbitrap Velos	roots		N-terminome (TAILS)	N-end rule
PXD001855	Venne et al. (2015)	yes	29980	9.24%	11099	Q Exactive	seedlings		N-terminome (ChaFRADIC)	
PXD002069	Linster et al. (2015)	yes	229132	6.92%	6433	LTQ Orbitrap Velos	rosette leaves		acetylation of N-term and lysine	drought, ABA
PXD002160	Correa-Galvis et al. (2016)	yes	67006	11.73%	2691	LTQ Orbitrap Elite	rosette leaves	chloroplast; PsbS interactome		
PXD002186	Nishimura et al. (2015)	yes	244861	43.11%	8681	LTQ Orbitrap	rosette leaves	chloroplast, Clp protease		
PXD002297	Walton et al. (2016)	no	18194	10.62%	6116	Q Exactive	seedlings		ubiquitination	
PXD003162	Lundquist et al. (2017)	yes	247256	30.21%	11321	LTQ Orbitrap Elite	rosette leaves	chloroplast; membrane complexes		BN-PAGE
PXD003516	Wang et al. (2016)	yes	38195	25.95%	14849	Q Exactive	rosette leaves	chloroplast		darkness
PXD003684	Bhuiyan et al. (2016)	yes	114273	28.47%	7004	LTQ Orbitrap	rosette leaves	chloroplast; plastoglobules; pgm48		senescence
PXD004025	Al Shweiki et al. (2017)	yes	463956	32.45%	19225	LTQ Orbitrap Velos	rosette leaves			variability

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

Table 1. continued

Data set identifier	Publication	Also in build 1	Matched # of MS/MS spectra	Matched MS/MS spectra (%)	# Distinct peptides	Instrument	Plant parts (ecotype; default is Col-0)	Subcellular fraction, complex, or interactome	N-Termini; enriched PTMs (S/T/Y-phos, K-ac, K-ubi)	(a) Biotic condition; development; hormone; other
PXD004276	Choudhary et al. (2016)	yes	62409	18.68%	12727	LTQ Orbitrap	seedlings		phosphorylation	cirdadian rhythm
PXD004599	Mattei et al. (2016)	yes	11521	15.10%	2145	LTQ Orbitrap	seedlings		phosphorylation	
PXD004742	Subramanian, Souleimanov, and Smith (2016)	yes	110661	30.87%	5596	LTQ Orbitrap Velos	rosette leaves			salt stress
PXD004896	Willems et al. (2017)	yes	66438	9.58%	24905	LTQ Orbitrap	cell culture (ler)		N-terminome (COFRADIC)	
PXD005600	Schonberg et al. (2017)	yes	50371	14.13%	2242	LTQ Orbitrap Velos	rosette leaves	chloroplast	phosphorylation	
PXD005740	Hander et al. (2019)	yes	1197	1.79%	771	Q Exactive	roots; rosettes			metacaspase
PXD006113	Brocard et al. (2017)	yes	126442	33.74%	10947	LTQ Orbitrap	rosette leaves	lipid droplet		
PXD006328	Strehmel et al. (2017)	yes	27755	8.78%	5167	Q Exactive	roots	exudate		
PXD006347	Née et al. (2017)	yes	3527	1.14%	746	Q Exactive	seed	DOG1 interactome		
PXD006651	Hartl et al. (2017)	yes	156348	59.20%	26635	Q Exactive	rosette leaves	chloroplast	lysine acetylation	
PXD006652	Hartl et al. (2017)	yes	116946	25.60%	15003	Q Exactive	rosette leaves	chloroplast; thylakoid	lysine acetylation	
PXD006694	McBride et al. (2017)	no	896288	64.10%	16941	Q Exactive; TripleTOF 5600	rosette leaves	microsome membrane complexes		
PXD006800	Brault et al. (2019)	yes	269151	52.68%	29671	Q Exactive	cell culture (ler)	total cell extract, plasmodesmata, plasma membrane, microsome and cell wall		
PXD006806	Brault et al. (2019)	yes	638600	73.78%	39245	Q Exactive	cell culture (ler)	plasmodesmata, plasma membrane, microsome and cell wall		
PXD006848	Seaton et al. (2018)	yes	864599	28.98%	28901	LTQ Orbitrap Velos	rosette leaves			light period
PXD007054	Rytz et al. (2018)	no	113434	37.67%	10499	LTQ Orbitrap Velos; Q Exactive	seedlings		sumoylation	heat stress
PXD007600	Uhrig et al. (2020)	no	616208	6.92%	25480	Orbitrap Fusion; Q Exactive	rosette leaves		phosphorylation	diurnal
PXD007630	Koskela et al. (2018)	yes	207912	43.98%	17031	Q Exactive	rosette leaves	chloroplast;KAT	N-terminal/lysine acetylation	
PXD008355	Van Leene et al. (2019)	yes	365300	27.09%	20308	Q Exactive	cell culture (ler)		phosphorylation	
PXD008663	Castrec et al. (2018)	yes	156183	4.58%	4772	LTQ Orbitrap Velos	rosette leaves		N-term and lysine acetylation	
PXD009016	Zhang et al. (2019)	yes	71216	11.00%	10511	Q Exactive	rosette leaves		phosphorylation	
PXD009274	Rytz et al. (2018)	no	101621	29.47%	8762	Q Exactive	seedlings		sumoylation	
PXD010324	Waltz et al. (2019)	yes	434505	47.48%	16691	Q Exactive	flowers; cell culture	mitochondria; ribosome		
PXD010545	Bouchnak et al. (2019)	yes	80068	23.05%	16460	Q Exactive	rosette leaves (ws)	chloroplast; envelope		
PXD010730	Wu et al. (2019)	yes	554183	39.04%	25341	Q Exactive	rosette leaves			gun1 and clpc1 mutants

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

Table 1. continued

Data set identifier	Publication	Also in build 1	Matched # of MS/MS spectra	Matched MS/MS spectra (%)	# Distinct peptides	Instrument	Plant parts (ecotype; default is Col-0)	Subcellular fraction, complex, or interactome	N-Termini; enriched PTMs (S/T/Y-phos, K-ac, K-ubi)	(a) Biotic condition; development; hormone; other
PXD011088	Rugen et al. (2019)	yes	799790	31.44%	24387	Q Exactive	rosette leaves; cell culture (col-0)	mitochondria; ribosome		
PXD011483	McLoughlin et al. (2019)	yes	3778156	49.67%	48727	Q Exactive	rosette leaves; seedling	protein aggregates; HSP101 interactome		
PXD011716	Kosmacz et al. (2019)	yes	105146	15.08%	21595	Q Exactive	seedlings	stress granule		
PXD011759	Wu et al. (2019)	yes	764134	40.48%	36970	Q Exactive	seedlings			gun1 mutant; lincomycin
PXD012708	Zhang et al. (2019)	yes	6971977	60.23%	234220	Orbitrap Fusion Lumos	10 plant parts (rosette leaves, cauline leaf, stems, flower, pollen, siliques, seeds, cotyledons, root, root cell culture)			large scale tissue atlas
PXD012710	Zhang et al. (2019)	yes	1629606	11.94%	89974	TripleTOF 5600 (Sciex)	11 plant parts (rosette leaves, cauline leaf, stems, flower, pollen, siliques, seeds, cotyledons, root, root cell culture)			large scale tissue atlas
PXD013005	Wu et al. (2019)	yes	1038417	49.73%	43943	Q Exactive	seedlings			gun1 mutant; lincomycin
PXD013264	McWhite et al. (2020)	no	344797	16.90%	23471	Orbitrap Fusion	seeds	complexes		ttg1−1 mutant
PXD013321	McWhite et al. (2020)	no	664021	22.82%	41207	Orbitrap Fusion Lumos; Orbitrap Fusion	seedlings	complexes		
PXD013325	Jiang et al. (2019)	yes	8753	13.08%	2587	LTQ Orbitrap Elite	rosette leaves	BSF interactome		
PXD013382	Smith et al. (2020)	no	676706	46.29%	30653	Q Exactive	rosette leaves		phosphorylation	aux; IAA
PXD013494	Montandon et al. (2019)	yes	27644	18.44%	2991	LTQ Orbitrap	rosette leaves	chloroplast; ClpC interactome		CEP5
PXD013495	Huang et al. (2019)	no	4115	3.15%	907	Orbitrap Fusion	cell culture (ler)		sulfenylation	H2O2
PXD013637	Hu et al. (2019)	yes	73524	12.86%	13550	Q Exactive	rosette leaves	CDKD, Cyclin H, H3 interactomes		GFP-TRAP; RFP-TRAP
PXD013646	Furtauer et al. (2019)	yes	2990818	26.55%	35710	Q Exactive; LTQ Orbitrap Elite	rosette leaves (ler)	non aqueous fractionation		cold, high light. gin2−1
PXD013868	Mergner et al. (2020)	yes	19758985	39.03%	391044	Q Exactive HF	30 tissue types		phosphorylation	large scale tissue atlas
PXD014008	Van Moerkercke et al. (2019)	no	1226474	26.19%	29177	Orbitrap Fusion	seedlings			
PXD014292	Fuchs et al. (2019)	no	122197	68.55%	26734	Q Exactive	cell culture	mitochondria		protein copy numbers
PXD014302	Nietzel et al. (2020)	no	32252	34.01%	2977	LTQ Orbitrap Velos	seedlings	mitochondria; cysteine oxidation		
PXD014610	Gemperline et al. (2019)	no	366368	20.14%	9979	LTQ Orbitrap Velos; Q Exactive	seedlings	proteasome subcomplexes		
PXD014617	McWhite et al. (2020)	no	301546	8.33%	15349	LTQ Orbitrap Velos; LTQ Orbitrap	rosette leaves	complexes		
PXD015135	Kretzschmar et al. (2019)	no	657992	56.50%	27068	Q Exactive	seed; seedlings	lipid droplet		seed germination

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

Table 1. continued

Data set identifier	Publication	Also in build 1	Matched # of MS/MS spectra	Matched MS/MS spectra (%)	# Distinct peptides	Instrument	Plant parts (ecotype; default is Col-0)	Subcellular fraction, complex, or interactome	N-Termini; enriched PTMs (S/T/Y-phos, K-ac, K-ubi)	(a) Biotic condition; development; hormone; other
PXD015161	Mair et al. (2019)	no	235942	36.45%	19450	Q Exactive	seedlings	epidermis, guard cells; proximity labeling		
PXD015162	Mair et al. (2019)	no	94840	35.78%	27874	Q Exactive	seedlings	guard cells; nuclei; proximity labeling		
PXD015212	Mair et al. (2019)	no	71984	25.43%	11159	Q Exactive	seedlings	guard cells; proximity labeling; FAMA interactome		
PXD015624	Berger et al. (2020)	no	2003312	48.69%	89519	Q Exactive	rosette leaves and roots	chloroplast; Fe−S clusters		
PXD015636	Berger et al. (2020)	no	9616	15.03%	650	Q Exactive	rosette leaves and roots	chloroplast; Fe−S clusters interactome		
PXD015919	Huang et al. (2020)	no	1676583	47.70%	60917	Q Exactive	seedlings	nuclear membrane; proxomity labeling		
PXD016263	Petereit et al. (2020)	no	6316	4.92%	1917	Orbitrap Fusion Lumos	seedlings	mitochondria	N-terminome (ChaFradic)	
PXD016315	Fila et al. (2020)	no	292510	52.00%	51094	Q Exactive	flowers			nac mutants
PXD016457	Sang et al. (2020)	no	121254	25.65%	23169	Q Exactive	leaf petiole	TF interactome		
PXD016507	Li et al. (2020)	no	14297	13.97%	1276	LTQ Orbitrap	seedlings		phosphorylation	carbon/nitrogen-nutrient stress,
PXD016575	Rodriquez et al. (2020)	no	566166	29.36%	104516	Q Exactive	seedlings			large scale. Autophagy; reprogramming
PXD016746	Petereit et al. (2020)	no	91088	21.35%	5406	Orbitrap Fusion	seedlings	mitochondria		ClpXP
PXD016883	Marondedze et al. (2019)	no	1895	10.93%	1329	Q Exactive	roots		mRNA binding proteins interactome	
PXD017189	Bhyuian et al. (2020)	yes	73870	40.64%	5053	LTQ Orbitrap	rosette leaves	chloroplast		cgep mutant
PXD017380	Data set with its publication pending	yes	425590	19.66%	28385	Q Exactive	rosette leaves	chloroplast; plastoglobules		abck
PXD017400	Liao et al. (2022)	yes	483662	23.00%	19617	Q Exactive	rosette leaves	chloroplast; ClpC interactome		
PXD017430	Armbruster et al. (2020)	no	13249	2.19%	1040	LTQ Orbitrap Velos	rosette leaves		N-terminome (SILProNAQ)	NAA50 mutant
PXD017443	Smith et al. (2020)	no	6753	16.30%	3002	Q Exactive HF	seedlings		phosphorylation	
PXD017444	Smith et al. (2020)	no	2504	11.60%	1123	Q Exactive HF	rosette leaves		phosphorylation	
PXD017663	Armbruster et al. (2020)	no	284977	41.63%	38070	Q Exactive	rosette leaves			
PXD018141	Bach-Pages et al. (2020)	no	27938	11.82%	4731	LTQ Orbitrap Elite	rosette leaves	RNA binding proteins		
PXD018911	Velanis et al. (2020)	no	224619	23.28%	19188	Orbitrap Fusion Lumos; Q Exactive	influorescence	APL2 polycomb complex		
PXD018987	Meteignier et al. (2021)	no	89778	41.45%	3615	Q Exactive	rosette leaves	chloroplast; mTERF interactome		
PXD019253	Rugen et al. (2021)	no	2009301	36.94%	24379	Q Exactive	rosette leaves	mitochondria; complexes BN-PAGE		light; dark
PXD019329	Firmino et al. (2020)	no	143289	18.47%	9924	Q Exactive	leaves, roots, seeds	70S and 80S ribosomes		

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

Table 1. continued

Data set identifier	Publication	Also in build 1	Matched # of MS/MS spectra	Matched MS/MS spectra (%)	# Distinct peptides	Instrument	Plant parts (ecotype; default is Col-0)	Subcellular fraction, complex, or interactome	N-Termini; enriched PTMs (S/T/Y-phos, K-ac, K-ubi)	(a) Biotic condition; development; hormone; other
PXD019330	Bassal et al. (2020)	no	3285309	28.00%	108927	LTQ Orbitrap Velos	multiple tissues			senescence
PXD019603	Escobar etal (2021)	no	1082918	66.62%	24113	orbitrap	rosette leaves	mitochondria		mHSP mutants
PXD019737	Junková et al. (2021)	no	1228828	57.80%	35580	Orbitrap Fusion Lumos	rosette leaves	microsomes		
PXD019904	Scarpin et al. (2020)	no	42620	12.74%	10280	Q Exactive	seedlings		phosphorylation	
PXD019928	Iannetta et al. (2021)	no	10234	4.90%	800	Q Exactive HF-X	rosette leaves			peptidome; peptidase mutant
PXD019942	Scarpin et al. (2020)	no	91	9.13%	19	Q Exactive	seedlings		phosphorylation	early development
PXD020480	Prerostova et al. (2021)	no	852658	57.34%	11650	Orbitrap Fusion Lumos	rosette leaves			cold treatments
PXD020588	Zhang et al. (2020)	no	83041	18.42%	4586	LTQ	rosette leaves	mitochondria; glycolytic interactome		
PXD020700	Bi et al. (2021)	no	9334	8.39%	3478	LTQ Orbitrap Velos	seedlings	spliceosome complex		
PXD020748	Bi et al. (2021)	no	12876	14.50%	3339	Q Exactive HF	seedlings	spliceosome complex		
PXD020749	Bi et al. (2021)	no	40608	45.09%	18191	Q Exactive HF	seedlings	spliceosome complex		
PXD020762	Wilson et al. (2021)	no	53806	17.09%	22713	Orbitrap Fusion Lumos	seedlings		phosphorylation	
PXD021518	Pipitone et al. (2021)	no	1219202	42.93%	49511	Q Exactive HF-X	seedlings; de-etiolation			
PXD021992	Grubbe et al. (2021)	no	143824	8.41%	13424	Orbitrap Fusion	seedlings		ubiquitination	
PXD022684	Parker et al. (2020)	no	15582	3.85%	2640	LTQ Orbitrap Velos	seedlings	RNA binding protein FPA interactome		
PXD023017	Ligas et al. (2019)	no	343086	28.90%	14964	Q Exactive	rosette leaves	mitochondria, OXPHOS complex		
PXD023022	Yperman et al. (2021)	no	14761	4.26%	777	Q Exactive HF	seedlings	TPLATE complex		
PXD023051	Yperman et al. (2021)	no	9644	11.74%	2119	Q Exactive	seedlings	TPLATE complex		
PXD026180	Dahhan et al. (2021)	no	505227	42.46%	50401	LTQ FT Ultra; Q Exactive; Q Exactive HF	cell culture (ler)	trans-golgi network		

[a]More details and breakdowns into individual experiments are provided in Supplemental Data Set 1 and the MetaData annotation system in PeptideAtlas.

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

set was then calculated and used as the threshold for the identification of expressed genes within the data set. Six data sets had a median of 0 and were removed from the analysis. Furthermore, 345 protein-coding genes were never expressed above the median. These genes and the data sets in which they are transcribed are described in the Supplemental Data Set 2.

Machine Learning—Developing Classification Models

The artificial neural network (ANN) model and the random decision forest (RDF) models are trained using Python 3.8.10 with TensorFlow 2.12.0 and TensorFlow Decision Forests 1.3.0, respectively. The input file used for both models is derived from a data set containing 23,674 *Arabidopsis* canonical and unobserved proteins and their attributes. Each entry in the data set includes the protein's identifier, gene symbol, the chromosome on which it is found, its status of being "canonical" or "not observed", number of recorded observations, a short description, molecular weight, gravy, pI, percentage of RNA-seq data sets detecting it, and highest TPM. Only the last five columns are selected for training in the input file. To accommodate the prediction tools, the status is denoted by a 1 or 0 that represents "canonical" or "not observed", respectively. All Python code with learned weights used for the modeling and the output files are available on GitHub at https://github.com/PlantProteomes/ArabidopsisDarkProteome. These models and weights have been uploaded to the GitHub site (https://github.com/PlantProteomes/ArabidopsisDarkProteome). We also added additional programs that load these models and perform the prediction on the original input without the training step. We did this for both the ANN and DF models.

RESULTS AND DISCUSSION

Selection of PXDs

In July 2022, there were ∼630 PXDs for *Arabidopsis* publicly available in ProteomeXchange (Figure 1A) most of which were submitted through PRIDE[59,60] (89%) and the remainder through MassIVE,[61] JPOST,[62] iProX,[63] or Panorama Public.[64] For most of these PXDs (84%), the MS data were acquired using an Orbitrap type instrument from the vendor Thermo (Thermo Fisher Scientific) (Figure 1B).[65,66] 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 a lower number of PXDs with more recent Orbitrap models (Lumos, Fusion, Exploris). The remainder of the PXDs (16%) were acquired on a variety of instruments (e.g., TripleTOF and Maxis/Impact II) from different vendors (Figure 1B). For build 2, we selected 63 new PXDs and analyzed these together with all 52 data sets from build 1 (Figure 1C). Table 1 summarizes key information for all 115 selected PXDs in build 2; additional information can be found in Supplemental Data Set 1. These new PXDs were selected because they appeared the most promising to identify new proteins and selected PTMs, as well as increase sequence coverage of proteins already identified at lower (noncanonical) confidence levels. For example, the selected PXDs concerned specific protein complexes (e.g., mitochondrial ribosomes PXD010324[67]), proximity labeling to target subcellular complexes (e.g., the nuclear pore complex PXD015919,[68] and subcellular localizations (e.g., clathrin-coated vesicles PXD026180[69] that were underrepresented.

We also selected two large studies involving affinity-enrichment for ubiquitination,[27,28] a study enriching for SUMOylated proteins,[70] as well as additional PXDs involving n-terminal or lysine acetylation or phosphorylation. We do note that most PXDs involved the Col-0 ecotype (for which most community resources are available), but one study used ecotype *Wassilewskija* (Ws) and six studies used cell cultures generated from *Landsberg erecta* (Ler). Because of the complexities of data processing and control of the overall false discovery rate (FDR), we excluded data sets obtained through data independent acquisition (DIA), targeted MS (MRM or SRM) and only considered data dependent acquisition (DDA). While DIA data sets 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 fragmented ions.[71] A large ensemble of DDA runs, especially when complex peptide mixtures are prefractionated (by, 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 (e.g., ref 72) and we are starting to develop a mechanism to integrate DIA data into the PeptideAtlas build process. However, we did include stable isotope labeled (multiplexed) proteome data sets, including isobaric iTRAQ and TMT,[73,74] dimethyl labeling,[75] as well as N-terminomics data sets using TAILS[76] or COFRADIC.[77] Finally, we also considered mass spectrometer type, and we note that ∼84% of all available PXDs in ProteomeXchange used Orbitrap-type instruments (Table 1; Figure 1B). We did select two PXDs with important contributions for which the spectra were acquired on other instruments than Orbitraps − TripleTOF5600 (PXD006694 with a large-scale tissue atlas, PXD012710 with microsome samples). Several other (older) instruments used in the ∼650 PXDs have too low resolution and/or mass accuracy for low FDRs in large scale analysis.

Assembly of a Comprehensive Protein Search Space to Maximize Protein Discovery

We assembled a comprehensive protein search space (Table 2) that included the two most recent *Arabidopsis* annotations (Araport 11 and TAIR10). These are still both used in recent proteomics studies even though Araport11 was released in 2017[37] and TAIR10 in 2010.[38] In addition, we added all other *Arabidopsis* (Col-0) sequences from the universal databases UniProtKB and RefSeq because these are widely used sequence resources. To help identify proteins not represented (or with alternative proteoforms) in these four main resources, we also included sequences generated by individual laboratories, including a large set of small Open Reading Frames (sORFs),[41] as well as the predicted protein sequences for 950 Araport11 pseudogenes. These pseudogenes are assumed to be untranslated, but we did previously find evidence that some do appear to produce stable proteins.[8] We also updated the set of the plastid- and mitochondrial-encoded proteins to address redundancies and mistakes in plastid- and mitochondrial ATGC and ATMG sequences and to allow a systematic analysis of nonsynonymous RNA editing for plastid- and mitochondrial encoded proteins. In a just published study,[43] we provided detail on the annotation and redundancy of plastid- and mitochondrial encoded proteins, the expression of organellar ORFs, and the impact of RNA editing. Table 2 shows the number of sequences for each sequence data set, their overlap, and unique protein sequences. All protein

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

Table 2. Summary of Source Databases for the Arabidopsis Search Space[a,b,c,d,e,f]

Source	Sequences	Distinct	Unique	Source (b)	Source (c)	Source (d)	Source (e)	Source (f)	Source (g)	Source (h)	Source (i)	Source (j)	Source (k)	Source (l)
PeptideAtlasAllOrganellar (a)	197	195	34	114	123	106	110	0	110	103	0	0	0	37
PeptideAtlasMinimalOrganellar (b)	114	114	0		114	64	65	0	93	79	0	0	0	27
AraportUpdated (c)	42,617	40,716	0			40,666	31,026	0	38,651	40,660	0	0	0	1112
Araport11 (d)	48,359	40,784	10				31,133	0	38,700	40,654	0	0	0	1147
TAIR10 (e)	35,386	32,785	1500					0	29,401	31,032	0	0	0	1057
Pseudogenes (f)	3720	3702	3701						0	0	0	0	1	0
UniProtKB (g)	39,342	39,273	373							38,669	0	0	0	1115
RefSeq (h)	48,265	40,709	5								0	0	0	1116
ARA-PEP:LW (i)	16,809	16,628	16,478									21	129	0
ARA-PEP:SIPs (j)	607	606	565										20	0
ARA-PEP:sORFs (k)	7901	7764	7614											0
IowaORFs (l)	7481	7270	6116											

[a]PeptideAtlasAllOrganellar includes all the PeptideAtlas_ATxGnnnnnnn.1 (original),.2 (RNA edits [major]),.3 (RNA edits [major and minor]),.4 (RNA edits [major, minor, and truncations]). [b]PeptideAtlasMinimalOrganellar includes one protein for each organellar gene, the RNA edited [major only] version if there are edits, or the original if no editing sites. [c]AraportUpdated begins with the Araport11 proteome with all organellar proteins replaced with PeptideAtlasMinimalOrganellar set and other corrections discussed in this article applied. [d]Araport11 represents the current set of Araport11 proteins as downloaded 2021−04−26. [e]TAIR10 represents the current set of TAIR10 proteins as downloaded 2020−12−22. [f]Pseudogenes represent an additional set of entries labeled as "pseudogenes" in Araport11 and are thus not exported as part of the proteome - downloaded 21−02−23

sequences in the search space can be downloaded from the PeptideAtlas Web site.

Protein Identification, Sequence Coverage, PTMs, and Overall Statistics in Build 2 (2023−10)

The 115 selected PXDs contained 259.4 million raw MSMS spectra from 10478 MS runs that we searched as 369 different experiments (Table 3 and Supplemental Data Set S1). We assigned these experiments based on the metadata associated with the PXDs, as well as associated publications. Importantly, this involved manual evaluation of experimental conditions, sample preparations and proteomics and MS workflows; this is a relatively time-consuming process requiring specific expertise which is currently hard to automate. This allowed us to search with appropriate parameters (parameters need to be assigned for specific PTMs, protease cleavage reagents, iTRAQ, TAILS, and COFRADIC) and also to associate the most relevant biological (e.g., dark vs light treatments) and technical metadata. The associated metadata (available through direct links with, e.g., experiments and matched spectra in PeptideAtlas) will facilitate discoveries of biological relevance (e.g., condition or cell-type specific accumulation patterns, the relation between alternative splicing and plant material) but also for analysis of technical features (e.g., sample-handling related PTMs such as off-target effects of iodoacetamide[78,79] or trypsin artifacts[80,81]).

Table 3. Comparison of the Summary Statistics of Arabidopsis PeptideAtlas Builds 1 and 2

Metric	Build 1	Build 2	Ratio of 2/1
Data sets (PXDs)	52	115	2.21
Experiments	266	369	1.39
MS Runs	6148	10,478	1.70
MS2 Spectra Acquired[a]	142,703,610	259,383,093	1.82
MS2 Spectra Scored[b]	125,181,633	210,655,824	1.68
PSM FDR	0.001	0.0008	0.80
PSMs passing threshold	39,480,811	70,470,125	1.78
AA sequence coverage	49.5%	51.6%	1.04
Distinct Peptides	535,340	596,839	1.11
Canonical proteins (Araport11[c])	17,858	18,267	1.02
Uncertain proteins (Araport11[c])	1942	1856	0.96
Redundant proteins (Araport11[c])	1600	1540	0.96
Not observed proteins (Araport11[a])	6255	5896	0.94
Araport11[c] proteins with peptides mapped	21,400	21,663	1.01

[a]Information in raw files. [b]Spectra of sufficient quality to be scored. [c]Araport11 but with updated plastid and mitochondrial encoded proteins (114 instead of 210 in original Araport11) and total size is 27,559 proteins.

In total there were 70.5 million peptide-spectrum matches (PSMs) to nearly 0.6 million distinct peptides, thereby identifying 18,267 Araport11 proteins at the highest confidence level (canonical proteins, two uniquely mapping non-nested peptides of ≥9 residues and with ≥18 residues of total coverage), 1856 "uncertain" proteins (too few uniquely mapping peptides of ≥9 aa to qualify for canonical status and may also have one or more shared peptides with other proteins), and 1540 "redundant" proteins (containing only

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

Table 4. Proteins Identified in Araport11[a] for Each of the Four Confidence Categories in Build 2 for Mitochondrial (M) and Plastid (C) Chromosomes and the Nuclear Chromosomes (1−5)

Chromosome*	Entries	Canonical, n (%)		Uncertain, n (%)		Redundant, n (%)		Not Observed, n (%)	
M	35	27	77.1%	5	14.3%	0	0.0%	3	8.6%
C	79	63	79.7%	12	15.2%	0	0.0%	4	5.1%
1	7156	4730	66.1%	502	7.0%	384	5.4%	1540	21.5%
2	4317	2762	64.0%	290	6.7%	240	5.6%	1025	23.7%
3	5460	3630	66.5%	353	6.5%	296	5.4%	1181	21.6%
4	4180	2788	66.7%	282	6.7%	247	5.9%	863	20.6%
5	6332	4267	67.4%	412	6.5%	373	5.9%	1280	20.2%
Total	27,559	18,267	66.3%	1856	6.7%	1540	5.6%	5896	21.4%

[a]Araport11 but with updated plastid and mitochondrial encoded proteins (114 instead of 210 in original Araport11) and total size is 27,559 proteins.

Table 5. Peptides Assigned to Proteins by a Hierarchy of Sources Ranging from Araport11 to DECOY, with Each Peptide Assigned Only to the Highest Source Possible and Then Not to Any Other Source

Hierarchy[a]	Primary protein match	No. of peptides	No. of PSMs	No. of primary proteins	No. of peptides (≥3 PSMs)	No. of PSMs (≥3 PSMs)	No. of primary proteins (≥3 PSMs)	No. of primary proteins (≥2 distinct peptides with ≥3 PSMs)
1	Araport11	595,346	70,409,850	20,876	411,364	70,166,505	18,860	17,056
2	TAIR10	438	33,123	69	271	32,908	43	25
3	PSEUDOGENE	205	1264	126	74	1104	54	9
4	UniProtKB	197	14,408	38	120	14,306	28	15
5	RefSeq	0	0	0	0	0	0	0
6	ARA-PEP:LW	101	519	82	30	440	21	3
7	ARA-PEP:SIPs	5	17	5	2	14	2	0
8	ARA-PEP:sORFs	75	404	60	29	352	18	3
9	IowaORFs	466	10,409	157	232	10,111	61	26
10	CONTAM[b]	5217	1,719,466	95	3577	1,717,256	88	83
11	DECOY[c]	728	8001	654	281	7470	269	10

[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.

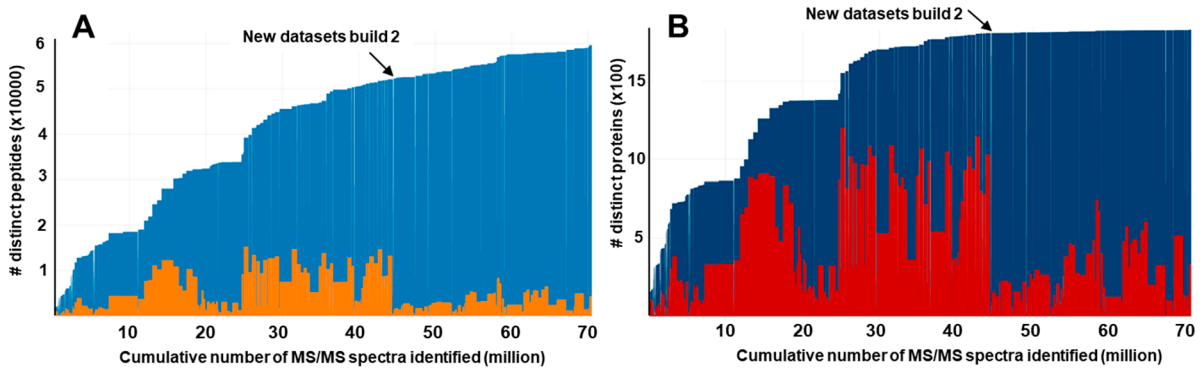


Figure 2. Contributions of individual experiments to the PeptideAtlas Build. (A) From the 369 experiments conducted, the graph displays the total number of distinct peptides for the build as well as the number of peptides contributed by each experiment. (B) The plot shows the cumulative number of distinct proteins and the number of proteins that were contributed to each experiment. The location where new data sets added since the first build is marked.

peptides that can be better assigned to other entries, and thus these proteins are not needed to explain the observed peptide evidence) (Table 3). The overall FDR of the PSMs was 0.08%. The "uncertain" proteins are needed to explain all the peptides identified above threshold, while "redundant" identifications have only peptides that already map to canonical or uncertain proteins; for more details on these definitions, see ref 8. These "redundant" proteins typically have significant sequence homology to these canonical proteins. Table 4 shows the breakdown of identifications at different confidence levels for each of the five nuclear chromosomes as well as the plastid and mitochondrial genomes. The percentage of identified predicted proteins per nuclear chromosome was on average 78.6% with only small differences between chromosomes. We identified nearly all predicted mitochondrial and plastid proteins and ORFs (91% and 95%, respectively); the few unobserved organellar proteins are either untranslated ORFs (likely pseudogenes) or very small proteins. In summary, build 2 has peptides mapping to 78.6% (21,663/27,559) of all predicted proteins in Araport11 (counting only one isoform per gene) and has a very low FDR (0.08%) for spectral matches (PSMs). The complete sets of identified proteins in

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

their respective confidence tiers can be downloaded at https://peptideatlas.org/builds/arabidopsis/.

In addition, there were 4342 peptides matching only to proteins in sources other than Araport11 with a total of 1.8 million PSMs (Table 5). These peptides were assigned to proteins by hierarchy of sources (ranked from 1 to 11), with each peptide assigned only to the highest-ranking source possible and then not to any other source. Table 5 also summarizes how many of these non-Araport11 proteins were identified when different thresholds were applied for the minimum number of PSMs and matched peptides. For example, when requiring at least 2 distinct peptides with each at least 3 observations (PSMs) there are 25 proteins identified in TAIR10 and nine pseudogenes, as well as 6 small proteins (LW or sORFs) from the ARA-PEP database. Supplemental Data Set S3 provides more information about these proteins not found in Araport11. These matched pseudogenes and non-Araport11 proteins should be considered for incorporation into the next *Arabidopsis* genome annotation. Finally, what this Table 5 also demonstrates is that samples also contain various contaminants (e.g., keratins from human skin, trypsin for autodigestion, BSA), as expected based on observations in other large-scale studies.[82,83]

Build 2 contains more than double the number of PXDs as build 1, and 68% more raw MSMS spectra were searched (Table 3). Whereas the number of PSMs increased by 78%, the number of distinct identified peptides increased by only 11% and the number of identified proteins (across all confidence levels) increased by just 1% (Table 3). Figure 2A shows the cumulative identified peptides as well as distinct peptides from the 369 experiments (each PXD can have more than one experiment), whereas Figure 2B shows the cumulative identified canonical proteins as well as distinct canonical proteins from the experiments. This shows that even though we deliberately selected PXDs to enrich for underrepresented proteins, this did only incrementally increase peptide and protein discoveries, despite the near doubling of matched PSMs. This clearly suggests that identification of the remaining 21% of the predicted proteome will require new approaches. Importantly, to maintain a high-quality protein identification and avoid accumulation of false discovery of peptides, we kept the PSM FDR low at 0.08% in build 2 (0.1% in build 2). This FDR is lower than used in many individual studies (often PSM FDR is set at 1%) but is needed because of the very high number of spectra searched compared to individual studies.

To better understand possible underlying causes for these diminished returns, we investigated the relationships between the number of matched spectra and identified distinct peptides or proteins for each PXD. This showed a wide PSM match rate for searched spectra between PXDs ranging from 1% to 74% (Table 1) mostly due to differences in spectral quality (due to, e.g., peptide abundance, instrument settings and sensitivities, sample preparation), but a strong positive linear correlation between the number of matched spectra and identified distinct peptides or distinct proteins (Supplemental Figures S1 and S2). Interestingly, plotting the % of matched spectra to identified distinct peptides or proteins showed a clear saturation, suggesting bottlenecks in the dynamic range for protein identification (Supplemental Figures S1 and S2). This suggests that dramatic innovations in mass spectrometry and/or proteomics workflows and sample selection are needed to identify the remaining 21.4% of the predicted proteome (Conclusions and Future Perspective).

Mapping Biological PTMs; N-Terminal and Lysine Acetylation, Phosphorylation, and Ubiquitination

We selected multiple PXDs that specifically enriched for the physiologically important PTMs of phosphorylation, N-terminal acetylation, lysine acetylation, or ubiquitination (Table 1). A sophisticated PTM viewer in PeptideAtlas allows detailed examination of these PTMs, including direct links to all spectral matches. PTM identification rates strongly depend on the confidence level (minimal probability threshold) of PTM assignment. We limited our summary in this publication on PTMs to canonical proteins, but PTMs for all confidence levels of protein identification are available in the PeptideAtlas web interface. Here we used localization probability $P \geq 0.95$ from PTMProphet[84] for each PTM, and also required at least 3 PSMs for a specific PTM at a specific residue to be included in the overall statistics. In general, higher numbers of repeat observations (PSMs) for a specific PTM at a residue improve the reliability of the assignment. Conversely, peptides with high PSM counts (e.g., hundreds or more) for which the vast majority (e.g., 99%) of peptide do not have a reported PTM at $P > 0.95$, are possibly false discoveries. We recommend therefore to use the PeptideAtlas to evaluate specific PTM sites if these are of particular interest to the reader. We evaluated the results for false positives and possible pitfalls in various ways, including spot checking matched spectra and proteins to which PTMs were mapped. Supplemental Data Sets S4, S5, S6, and S7 provide the results for these four PTMs and Supplemental Data Set S8 provides the combined results of these PTMs per canonical protein to analyze for possible cross-talk between PTMs. In total, there are 5764 canonical proteins with one to four PTM types and a total of 17,675 modified amino acids identified by 0.58 million PSMs. We briefly summarize the results below:

N-Terminal Acetylation (NTA)

Proteins are synthesized with an initiating N-terminal methionine that can be N-terminally acetylated. However, a large portion of cellular proteins undergo removal of the initiating methionine residue by methionine amino peptidases (MAPs) if the side chain of the second residue is small enough.[85,86] If the N-terminal methionine is removed, NTA can occur on the second residue of the predicted protein. Both methionine removal and NTA are cotranslational processes that occur in the cytosol and plastids.[87−89] However, nuclear-encoded proteins synthesized in the cytosol and then sorted into chloroplasts can undergo post-translational NTA after removal of the cleavable chloroplast transit peptide (cTP) by several N-terminal acetyltransferases (NATs) in the chloroplast.[87,88] Indeed, intrachloroplast NTA has been documented by several studies mostly involving N-terminal labeling with stable isotopes followed by fractionation (TAILS, SILProNAQ, COFRADIC)[89−92] and will not be further addressed in this study. The presence of NATs in the nucleus (NAA50), ER (NAA50) and plasma membrane (NAA60) allows for additional post-translational NTA after sorting to these respective subcellular locations,[88] thus adding to the complexity of NTA patterns. Finally, proteins sorted to mitochondria with cleavable N-terminal sorting signals typically do not accumulate with an acetylated N-terminus[93] and indeed no NAT has been reported to localize to mitochondria. When peptides are identified matching to the initiating methionine or the immediate downstream residue of a protein, this is important support for the lack of cleavable N-

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

terminal sorting signals (because the sorting and cleavage process and subsequent degradation of the cleaved signal peptide are typically very efficient).

After removal of false positives (see below), the search process identified 3185 Araport11 canonical proteins (including 18 chloroplast- and 5 mitochondrial-encoded proteins) containing 3258 NTA sites mostly at position 1 (M) or position 2, and the remainder further downstream (Supplemental Data Set S4). 98% of these NTA proteins contained a single NTA site. The 2% of cases where more than one NTA site per protein was found could be due to alternative splice forms or translation start sites,[89] proteins sorted to one or more subcellular location or false discovery of the PTM (there is no known sample preparation induced NTA). Interestingly, we found 30 false positive NTAs in four (iso)leucine-repeat peptides (sequences: IIIIIIIII or VIIIII or VIIIIII or VVLLIIL matching to 27 canonical proteins). [Acetyl]-V has an identical mass as [Formyl]-L or I (L and I are isobaric), and these false positives stem from this misassignment. Formylation can occur at any peptide N-terminus (and the side chain of T and S) and is a common PTM induced by formic acid (even at low concentrations).[94,95] We also noted false positives due to combinatorial (assigned or real) mass modifications, involving deamidation (+0.98402 Da), carbamylation (+43.00582 Da), and C12/C13 isotopes (+1 Da), especially when the assigned NTA (+42.01056 Da) was observed with an absolute low number of PSMs or a relative low number of PSMs compared to the total number of PSMs for that peptide (for highly abundant proteins).

There were 1493 nuclear-encoded canonical proteins with matched peptides starting exclusively with the initiating methionine, of which 1164 were observed with NTA. There were 2810 nuclear-encoded canonical proteins with matched peptides starting exclusively at position 2, of which 1912 were observed with NTA. These acetylated residues were mostly for proteins without predicted N-terminal signal peptides (sP, cTP, or mTP). We created sequence logo plots for each of these four groups (Figure 3A–D) to show the methionine amino peptidase activity (to remove the initiating M) and the NAT activity. The logos show that proteins that retain the methionine have mostly the acidic amino acids residues (D,E) and N in the second position (Figure 3A). NTA occurs on the initiating M (Figure 3C), as well as on A, S, V, and G (Figure 3D). The iceLogo[96] (Figure 3E) comparing the sets in panel B and D shows that the dominant NAT activity for this set of identified proteins is to acetylate A and S residues. NTA is the result of the activity of multiple NATs each with their own set of preferred substrates and NATA has been reported to be responsible for N-terminal acetylation of ~50% of the plant proteome.[88]

Lysine Acetylation

Identification of K-acetylation required a targeted search that was applied on the raw files from three PXDs with enriched lysine acetylome samples from the Finkemeier lab (PXD006651, PXD006652, and PXD007630) (Table 1). After application of our postsearch selection criteria (PTM localization $P > 0.95$; ≥ 3 PSMs per PTM site) and removal of false positives, we identified 864 core canonical proteins containing K-Acetyl modifications representing 1750 K-sites (Supplemental Data Set S5). 512 proteins (59%) contained a single K-acetyl site, whereas others are more heavily K-acetylated. The acetylated proteins were distributed across

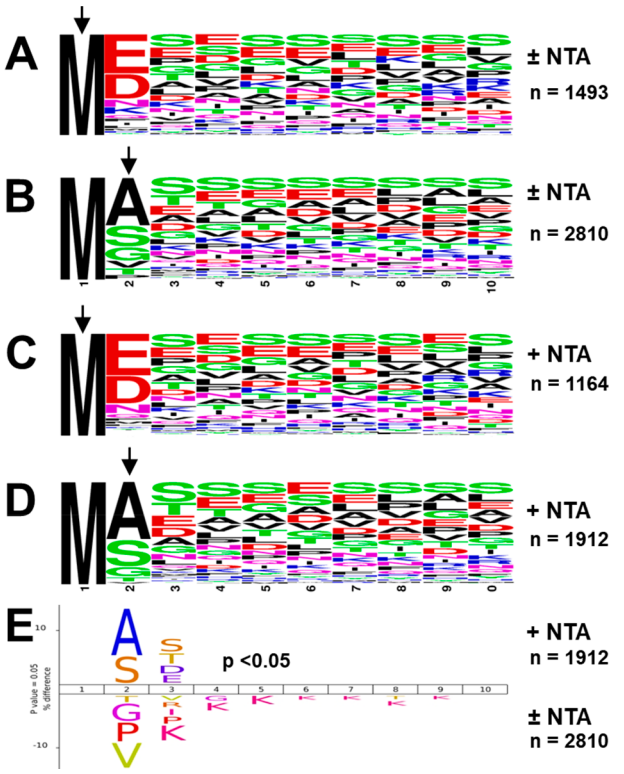


Figure 3. N-Terminal consensus sequence patterns of canonical nuclear-encoded proteins accumulating with the initiating methionine or the 2nd residue (after methionine excision) with or without NTA. (A, B) Sequence logos of proteins (first 10 residues are shown) that are exclusively found with the initiating methionine (A) or exclusively found with just this methionine removed (B), irrespective of NTA. (C, D) Sequence logos of NTA proteins (first 10 residues are shown) exclusively accumulating with the initiating methionine (C) or exclusively found with the second residue (methionine removed). (E) Icelogo for NTA canonical proteins exclusively starting at position 2 using all canonical proteins starting exclusively at position 2, but irrespective of the NTA status. Arrows indicate the observed N-terminal residue.

multiple subcellular locations and functions supporting recent findings in *Arabidopsis*,[97] but also other plant species,[98] the green algae *Chlamydomonas reinhardtii*[99] as well as the moss *Physcomitrium patens*.[100]

Phosphorylation

After application of our postsearch selection criteria (PTM localization score $P > 0.95$; ≥ 3 PSMs per PTM site), there are 5198 canonical phosphoproteins (p-proteins) representing 14,748 phosphosites (p-sites) (86% S, 13% T, 0.6% Y) (Supplemental Data Set S6). 45% of the 5198 p-proteins contained only a single p-site, and 20%, 11% and 7% contained 2, 3, or 4 p-sites, respectively. This ratio between pS, pT and pY is consistent with published literature for large scale phosphorylation data sets in *Arabidopsis*.[26,101] When considering all p-sites in PeptideAtlas with p-site score of $p > 0.80$ and without applying a minimum PSMs repeat observation, PeptideAtlas records 206,660 p-sites considering all protein identification tiers (for more information about the number of p-sites at different threshold levels see the PeptideAtlas home page). This illustrates that the number of reported p-sites (and other PTMs) greatly depends on minimal thresholds (FDR PSM, PTM site score, repeat PSMs) as well as well the protein search space and parameters. This has also been addressed in a

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

previous meta-data analysis of p-proteomics data sets in *Arabidopsis*,[101] and also in ref 102 for a re-evaluation of tyrosine phosphorylation sites in *Arabidopsis* chloroplasts.

Ubiquitination

We found 668 ubiquitinated core canonical proteins based on 765 single K-glycine (KG) sites[27] and 412 K-diglycine (KGG) sites,[28] totaling 1177 ubi-sites (Supplemental Data Set S7). The two PXDs that contained enriched ubiquitinated sites were from large scale studies[27,28] that applied different methods (resulting in K-G or K-GG) to identify the ubiquitinated sites. 449 proteins (67%) contained a single G or GG PTM site. By far the most PSMs for G or GG were found for nine ubiquitin (extension) proteins (>1000 PSMs), followed (albeit at far lower PSM levels) by several plasma membrane proteins and histones. We note that there are no mitochondrial-encoded proteins and one chloroplast-encoded protein PeptideAtlas_ATCG00900.1 (30S ribosomal protein S7A/B) with just three PSMs for one site (K13-G). 45 sites across 18 proteins exhibited both a Gly and a GG PTM. Since the G and GG studies were independent, this might indicate that these sites have a lower FDR than sites which were only detected by one of the methods. These 18 proteins are the nine ubiquitin or ubiquitin extension proteins which is logical since they form polyubiquitination chains. The others are abundant glycolytic enzymes (aldolases), cytosolic ribosomal proteins, an elongation factor involved in cold-induced translation (LOS1),[103] the SNARE protein AtVAM3p,[104] and two enzymes involved in amino acid metabolism (Supplemental Data Set S7). It is perhaps not surprising that there was so little overlap between ubiquitination sites between these two studies because ubiquitination is generally a transient PTM, and in the case of polyubiquitination this leads to rapid degradation. Furthermore, plant materials, sampling, and methodologies were very different across these two studies.

Summary of the PTMs

At the chosen minimal thresholds, we identified 5764 canonical proteins with one or more of these four PTMs (NTA, Kac, P, or UBI) based on 0.582 million PSMs for 17675 PTM sites (Supplemental Data Set S8). 4952 proteins contain only one type of PTM, 635 proteins contain two types of PTMs, 160 proteins contain three types of PTMs, and 17 proteins contain all four types of PTMs.

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, many of which are induced during sample processing (see Methods). These modifications generally have 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 4. This shows that the oxidation of methionine is by far the most frequent (12.8% of all PSMs), followed by deamidation of asparagine (5%) and

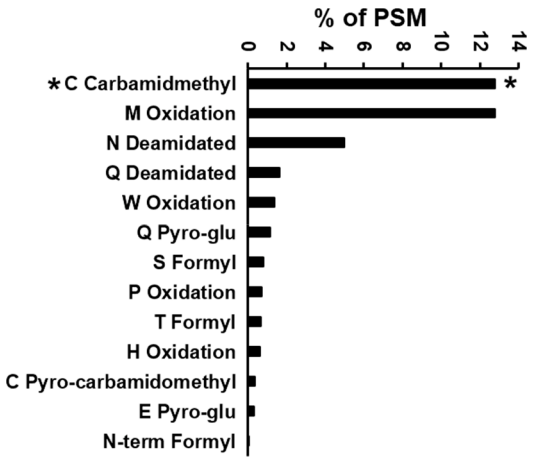


Figure 4. Percentage of PSMs (of total) with mass modifications is mostly associated with sample preparations. Numbers are computed as the total number of PSMs that include at least one instance of the listed mass modification. Some PSMs contain more than one mass modification of the same type (not multiple counted) or different type (multiple counted). * fixed modification.

glutamine (1.6%), tryptophan oxidation (1.4%), pyroglutamate from N-terminal glutamate (1.2%), formylation of serine (0.8%) and threonine (0.7%), oxidation of proline (0.7%), and a few others. Carbamidomethylation of cysteine (searched as fixed modification) from alkylation was high, at 12.8%. These mass modifications are also visible in the PeptideAtlas web interface, with viewable spectra and their interpretations.

Understanding the Nature of the Unobserved Proteomes in the New Release of *Arabidopsis* PeptideAtlas

Of the 27,559 predicted nuclear and organelle protein coding genes of the *Arabidopsis* in Araport11, we identified 18,267 (66.3%) corresponding proteins as meeting the canonical criteria (canonical proteins) and 5896 proteins (21.3%) having no observations at all (dark proteins) in our PeptideAtlas build. The remaining identified proteins are in uncertain or redundant categories. Our working hypotheses is that the dark proteins are not observed because they: (i) are generally expressed at too low levels for detection, (ii) are expressed only under very specific conditions or in specific cell types, (iii) have short half-life preventing proteins to accumulate at significant steady state levels, (iv) have physicochemical properties (very small and/or very hydrophobic) that make them difficult to detect using standard proteomics and mass spectrometry workflows,[8] or (v) simply not translated at all under any conditions.

Figure 5A displays the histograms of molecular weight (between 0 and 80 kDa) for the canonical and dark proteins. Figure 5B displays the relative proportion of the canonical and dark proteins in each kDa bin. Below 4 kDa all proteins are dark proteins. Between 14 and 16 kDa, ~50% of the proteins are canonical and ~50% are dark. With increasing molecular weight, the proportion that are canonical proteins increases to ~90%. There are a substantial number of proteins above 80 kDa, but the proportion of proteins that are canonical is generally constant above 80 kDa at ~95%. Figure 5C and D display the distribution of hydrophobicity computed as the gravy score based on the algorithm of Kyte and Dolittle.[105] Values above 0 are considered hydrophobic, with values above 0.5 being very hydrophobic. Figure 5C shows the absolute number of proteins per bin, whereas panel 3D shows the

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

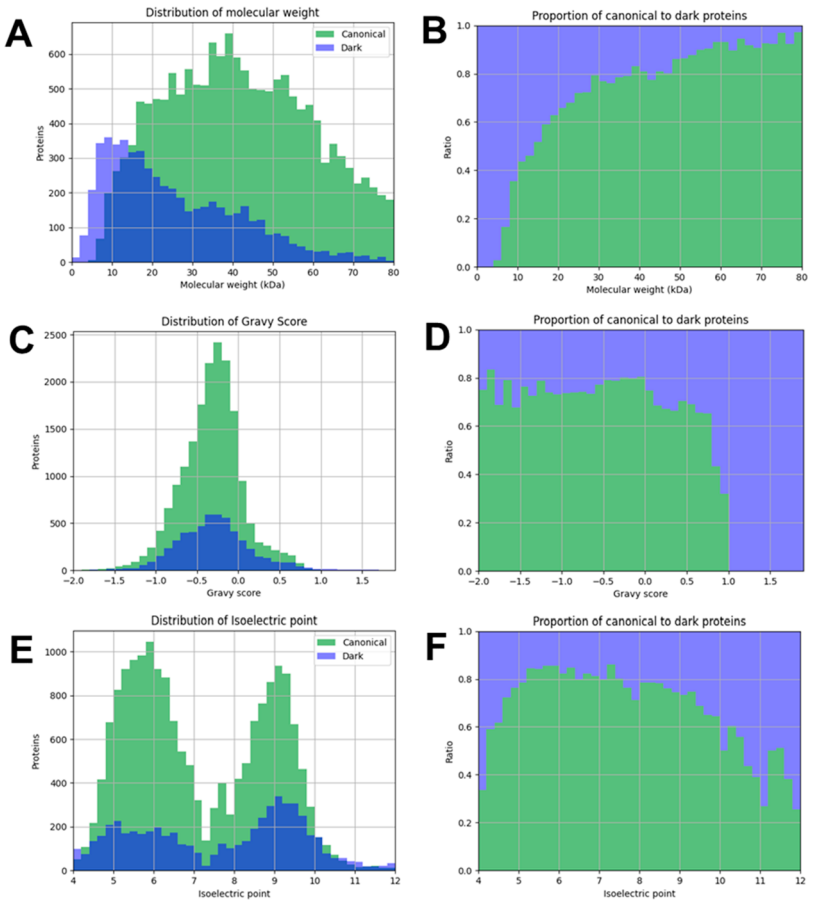


Figure 5. Distributions of physical−chemical properties of the 18,079 canonical (green) and 5595 dark (purple) proteins. (A−E) Absolute counts of proteins within each bin for canonical and dark proteins. (B, D, F) The proportion of canonical and dark proteins within each bin. (A, B) Distributions and proportions of the molecular weight (kDa) of canonical (green) and dark (purple) proteins. Proteins with molecular weights between 0 and 80 kDa are shown. (C, D) Distributions and proportions of the hydrophobicity (gravy score) of canonical (green) and dark (purple) proteins. Proteins with gravy score between −2.0 (hydrophilic) and 2.0 (very hydrophobic) are shown. (E, F) Distributions and proportions of the isoelectric point (pI) of canonical (green) and dark (purple) proteins. Proteins with pI between 4.0 (acidic) and 12 (very basic) are shown.

relative proportion of canonical and dark proteins per bin. The two distributions are broadly similar between gravy scores −2.0 to +0.8, with a sharp decline in the proportion of canonical proteins above a gravy score of +0.8. All 64 proteins with a gravy score greater than +1.0 are dark (i.e., undetected) and most of these proteins are small with a predicted signal peptide for secretion to the ER. Furthermore, most of these very hydrophobic proteins have no known function, but also include seven arabinogalactan proteins (AGPs)[106] and four plasma membrane RCI2 proteins.[107] Figure 5E and F display the distribution of the isoelectric point (pI) for proteins. Both canonical and dark proteins exhibit the typical bimodal distributions peaking at just below 6.0 and again at just above 9.0 based on their total counts (Figure 5E). The distribution in the relative proportion of canonical to dark proteins is complex (Figure 5F), but in general, the proportion of canonical proteins is substantially reduced at the two extremes. Very basic proteins (pI) are enriched for ribosomal proteins and "hypothetical" proteins.

In addition to these inherent properties of the canonical and dark proteins, we also explored the distributions of computed RNA abundances of the transcripts across 5673 single- and paired-end RNA-seq quality-controlled and filtered data sets from (Palos et al.,[57] 2022) with reads aligned to the *Arabidopsis* genome (Methods). We excluded 345 protein coding genes that were never expressed above the median, as well as 309 undetected genes from the remaining analyses which were likely undetected due to mapping limitations with overlapping or highly similar genes (Supplemental Data Set S2). To evaluate mRNA expression patterns for the canonical and dark proteins, we considered two metrics, i.e. the percentage of RNA-seq data sets in which the transcript for a gene was detected (Figure 6A,B) and the maximum transcripts per million (TPM) for each expressed gene in any one of the RNA-seq data sets (Figure 6C,D). Figure 6A displays the distribution of the percentage of the 5673 RNA-seq data sets in which each transcript was detected. The highest bin (99−100%) is truncated at 1000 genes to better show details of the other bins (the true height of this highest bin is 12,000). The relative proportions of canonical and dark proteins in each transcript bin are more easily seen in the proportion plot (Figure 6B) which shows that the proportion increases linearly across most of the range of transcript detection, except for the extremes at the ends.

In other words, the more often a transcript for a gene is detected in one of the RNA-seq data sets, the higher the chance that the protein is canonical. For genes where this RNA detection percentage was below ∼5%, the predicted protein was typically not detected (i.e., dark), whereas for genes where the transcript was detected in >98% of the RNA-seq data sets,

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

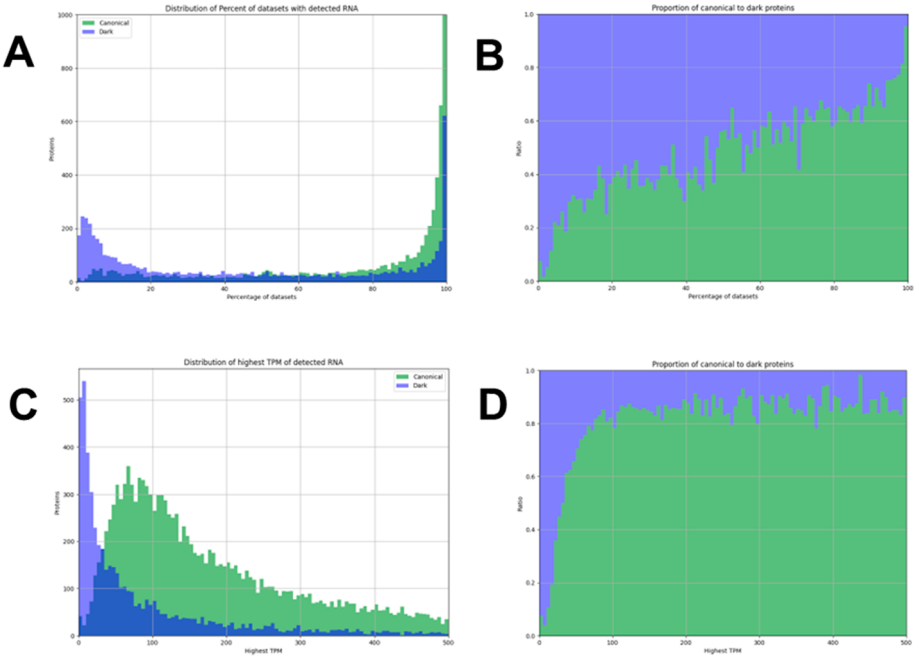


Figure 6. Transcript abundance and observation frequency of 26,975 nuclear-encoded protein coding genes in 5673 high quality RNA-seq data sets. (A, B) Distributions of the percentage of RNA-seq data sets with detected transcripts associated with the canonical (green) and dark (purple) proteins. (A) Absolute counts of proteins within each bin and (B) proportion of light and dark proteins within each bin. (C, D) Distributions of the maximum transcripts per million (TPM) among all RNA-seq experiments for the detected transcripts associated with the canonical (green) and dark (purple) proteins. Absolute counts of proteins within each bin (C) and the proportion of light and dark proteins within each bin (D). The number of TPM extends as high as 207,000 for seed storage protein albumin 3 (AT4G27160), followed by seed storage cruciferin 1 and 3 (AT5G44120 and AT4G28520), Rubisco small subunit 1A (AT1G67090) and the hypothetical very small (33 aa) protein AT2G01021.

the predicted protein was nearly always canonical. Figure 6C depicts the distribution of the highest TPM among the analyzed RNA-seq experiments for each of the canonical and dark proteins. The TPM values extend as high as 207,000 (for seed storage protein albumin 3 - At4G27160) but the proportion does not change substantially above 100 TPM, and we only depict the range 0 to 500 TPM. Clearly the proportion of dark proteins rapidly increases when the maximum TPM falls below ∼100 TPM, but many proteins were still identified even if the TPM for the gene was well below 100. As is also well-known from other studies, this demonstrates that the relation between mRNA abundance and MSMS-based protein identification is complex and impacted by other factors.[26,108,109] A major confounding issue is that the number of protein copies generated per mRNA transcript across genes is not fixed, i.e. some mRNAs are translated very often, whereas others are not, an extreme example is mRNA stored in stress granules. Also, proteins with short half-lives as compared to the half-lives of the corresponding mRNAs are likely to show a poor correlation.

Machine Learning Models to Predict and Understand MS-Based Detection of *Arabidopsis* Proteins

Figures 4 and 5 show that each of the protein and RNA attributes has a substantial influence on whether proteins are canonical or dark. Taking advantage of these attributes to better understand why these dark proteins are not observed, we trained both an artificial neural network (ANN) model and a random decision forest (RDF) model for the canonical and dark proteins based on physicochemical protein properties and RNA expression patterns. The quantitative output of these models was the probability for proteins to be canonical. The starting point was a table of 18,079 nuclear-encoded canonical proteins and 5595 nuclear-encoded dark proteins for a total of 23,674 proteins (uncertain and redundant proteins are left out for the training of the models; proteins without RNA values are also left out), as well as the computed physicochemical and RNA attributes discussed above. Figure 7 shows the receiver operating characteristic (ROC) curves to visualize the RDF (A,B) and the ANN (C,D) model performances trained on each of the features individually and collectively. ROC curves measure the ability of the model to distinguish between canonical and dark proteins. Figure 7E shows that the % detected transcript (i.e., in what portion of the >6000 RNaseq data sets the transcript was detected) made the most important contribution to the RDF model followed by highest TPM and molecular weight. The overall accuracy of the RDF model when trained on all attributes was slightly better to the ANN model with area under the curve (AUC) values of 0.94 vs 0.91. Both the RDF and ANN models were robust, as their ROC curves did not depend on which subset of the input data was used for training (Figure 7C,D). Supplemental Data Set S9 provides the protein and RNA features (input) for the models as well as the output (probability to be canonical).

Even though the AUCs in the ROC curves were high, there is a substantial number of predicted canonical proteins that were in fact dark proteins and vice versa. To better understand possible reasons for these false predictions (outliers), we assembled two sets of outliers using the combined outcomes of both machine learning models, as follows. For dark protein outliers (predicted to be canonical, but dark), we required that both models calculated a probability (to be canonical) of >0.80; this resulting in 222 outliers. These outlier dark proteins had average physiochemical properties (47 kDa, −0.4 Gravy, 7.3 pI) and moderate average RNA expression values (96%

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

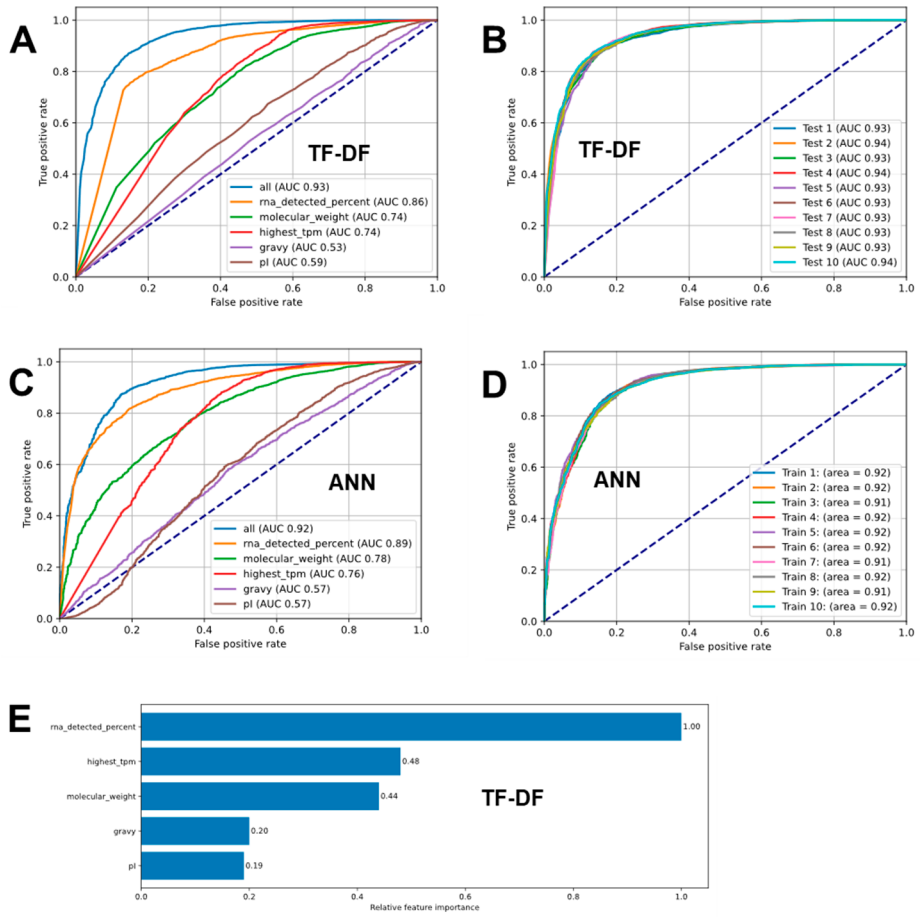


Figure 7. Machine learning models (ANN and TF-DF) to predict the probability of *Arabidopsis* proteins detected at the canonical levels in build 2. (A−D) ROC curves for TF-DF models (A, B) or ANN (C, D) models trained on protein physicochemical properties and RNA expression data. A higher percentage of area under the curve (AUC) signifies better accuracy whereas an AUC of 0.5 (denoted by the dotted navy line) signifies near random prediction. As shown, % RNA detected, molecular_weight, and highest TPM enhance the performance of an ANN model, whereas pI and gravy barely impact it. (B, D) ROC curves of TF-DF (B) and ANN (D) models trained on 10 randomized subsets of the same size from the input data. The accuracy of the TF-DF and ANN models are consistently around 93% and 92%, respectively. (E) Feature importance. The TF-DF model has several built-in methods that calculate the significance of features to a model's performance.

RNA detected, highest TPM 361). Hence, these undetected proteins appeared to have favorable properties (not very low molecular weight, not hydrophobic, not very basic and significant transcript levels and detection across RNA-seq data sets), yet were not detected by MS. For canonical protein outliers (predicted to be dark, but canonical) we required that both models calculated a probability (to be canonical) of <0.20; this resulted in only 42 outliers; these outliers had the average physiochemical properties of 24 kDa MW, −0.3 Gravy, 7.9 pI and low average RNA expression values (19% RNA detected, highest TPM 33). Hence, these unexpected canonical proteins have very low transcript levels and were often not detected in RNA-seq experiments yet were detected at high confidence levels. We then further explore the underlying scenarios for this unexpected behavior based on functional annotations and manual inspection, as described in a section below (Explanations for Unexpected Dark Proteins).

Biological Properties and Functions of the Dark Proteome

Based on the description of the proteins in TAIR and text mining, we observed that proteins annotated as "hypothetical proteins" (some have a Domain of Unknown Function, DUF) were highly overrepresented at 24% of all dark proteins (1349 out of 5595), compared to just 2.6% of the canonical proteins (476 out of 18,079) (Figure 8A). These hypothetical proteins are annotated in TAIR as "protein coding" and not as pseudogenes. On average, the predicted observability to be canonical for these hypothetical proteins was indeed much lower for the dark proteins than the canonical proteins (Figure 8B). Proteins annotated as "unknown" and/or proteins with a DUF' represented 5% of the dark proteins and 4.3% of the canonical proteins (Figure 8A), thus lacking this over-representation in dark proteins.

To take an unbiased approach to determine if the dark proteome is enriched for particular types of proteins, we used the *Arabidopsis* Gene Ontology (GO) enrichment analysis[110−112] for the three GO categories biological process (BP), molecular function (MF), and cellular component (CC). GO analysis was done by comparing all dark proteins to either the sum of canonical and dark proteins or all predicted Araport11 proteins; the results were similar for both comparisons, and we show, therefore, the results of the latter. We did not observe any significant enrichment for the CC categories suggesting that build 2 did not undersample any

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

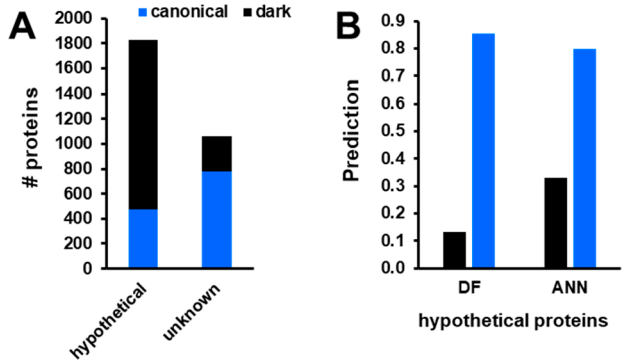


Figure 8. Hypothetical and unknown/DUF proteins in the dark and canonical proteome and their predictions to be canonical. All canonical and unobserved proteins were scored for the presence of the words "hypothetical", "unknown" or "Domain of Unknown Function (DUF)" in their description from Araport11/TAIR. (A) Hypothetical and unknown proteins in the dark and canonical proteome. (B) Predicted observability for the hypothetical proteins to be canonical using the two machine leaning models (DF and ANN).

particular subcellular localization. Indeed, the PXDs that are included in build 2 deliberately include all plant parts and most subcellular fractions such as chloroplasts, mitochondria, etc. However, significant enrichment was observed for BP and MF with the 20 most significant GO terms (lowest FDR) for BP or MF shown in Figure 9A and B. A protein can have several GO terms for each category, and different GO terms can relate to similar processes or functions (Supplemental Data Set S10). There were 520 proteins in the top20 GO terms for BP and 739 proteins for the top20 GO terms for MF, with 271 found in both.

Upon analysis of the enriched BP GO terms (Figure 9A) and the protein IDs, we determined that there are mainly three broad types of protein functions enriched in the dark proteome. These are (i) 149 signaling peptides/peptide hormones such as members of the clavata family, defensins, root meristem growth factor (GO terms: Cell signaling (involved in cell fate commitment), Cell−Cell signaling, Cell fate commitment, Signaling receptor activity, Signaling receptor binding, Regulation of asymmetric cell division, nitrate import, cell killing, killing of other cells of other organisms, phloem development, regulation of cell differentiation), (ii) ∼236 proteins involved in the ubiquitination pathway, including 160 E3 ligases, one E2 conjugating enzyme, 8 ubiquitin(-like) proteins (GO terms: Protein ubiquitination, protein modification by small conjugation (or removal), Ubiquitin(-like) protein ligase activity, Positive regulation of (proteasome) ubiquitin-dependent protein catabolic process), and (iii) ∼130 proteins associated with DNA & RNA related processes ∼130 proteins associated with DNA & RNA related processes (GO terms: RNA/Nucleic acid phosphodiester bond

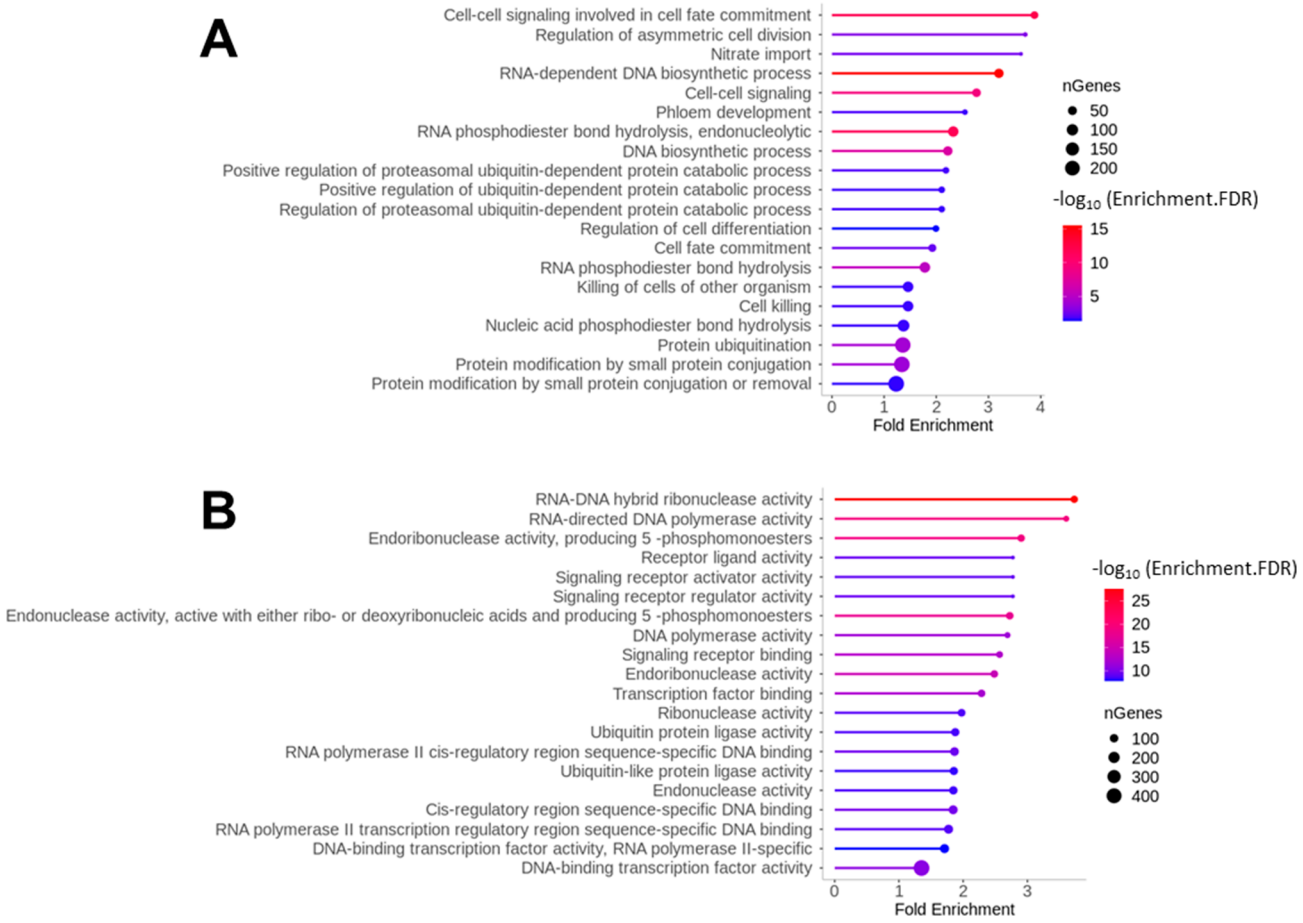
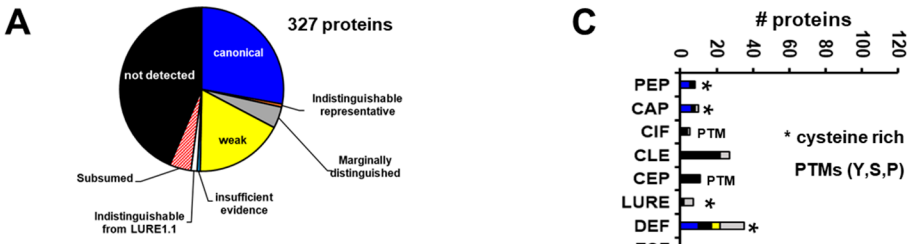


Figure 9. GO enrichment of the 5595 dark proteins compare to all predicted *Arabidopsis* proteins for Biological Process and Molecular function. (A, B) The 20 most significant GO terms (lowest FDR) are shown, ordered by fold enrichment for biological process (A) and molecular function (B).

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214



B

Family	canonical %	dark %	weak %	other %	precursor length (aa)	median precursor length
PEP	63%	38%	0%	0%	81-109	89
CAP	67%	11%	11%	22%	160-210	166
CIF	0%	75%	25%	25%	76-102	83
CLE	0%	81%	4%	19%	74-121	97
CEP	0%	90%	10%	10%	82-229	103
LURE	0%	29%	0%	71%	90-94	91
DEF	33%	23%	17%	43%	60-129	76
ESF	0%	33%	33%	67%	82-90	83
EPF	9%	64%	9%	27%	99-230	120
IDA	0%	83%	0%	17%	77-102	94
LTP	87%	13%	0%	0%	112-126	118
LCR	34%	26%	37%	40%	70-127	79
SCRL	44%	22%	30%	33%	86-109	93
PAMP	0%	100%	0%	0%	72-86	84
PSY	0%	75%	13%	25%	71-94	82
RALF	31%	39%	17%	31%	72-138	90
RTFL/DVL	5%	86%	5%	10%	40-144	53
RGF	0%	92%	0%	8%	79-163	110
EGG CELL	20%	20%	0%	60%	125-158	127
THIONIN	67%	33%	0%	0%	115-134	134
PSK	29%	29%	43%	43%	77-109	87
OTHER	33%	33%	33%	33%	36-81	81

Figure 10. Identification status of members of different signaling peptide families in build 2. (A) Overall identification status across 8 confidence tiers of the 327 signaling peptide producing proteins (Supplemental Data Set S11). The tiers system is described in more detail in ref 8. Identified proteins with status "weak" have at least one uniquely mapping peptide of 9 amino acid residues but do not meet the criteria for canonical (at least 2 uniquely mapping non-nested peptides of at least 9 residues with at least 18 residues of total coverage). (B) Bar diagrams of proteins within each of the peptide signaling families. Color coding using a continuous scale within each column indicates the number of proteins not-observed (black), weak (yellow), canonical (blue), or in other tiers (gray). * indicates cysteine rich peptides. PTMs indicates known presence of PTMs of signaling peptides. (C) Listing all families, identification level, and precursor length (range and median) size mature bioactive peptides.

hydrolysis (endonucleolytic), RNA-dependent DNA biosynthetic process, and DNA biosynthetic process). Many of these proteins belong to superfamilies such as RNA-directed DNA polymerase (reverse transcriptase)-related family (it is not clear what function these have in *Arabidopsis*), non-LTR retroelement reverse transcriptase, reverse transcriptase zinc-binding protein, Polynucleotidyl transferase ribonuclease H-like superfamily, ribonuclease H superfamily polynucleotidyl transferase. Many of these proteins seem to have no defined function.

Analysis of the top 20 enriched MF GO terms (Figure 9B) showed 115 UBI-related proteins and 70 signaling peptides as in the BP GO analysis above. But transcription factor proteins represent by far the most enriched molecular function, with a total of over 400 members of different TF families (e.g., AP2/EREBP, ARF, Auxx/IAA, bHLH, bZIP, C2C2(Zn), C2H2, MADS box, MYB, CCAAT, WRKY) (GO terms: DNA-binding transcription factor activity, Transcription factor binding, RNA polymerase II cis-regulatory region sequence-specific DNA binding, Cis-regulatory region sequence-specific DNA binding, RNA polymerase II transcription regulatory region sequence-specific DNA binding, DNA-binding transcription factor activity, RNA polymerase II-specific, DNA-binding transcription factor activity). The second largest molecular function was for various endonuclease activities with ~83 proteins, including several types of reverse transcriptases and ribonuclease H family members (GO terms: Endonuclease activity, Endonuclease activity, active with either ribo- or DNAs and producing 5-phosphomonoesters, Ribonuclease activity, Endonuclease activity, RNA-DNA hybrid ribonuclease activity). Finally, there were 27 proteins associated with the GO terms RNA-directed DNA polymerase activity and DNA polymerase activity; most of these were also annotated as reverse transcriptases.

Signaling Peptides/Peptide Hormones Are Highly over-Represented in the Dark Proteome

The GO enrichment analysis (Figure 9A,B) suggested that proteins encoding plant signaling peptides or peptide hormones are strongly overrepresented in the dark proteome. Most are inactive precursors (preproproteins of ~7 to ~12 kDa) that undergo a multistep proteolytic processing to result in the relatively small (between ~5 and ~100 amino acids) bioactive peptide signals.[113−116] These small proteins are of great importance in many aspects of plant life. Most of these precursors are secreted through a cleavable N-terminal signal peptide (sP) for targeting into the ER, followed by traveling through the Golgi, plasma membrane, and into the apoplast.

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

Table 6. PeptideAtlas Detection of the ERFVII Transcription Factor Members Involved in Oxygen Sensing

Accession	Name	PA status	# PSMs and plant materials	MW	PI	GRAVY	% RNA detected	Average TPM	Highest TPM	Probability DF	Probability ANN
AT3G16770.1	RAP2.3	Weak	one Phosphopeptide −2 PSMs (cell culture-phospho, callus-phospho)	27.76	5.21	−0.73	99.98	329	7877	0.997	0.939
AT3G14230.1	RAP2.2	Marginally Distinguished AT1G53910.1	two peptides−total 2 PSMs (cell culture)	42.53	4.91	−0.78	100.00	136	1932	1.000	0.969
AT2G47520.1	HRE2	Weak	one peptide −3 PSMs (cell culture, callus)	19.35	6.41	−0.86	82.78	7	1202	0.883	0.793
AT1G72360.1	HRE1	Weak	one peptide −2 PSMs (cell culture, flower)	23.66	4.83	−0.73	99.59	16	1375	0.737	0.928
AT1G53910.1	RAP2.12	Canonical	five peptides−total 5 PSMs (Cell culture)	39.8	5.19	−0.74	100.00	148	1538	0.990	0.972

However, the mode of bioactive peptides can be extracellular or intracellular. We note that there are also bioactive peptides derived from different types of short open reading frames (sORFs, uORFs, lncRNA, pri-miRNA), most of which do not yet have an ATG identifier in the current TAIR annotation.[117,118] Bioactive plant peptides have traditionally been grouped into (i) cysteine-rich peptides that form internal disulfide bonds, and (ii) post-translationally modified small peptides that undergo one or more PTMs during their passage through the ER or Golgi (e.g., tyrosine sulfation,[119] proline hydroxylation, etc.).[114,115]

Many peptide families have been recognized,[113−115,120] including Clavata/embryo-surrounding region (CLE),[121,122] Epidermal Patterning Factor (EPF),[122] phytosulfokine-alpha (PSK),[114] cysteine-rich peptides of the LURE family,[123] Embryo Surrounding Factor (ESF), PAMP-induced secreted peptides (PIP), Plant Peptides containing Tyrosine sulfation family (PSY),[124] root meristem growth factor (RGF), caesarian strip integrity factor (CIF),[125] inflorescence deficient in abscission (IDA), precursor of plant elicitor peptide (PROPEP),[126,127] defensin-like (DFL), and POLARIS which is not part of a larger family. We assembled a tentative list of their protein ATG identifiers (327 genes) to get a better understanding of to what extent they were identified in the new PeptideAtlas build (Supplemental Data Set S11). PeptideAtlas identified 92 (28%) at the canonical level, and 144 (44%) were part of the dark proteome (Figure 10A). The remainder of these 327 proteins were identified at various lower confidence levels often as part of a group of homologues (48 weak, 2 insufficient evidence, 14 subsumed, 14 marginally distinguished, 6 indistinguishable representative) (Figure 10A). The identification level within each family (Figure 10B,C) shows that the majority of members of some families were identified at the canonical level (PEP, CAP, LTP and THIONIN), whereas the identification rate in other families was very low (CIF, CLE, CEP, EPF, IDA, PAMP, PSY, RTFL/DVL, RGF) with >64% members unobserved (dark). The correlation between average (or median) precursor length for each family and identification status is weak. This is logical because these proteins are synthesized as precursors followed by one or more proteolytical cleavages. Furthermore, for family members decorated with PTMs on the amino acid residues Y, S, or P (see Figure 10C) identification rates should be lower since our database search does not include these PTMs because they are relatively rare. Inclusion of such PTMs in regular searches is not appropriate and would result in many false discoveries.

Interestingly, PSMs of the identified proteins ranged from just a few to several thousand for several LTP family members (LPT1,2,3,4) and DEF members (PDF1.1, 1.2A/B/C. 1.3). Sequence coverage was >60% for some 22 preproteins, including several THIONINS, CAPs and a few PEPs; further close inspection of the matched peptides in PeptideAtlas showed that the sequence coverage started downstream of the cleavable signal peptide and mostly or completely included the predicted C-termini. More biological insight into the accumulation and maturation of these signaling peptides can be derived by exploring the associated metadata (stored and linked in PeptideAtlas) and relate that to identification status, protein coverage, and abundance as measured by PSMs in PeptideAtlas. Identifications of the unobserved and low confidence peptides will require targeted experimental approaches, and specific search strategies (e.g., allowing for specific PTMs).

E3 Ligases Are Highly over-Represented in the Dark Proteome

The GO enrichment, and inspection of the associated protein IDs, showed that E3 ligases were over-represented in the dark proteome. *Arabidopsis* has some ∼1400 E3 ligases that each target one or several substrates for polyubiquitination and subsequent degradation by the proteasome. The required amount of an E3 ligase in a cell greatly depends on the number and abundance of its substrates. The dark proteome included 601 E3 ligases (10.7% of the dark proteome), whereas the canonical proteome included 429 E3 ligases (2.4% of the canonicals). Comparing the dark and canonical E3 ligases shows that these 2 groups do not differ in the three physicochemical properties (size, gravy, pI) but that dark proteins have on average much lower transcript detectability and maximum levels.

Proteins with Short Half-Life or Extensive Proteolytic Processing—Protein Features Not Considered in the Machine Learning

There are two protein features (attributes) that were not considered in the machine learning models. These features are (i) proteolytic trimming of the preproteins or (pre)proproteins and (ii) short protein half-life resulting in net low abundance under most conditions. Both scenarios make it harder to detect such proteins by MSMS than predicted by the machine learning models. We already described examples for extensive proteolytic trimming for plant signaling peptides and peptide hormones, which are indeed overrepresented in the dark proteome.

Proteins that are predicted to be canonical but with a conditional short half-life might go undetected (dark proteins)

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

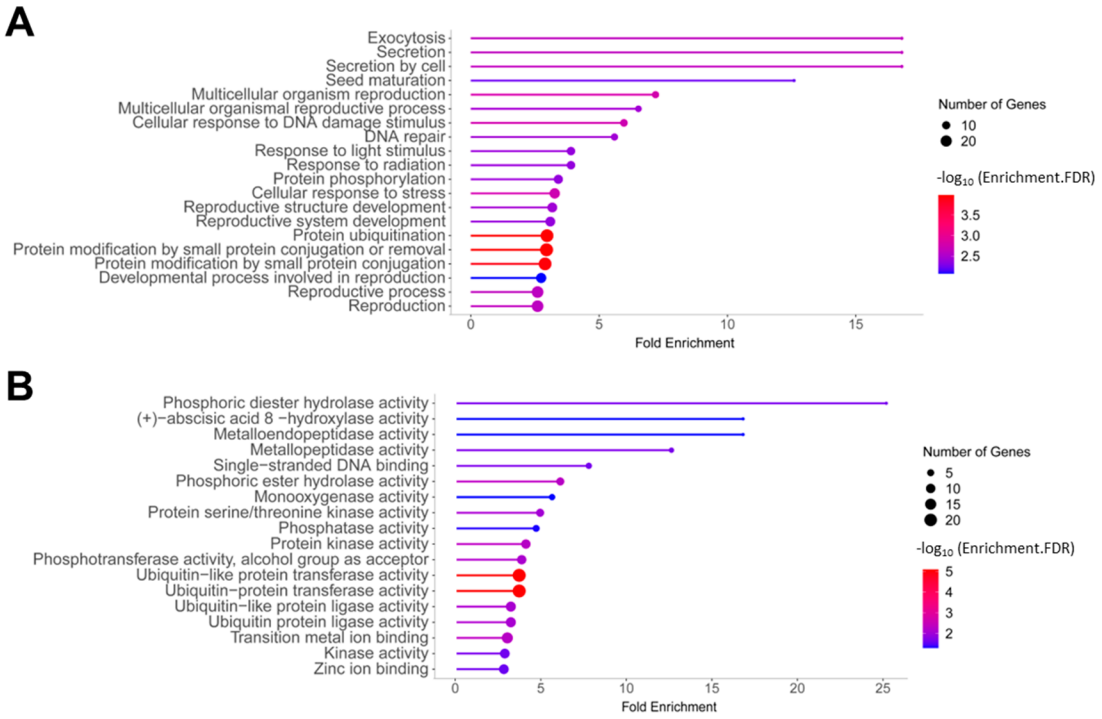


Figure 11. GO enrichment of 222 outlier dark proteins compared to all 5595 dark proteins or Biological Process and Molecular function. The outliers are defined as dark proteins having a predicted probability to be canonical of >0.8 by both machine learning models. (A,B) The 20 most significant GO terms (lowest FDR) are shown, ordered by fold enrichment for biological process (A) and molecular function (B).

or with a very low number of PSMs, because they are continuously degraded under most circumstances. However, the half-lives of most proteins is unknown. One of the exceptions is the set of five transcription factors in the group VII of the Ethylene Response Factor (ERF-VII) family involved in oxygen sensing[128−132] (Table 6). These proteins have a short half-life under normal oxygen concentration (normoxia) because they are degraded by the proteasome through the N-degron pathway but become stabilized during hypoxia or anoxia. These proteins have a cysteine at position second from the N-terminus. After removal of the start methionine by methionine amino peptidases, these N-terminal cysteines are enzymatically oxidized by Plant Cysteine Oxidases (PCDs) which is then followed by enzymatic arginylation (i.e., additional of an arginine residue).[132,133] The arginylated N-terminus is then recognized by specific E3 ligases, resulting in polyubiquitination and degradation by the proteasome. At low cellular oxygen concentrations (hypoxia) due to respiration or environmental conditions (e.g., flooding, high altitude), these transcription factors stabilize because the enzymatic oxidation is slowed down.[134] In *Arabidopsis*, there are five members of this ERF-VII family, i.e., hypoxia response 1 (HRE1; AT1G72360), HRE2 (AT2G47520), related to apetala 2.12 (RAP2.12; AT1G53910), RAP2.2 (AT3G14230), and RAP2.3 (AT3G16770).

Table 6 summarizes the PeptideAtlas findings and protein attributes of this ERF-VII family. Whereas there was MSMS support for all five proteins, the number of PSMs was very low (between 2 and 5). All but one peptide was from callus or cell culture−callus is known to have low internal $[O_2]$[132] explaining why the proteins were observed in callus. It seems quite plausible that plant cell cultures also might experience hypoxia (due to high respiration and low/no photosynthesis). The ERVII TF proteins are predicted to be canonical (predicted observability between 0.7 and 1) (Table 6). However, only RAP2.12 was identified at the canonical level but only in one specific experiment using cell cultures (PXD013868, experiment 8213 https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/ManageTable.cgi?TABLE_NAME=AT_SAMPLE&sample_id=8213). Furthermore, RAP2.3 was only identified with a phosphorylated peptide identified in callus and in cell cultures. Their transcripts were detected in the majority (>82%) of the 5673 RNA-seq data sets and all proteins have very high maximum TPM values (1202−7877). This is a great example where the correlation between predicted probability to be canonical (from the machine learning models) and observed overall number of PSMs suggest unusual properties of the proteins, in this case short half-life. The associated metadata help to provide biological context as the findings for these ERF-VII proteins illustrate.

Explanations for Unexpected Dark Proteins

A small subset of dark proteins (222 out of 5595) were predicted by both machine learning models to be canonical (p > 0.8). To explore biological scenarios for these unexpected dark or unexpected canonical proteins we used both GO enrichment and manual evaluation. We compared GO distributions of the 222 unexpected dark proteins and 5595 dark proteins (Figure 11 and Supplemental Data Set S10). The highest number of unexpected proteins were found for GO terms associated with ubiquitination (Protein ubiquitination, protein modification by small conjugation (or removal), Ubiquitin(-like) protein transferase activity, Ubiquitin(-like) protein ligase activity). Upon further inspection, these were mostly E3 ligases, in particular RING ligases. Other GO terms pointed to enrichment in kinases, terms associated with reproduction, DNA repair, and response to light stimulus or response to radiation, but the genes associated with these GO

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

terms have quite broad range of functions (e.g., transcription factors, some E3 ligases).

Two of the unexpected dark proteins were chloroplast sigma factors 1 and 3 (SIG1 and SIG3; AT1G64860 and AT3G53920) with a predicted probability to be canonical between 0.84 and 0.98. Both are very basic proteins (9.5 and 9.8 pI) with have relatively high molecular weight of the precursors (56 and 65 kDa) and were detected in nearly all 5673 RNA-seq data sets with the highest TPM of 383 and 105; hence it is therefore surprising that they were not detected by MSMS. *Arabidopsis* has six sigma factors (SIG1−6)[135,136] and also SIG4 and SIG5 were unobserved (but with lower probabilities to be canonical than the other sigma factors), whereas SIG2 and SIG6 were canonical. Protein sequence coverage by matched peptides for SIG2 and SIG6 were 45% and 20%, respectively, with 16 and 7 PSMs respectively, showing that also SIG2 and SIG6 are of low general abundance. The most logical explanation is that the half-lives of all sigma factors are relatively short. Chloroplast GUN1 (AT2G31400) is a large PPR protein (100 kDa) is known to have a short half-life of just several minutes because it is degraded by the Clp chaperone-protease system.[137] GUN1 was identified at the canonical level with 12% sequence coverage but only 9 PSMs which is relatively low given its large size and high TPM (596). These examples suggest that many of these unexpected dark proteins have short half-lives or are expressed at high levels but only under undersampled conditions or cell types. These proteins offer an excellent opportunity to learn more about the control of protein half-life and conditional expression patterns.

Lessons from New PXDs in Build 2 That Contribute Most Effectively to Identifying New Canonical Proteins

To inform possible strategies to efficiently identify the remaining 21% of the predicted proteome, we evaluated which of the new PXDs that we selected had the most impact. Figure 12 shows the relation between the number of identified spectra and newly identified canonical proteins (not identified at the canonical level based on earlier data sets) for each of the 63 new PXDs that we added for build 2. Six PXDs that each added the most new canonical proteins are annotated in the figure, together identifying 146 new canonical proteins. Reviewing these new proteins within each of these six PXDs for protein features, including function and molecular weight, identified clear patterns consistent with sample types.

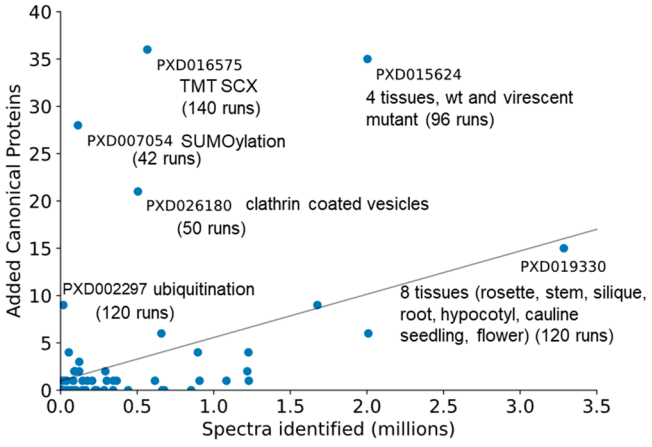


Figure 12. Relation between the number of identified spectra and newly identified canonical proteins for each of the 63 new PXDs that we added for build 2. Key information on the sample type is shown. Newly identified canonical proteins are proteins that were not yet identified as canonicals in build 1 or PXDs in build 2 with a lower number. MS instruments used are PXD016575 − Q Exactive HF-X; PXD007054 − LTQ Orbitrap Velos; PXD026180 − LTQ, Q Exactive HF, Q Exactive and LTQ FT Ultra; PXD015624 − Q Exactive, PXD0119330 - Orbitrap Velos Pro; and PXD0002297 − Q Exactive.

PXD002297 contained 120 MS runs using a Q Exactive instrument from which we matched ∼18,000 MSMS spectra yielding 9 new canonical proteins. This study used COFRADIC technology to map ubiquitination sites reporting 3009 ubiquitination sites in 1,607 proteins.[27] In PXD007054 we identified only 0.11 million MSMS spectra based on 42 MS runs, yet this resulted in 28 new canonical proteins. This study was focused on identification of SUMOylated proteins using a three-step purification protocol based on 6His-tag and anti-SUMO1 antibodies from 8-day old *Arabidopsis* seedlings expressing a 6His-SUMO1(H89-R) transgene in wildtype and SUMO E3 ligase mutants *siz1*−*2* and *mms21*−*1*.[70] Interestingly, the new canonical proteins were highly enriched for transcription factors (17 out of these 28). PXD015624 provided 96 MS runs from which we matched 2 million MSMS spectra resulting in 35 new canonical proteins.[138] The experiments involved label free proteomics of rosettes and roots from 8 week old plants and 2 week old seedlings of wild-type and *nfu2* plants (small and virescent) using a standard workflow (four replicates) involving protein separation by SDS-PAGE (4 slices per lane, tryptic digestion) and an Q Exactive Plus mass spectrometer. More than half of these new canonical proteins were larger proteins over 55 kDa, including five LRR kinases (98−106 kDa) and the glutamate receptor 2.3 (101 kDa). From PXD016575 we identified 0.57 million MSMS spectra and 36 new canonical proteins from 140 MS runs. The experiments involved the analysis of seedlings of wildtype and the autophagy-deficient mutant *atg2*−*2* upon consecutive, temporary reprogramming inducing stimuli ABA and flg2.[139] The proteomics workflow involved SDS extracted total seedling proteomes, TMT labeling followed by SCX chromatography, and standard nanoLC-MSMS using a Q Exactive instrument. The new canonical proteins from this set included 19 proteins below 20 kDa, including several RALF signaling peptides; these small proteins are often missed in SDS-PAGE separated samples. PXD019330 was a truly large-scale proteomics study sampling multiple tissue types (roots, leaves, cauline leaves, stems, flowers, siliques/seeds, whole plant seedlings) at different developmental stages.[140] A standard workflow was used involving protein separation by SDS-PAGE (5 slices per lane, tryptic digestion) and an LTQ-Orbitrap Velos instrument and notably a long C18 column (50 cm) and long (9 h) elution with a total of 120 MS runs. We matched 3.29 million MSMS spectra, resulting in just 15 new canonical proteins. These new canonicals included several chloroplast membrane proteins (FAX4 and Lil1.2), a nitrate transporter, and two very small metallothioneins. PXD026180 contained 50 MS runs from four different MS instruments (LTQ, Q Exactive HF, Q Exactive, LTQ FT Ultra) from which we mapped 0.5 million MSMS spectra, yielding 21 new canonical proteins. This study analyzed purified clathrin coated vesicles (CCVs) from undifferentiated *Arabidopsis* suspension cultured cells using both SDS-PAGE and in-solution digests, followed by nanoLC-MSMS on the extracted peptides.[69] These six PXDs utilize a wide range of methods and plant materials,

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

some highly affinity enriched (SUMOylation, ubiquitination, CCV) and others including a range of different plant parts.

As this snapshot of six PXDs illustrates, the proteomics-MS workflows showed a wide range of techniques (e.g., from SDS-PAGE with in-gel digests, to in-solution digest, TMT labeling, and SXC chromatography) in all cases followed by reverse-phase nanoLC-MSMS but with different generations of MS instruments. Considering the total number of matched MSMS spectra, those PXDs that used affinity enrichment based on specific PTMs or isolation of highly specialized subcellular structures, clearly identified the most new canonical proteins when normalized to the number of matched spectra. This suggests that the identification of the remaining 21% of the predicted *Arabidopsis* proteome will be most effective when this will also include targeting specific subcellular structures and specific PTMs.

CONCLUSIONS AND FUTURE PERSPECTIVE

This second release of the *Arabidopsis* PeptideAtlas is based on ∼259 million searched raw MSMS spectra from 115 PXDs and includes 21,017 protein identifications based on ∼70 million matched spectra (PSMs) and nearly 0.6 million distinct matched peptides. Compared to the first release[8] this represents an increase of 78% more PSMs, 11% more distinct peptides, 1.2% more proteins and an increase from 49.5% to 51.6% in global proteome sequence coverage. Furthermore, this new PeptideAtlas release includes 5198 phosphorylated proteins, 668 ubiquitinated proteins, 3050 N-terminally acetylated proteins, and 864 lysine-acetylated proteins. The majority of predicted *Arabidopsis* proteins have now been identified by MS, and users can explore the PeptideAtlas to readily determine if their proteins of interest have been identified, in which type of tissues or samples, obtain a sense of abundance, and evaluate if these proteins undergo any of the known major PTMs (phosphorylation, N-terminal or lysine acetylation, ubiquitination). Through GO enrichment analysis, machine learning, meta-data curation and analysis, as well as manual evaluation, we identified multiple reasons why proteins have not yet been identified in this new PeptideAtlas build. These reasons include (i) small size (either because the gene encodes for a small protein or due to extensive proteolytic processing as in the case of signaling peptides), (ii) high hydrophobicity, (iii) very high pI, (iv) low steady state abundance due to low gene expression or short protein half-life), (v) unusual PTMs, or (vi) only presence in very specific conditions or cell types that were not included in the selected PXDs. All but seven of the PXDs used trypsin as the enzyme to convert proteins to peptides prior to MSMS analysis. Trypsin by far is the most used enzyme because it is efficient and reliable and produces peptides with a positive C-terminal amino acid (R or K) which helps generation of good y and b ions in positive ionization mode. However, there are demonstrated benefits to complement the MSMS peptide analysis of tryptic digests with digests by proteases (e.g., GluC, AspN).[141−144] The main benefit of multiprotease digests is increased protein sequence coverage, whereas the increase in newly identified proteins (compared to just trypsin) tends to be quite limited or even incremental.[143] For very hydrophobic proteins nonenzymatic digestion by limited acid hydrolysis or cyanogen bromide (cleaving C-terminal of methionine) can be helpful as we demonstrated for small and hydrophobic *Arabidopsis* thylakoid proteins.[145]

The challenge now is to identify the remaining 20% of the predicted *Arabidopsis* proteome. Furthermore, this new build also mapped peptides to an additional ∼80 proteins not represented in the current *Arabidopsis* genome. These additional proteins should be considered in the community effort led by to Tanya Berardini at TAIR to generate a new annotation for Col-0..

This PeptideAtlas was built using about ∼20% of the currently (July 2022) available PXDs for *Arabidopsis*; incorporation of the vast majority of the unused PXDs is likely to only marginally increase the number of identified proteins, as inferred from our comparison between build 1 and build 2. It is also not feasible to incorporate all of these available raw data given the necessary time and expertise required. Furthermore, in case of several older PXDs in ProteomeXchange, low resolution instruments (e.g., LCQs or LTQs) instruments were used; data from such PXDs are unlikely to contribute much to the PeptideAtlas. (We note that even older data sets from 2005−2012 originally submitted to PRIDE are not available in ProteomeXchange).

To increase the number of protein identifications in PeptideAtlas, a strategic approach will be needed, by very carefully selecting data sets with the most sophisticated workflows (including selective enrichment for PTMs, multi-protease digestions) and acquisition using the very latest generation of MS instruments (high mass accuracy, sensitivity and high dynamic range, very fast acquisition rates). Finally, a targeted approach to identify the missing (dark) proteome might be most effective using the combined insights from the machine learning models and the predicted protein properties and large-scale RNA-seq analysis across cell and tissue types as well as developmental stages and biotic and abiotic conditions. The plant community can take inspiration from sustained efforts to map the human proteome in PeptideAtlas with 85.9% of the predicted proteins now identified at the highest confidence level[146] (https://peptideatlas.org/builds/human/). We do note that the first human PeptideAtlas was released in 2012 with annual updates gradually increasing the % of predicted proteins identified at the canonical level from 80.0% in 2018, 82.5% in 2019, 84.3% in 2020, 84.6% in 2021, and 85.9% in 2022. We will use the experience, strategies, and insights for the human PeptideAtlas project, combined with feedback PXD submissions by the *Arabidopsis* research community to increase proteome detection for *Arabidopsis*.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.3c00536.

SI Table of Contents (PDF)

Supplemental Figure S1: Correlation between the number of spectra search and identified spectra number of MS runs, identified peptides, or identified canonical proteins based on values from Table 2; Supplemental Figure S2: Correlation between the number of spectra identified and identified peptides or identified canonical proteins based on values from Table 2 (PDF)

Supplemental Data Set S1: Comprehensive overview of the 115 PXDs and their 369 experiments used for build 2 (XLSX)

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

Supplemental Data Set S2: Transcript per million (TPM) expression values of nuclear protein coding genes in Araport11 and RNA-seq data sets (XLSX)

Supplemental Data Set S3: Proteins identified in non-Araport11 sources by hierarchy of sources (XLSX)

Supplemental Data Set S4: Identification of N-terminal acetylation (NTA) sites in canonical proteins in PeptideAtlas (XLSX)

Supplemental Data Set S5: Identification of lysine acetylation (Kac) sites in canonical proteins in PeptideAtlas (XLSX)

Supplemental Data Set S6: Identification of phosphorylation (S,T,Y) sites in canonical proteins in PeptideAtlas (XLSX)

Supplemental Data Set S7: Identification of Ubiquitination sites in canonical proteins in PeptideAtlas (XLSX)

Supplemental Data Set S8: Combined PTM results for the canonical proteins in PeptideAtlas with identified PTM sites for N-terminal acetylation, lysine acetylation, phosphorylation, and/or ubiquitination (XLSX)

Supplemental Data Set S9: Nuclear-encoded proteins Araport11 identifiers (26,977) with annotations, protein properties, RNA-seq-based transcript information, machine learning predicted probability to be canonical, and identification status in PeptideAtlas (XLSX)

Supplemental Data Set S10: GO enrichment results of dark proteins (XLSX)

Supplemental Data Set S11: Proteins coding for signaling peptides, their annotations, physicochemical properties, RNA expression patterns, and identification status in PeptideAtlas (XLSX)

AUTHOR INFORMATION

Corresponding Authors

Klaas J. van Wijk − *Section of Plant Biology, School of Integrative Plant Sciences (SIPS), Cornell University, Ithaca, New York 14853, United States;* orcid.org/0000-0001-9536-0487; Email: kv35@cornell.edu

Eric W. Deutsch − *Institute for Systems Biology (ISB), Seattle, Washington 98109, United States;* orcid.org/0000-0001-8732-0928; Email: edeutsch@systemsbiology.org

Authors

Tami Leppert − *Institute for Systems Biology (ISB), Seattle, Washington 98109, United States*

Zhi Sun − *Institute for Systems Biology (ISB), Seattle, Washington 98109, United States;* orcid.org/0000-0003-3324-6851

Alyssa Kearly − *Boyce Thompson Institute, Ithaca, New York 14853, United States*

Margaret Li − *Institute for Systems Biology (ISB), Seattle, Washington 98109, United States*

Luis Mendoza − *Institute for Systems Biology (ISB), Seattle, Washington 98109, United States;* orcid.org/0000-0003-0128-8643

Isabell Guzchenko − *Section of Plant Biology, School of Integrative Plant Sciences (SIPS), Cornell University, Ithaca, New York 14853, United States*

Erica Debley − *Section of Plant Biology, School of Integrative Plant Sciences (SIPS), Cornell University, Ithaca, New York 14853, United States*

Georgia Sauermann − *Section of Plant Biology, School of Integrative Plant Sciences (SIPS), Cornell University, Ithaca, New York 14853, United States*

Pratyush Routray − *Section of Plant Biology, School of Integrative Plant Sciences (SIPS), Cornell University, Ithaca, New York 14853, United States*

Sagunya Malhotra − *Institute for Systems Biology (ISB), Seattle, Washington 98109, United States*

Andrew Nelson − *Boyce Thompson Institute, Ithaca, New York 14853, United States*

Qi Sun − *Computational Biology Service Unit, Cornell University, Ithaca, New York 14853, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.3c00536

Author Contributions

TL and ZS carried out the MS searches and PeptideAtlas data loading, supervised by EWD, and assembled the search results. ML developed the machine learning code. AK and AN assembled and analyzed the RNA-seq data. ML and SM contributed to data analysis and created figures. LM and ZS developed the PeptideAtlas web interface. IG, ED, GS, and PR helped annotated the metadata in PeptideAtlas. QS helped assemble the protein search space. EWD and KJVW developed, coordinated, and oversaw the project, evaluated outcomes, and wrote the paper.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was primarily funded by the National Science Foundation IOS-1922871 (KJVW and ED) and in part by DBI-1933311 (EWD), MCB-2120131 (ADLN), IOS-2023310 (ADLN), and by the National Institutes of Health grant R01 GM087221 (EWD).

REFERENCES

(1) Koornneef, M.; Meinke, D. The development of Arabidopsis as a model plant. *Plant J.* **2010**, *61* (6), 909−21.

(2) Meinke, D. W.; Cherry, J. M.; Dean, C.; Rounsley, S. D.; Koornneef, M. Arabidopsis thaliana: a model plant for genome analysis. *Science* **1998**, *282* (5389), 662.

(3) Somerville, C. R.; Ogren, W. L. Inhibition of photosynthesis in Arabidopsis mutants lacking leaf glutamate synthase activity. *Nature* **1980**, *286*, 257−259.

(4) Somerville, C. R.; Ogren, W. L. Mutants of the cruciferous plant Arabidopsis thaliana lacking glycine decarboxylase activity. *Biochem. J.* **1982**, *202* (2), 373−80.

(5) Provart, N. J.; Brady, S. M.; Parry, G.; Schmitz, R. J.; Queitsch, C.; Bonetta, D.; Waese, J.; Schneeberger, K.; Loraine, A. E. Anno genominis XX: 20 years of Arabidopsis genomics. *Plant Cell* **2021**, *33* (4), 832−845.

(6) Parry, G.; Provart, N. J.; Brady, S. M.; Uzilday, B. Current status of the multinational Arabidopsis community. *Plant Direct* **2020**, *4* (7), No. e00248.

(7) Alex Mason, G.; Canto-Pastor, A.; Brady, S. M.; Provart, N. J. Bioinformatic Tools in Arabidopsis Research. *Methods Mol. Biol.* **2021**, *2200*, 25−89.

(8) van Wijk, K. J.; Leppert, T.; Sun, Q.; Boguraev, S. S.; Sun, Z.; Mendoza, L.; Deutsch, E. W. The Arabidopsis PeptideAtlas: Harnessing worldwide proteomics data to create a comprehensive community proteomics resource. *Plant Cell* **2021**, *33* (11), 3421−3453.

(9) San Clemente, H.; Jamet, E. WallProtDB, a database resource for plant cell wall proteomics. *Plant Methods* **2015**, *11* (1), 2.
(10) Salvi, D.; Bournais, S.; Moyet, L.; Bouchnak, I.; Kuntz, M.; Bruley, C.; Rolland, N. AT_CHLORO: The First Step When Looking for Information About Subplastidial Localization of Proteins. *Methods Mol. Biol.* **2018**, *1829*, 395−406.
(11) Sun, Q.; Zybailov, B.; Majeran, W.; Friso, G.; Olinares, P. D.; van Wijk, K. J. PPDB, the Plant Proteomics Database at Cornell. *Nucleic Acids Res.* **2009**, *37*, D969−74.
(12) Tanz, S. K.; Castleden, I.; Hooper, C. M.; Vacher, M.; Small, I.; Millar, H. A. SUBA3: a database for integrating experimentation and prediction to define the SUBcellular location of proteins in Arabidopsis. *Nucleic Acids Res.* **2012**, *41*, D1185−91.
(13) Schulze, W. X.; Yao, Q.; Xu, D. Databases for plant phosphoproteomics. *Methods Mol. Biol.* **2015**, *1306*, 207−16.
(14) Willems, P.; Horne, A.; Van Parys, T.; Goormachtig, S.; De Smet, I.; Botzki, A.; Van Breusegem, F.; Gevaert, K. The Plant PTM Viewer, a central resource for exploring plant protein modifications. *Plant J.* **2019**, *99* (4), 752−762.
(15) Omenn, G. S.; Lane, L.; Overall, C. M.; Paik, Y. K.; Cristea, I. M.; Corrales, F. J.; Lindskog, C.; Weintraub, S.; Roehrl, M. H. A.; Liu, S.; Bandeira, N.; Srivastava, S.; Chen, Y. J.; Aebersold, R.; Moritz, R. L.; Deutsch, E. W. Progress Identifying and Analyzing the Human Proteome: 2021 Metrics from the HUPO Human Proteome Project. *J. Proteome Res.* **2021**, *20* (12), 5227−5240.
(16) Hesselager, M. O.; Codrea, M. C.; Sun, Z.; Deutsch, E. W.; Bennike, T. B.; Stensballe, A.; Bundgaard, L.; Moritz, R. L.; Bendixen, E. The Pig PeptideAtlas: A resource for systems biology in animal production and biomedicine. *Proteomics* **2016**, *16* (4), 634−44.
(17) McCord, J.; Sun, Z.; Deutsch, E. W.; Moritz, R. L.; Muddiman, D. C. The PeptideAtlas of the Domestic Laying Hen. *J. Proteome Res.* **2017**, *16* (3), 1352−1363.
(18) Nissa, M. U.; Reddy, P. J.; Pinto, N.; Sun, Z.; Ghosh, B.; Moritz, R. L.; Goswami, M.; Srivastava, S. The PeptideAtlas of a widely cultivated fish Labeo rohita: A resource for the Aquaculture Community. *Sci. Data* **2022**, *9* (1), 171.
(19) King, N. L.; Deutsch, E. W.; Ranish, J. A.; Nesvizhskii, A. I.; Eddes, J. S.; Mallick, P.; Eng, J.; Desiere, F.; Flory, M.; Martin, D. B.; Kim, B.; Lee, H.; Raught, B.; Aebersold, R. Analysis of the Saccharomyces cerevisiae proteome with PeptideAtlas. *Genome Biol.* **2006**, *7* (11), R106.
(20) Gunaratne, J.; Schmidt, A.; Quandt, A.; Neo, S. P.; Sarac, O. S.; Gracia, T.; Loguercio, S.; Ahrne, E.; Xia, R. L.; Tan, K. H.; Lossner, C.; Bahler, J.; Beyer, A.; Blackstock, W.; Aebersold, R. Extensive mass spectrometry-based analysis of the fission yeast proteome: the Schizosaccharomyces pombe PeptideAtlas. *Mol. Cell Proteomics* **2013**, *12* (6), 1741−51.
(21) Michalik, S.; Depke, M.; Murr, A.; Gesell Salazar, M.; Kusebauch, U.; Sun, Z.; Meyer, T. C.; Surmann, K.; Pfortner, H.; Hildebrandt, P.; Weiss, S.; Palma Medina, L. M.; Gutjahr, M.; Hammer, E.; Becher, D.; Pribyl, T.; Hammerschmidt, S.; Deutsch, E. W.; Bader, S. L.; Hecker, M.; Moritz, R. L.; Mader, U.; Volker, U.; Schmidt, F. A global Staphylococcus aureus proteome resource applied to the in vivo characterization of host-pathogen interactions. *Sci. Rep* **2017**, *7* (1), 9718.
(22) Malmstrom, J.; Beck, M.; Schmidt, A.; Lange, V.; Deutsch, E. W.; Aebersold, R. Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans. *Nature* **2009**, *460* (7256), 762−5.
(23) Reales-Calderon, J. A.; Sun, Z.; Mascaraque, V.; Perez-Navarro, E.; Vialas, V.; Deutsch, E. W.; Moritz, R. L.; Gil, C.; Martinez, J. L.; Molero, G. A wide-ranging Pseudomonas aeruginosa PeptideAtlas build: A useful proteomic resource for a versatile pathogen. *J. Proteomics* **2021**, *239*, 104192.
(24) Deutsch, E. W.; Bandeira, N.; Perez-Riverol, Y.; Sharma, V.; Carver, J. J.; Mendoza, L.; Kundu, D. J.; Wang, S.; Bandla, C.; Kamatchinathan, S.; Hewapathirana, S.; Pullman, B. S.; Wertz, J.; Sun, Z.; Kawano, S.; Okuda, S.; Watanabe, Y.; MacLean, B.; MacCoss, M. J.; Zhu, Y.; Ishihama, Y.; Vizcaino, J. A. The ProteomeXchange consortium at 10 years: 2023 update. *Nucleic Acids Res.* **2023**, *51* (D1), D1539−D1548.
(25) Plant Cell Atlas, C.; Jha, S. G.; Borowsky, A. T.; Cole, B. J.; Fahlgren, N.; Farmer, A.; Huang, S. C.; Karia, P.; Libault, M.; Provart, N. J.; Rice, S. L.; Saura-Sanchez, M.; Agarwal, P.; Ahkami, A. H.; Anderton, C. R.; Briggs, S. P.; Brophy, J. A.; Denolf, P.; Di Costanzo, L. F.; Exposito-Alonso, M.; Giacomello, S.; Gomez-Cano, F.; Kaufmann, K.; Ko, D. K.; Kumar, S.; Malkovskiy, A. V.; Nakayama, N.; Obata, T.; Otegui, M. S.; Palfalvi, G.; Quezada-Rodriguez, E. H.; Singh, R.; Uhrig, R. G.; Waese, J.; Van Wijk, K.; Wright, R. C.; Ehrhardt, D. W.; Birnbaum, K. D.; Rhee, S. Y. Vision, challenges and opportunities for a Plant Cell Atlas. *Elife* **2021**, *10*, e66877.
(26) Mergner, J.; Frejno, M.; List, M.; Papacek, M.; Chen, X.; Chaudhary, A.; Samaras, P.; Richter, S.; Shikata, H.; Messerer, M.; Lang, D.; Altmann, S.; Cyprys, P.; Zolg, D. P.; Mathieson, T.; Bantscheff, M.; Hazarika, R. R.; Schmidt, T.; Dawid, C.; Dunkel, A.; Hofmann, T.; Sprunck, S.; Falter-Braun, P.; Johannes, F.; Mayer, K. F. X.; Jurgens, G.; Wilhelm, M.; Baumbach, J.; Grill, E.; Schneitz, K.; Schwechheimer, C.; Kuster, B. Mass-spectrometry-based draft of the Arabidopsis proteome. *Nature* **2020**, *579* (7799), 409−414.
(27) Walton, A.; Stes, E.; Cybulski, N.; Van Bel, M.; Inigo, S.; Durand, A. N.; Timmerman, E.; Heyman, J.; Pauwels, L.; De Veylder, L.; Goossens, A.; De Smet, I.; Coppens, F.; Goormachtig, S.; Gevaert, K. It's Time for Some "Site"-Seeing: Novel Tools to Monitor the Ubiquitin Landscape in Arabidopsis thaliana. *Plant Cell* **2016**, *28* (1), 6−16.
(28) Grubb, L. E.; Derbyshire, P.; Dunning, K. E.; Zipfel, C.; Menke, F. L. H.; Monaghan, J. Large-scale identification of ubiquitination sites on membrane-associated proteins in Arabidopsis thaliana seedlings. *Plant Physiol* **2021**, *185* (4), 1483−1488.
(29) Perdigao, N.; Heinrich, J.; Stolte, C.; Sabir, K. S.; Buckley, M. J.; Tabor, B.; Signal, B.; Gloss, B. S.; Hammang, C. J.; Rost, B.; Schafferhans, A.; O'Donoghue, S. I. Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (52), 15898−903.
(30) Wright, B. W.; Yi, Z.; Weissman, J. S.; Chen, J. The dark proteome: translation from noncanonical open reading frames. *Trends Cell Biol.* **2022**, *32* (3), 243−258.
(31) Skinner, O. S.; Kelleher, N. L. Illuminating the dark matter of shotgun proteomics. *Nat. Biotechnol.* **2015**, *33* (7), 717−8.
(32) Birnbaum, K. D.; Otegui, M. S.; Bailey-Serres, J.; Rhee, S. Y. The Plant Cell Atlas: focusing new technologies on the kingdom that nourishes the planet. *Plant Physiol* **2022**, *188* (2), 675−679.
(33) Hooper, C. M.; Castleden, I. R.; Tanz, S. K.; Aryamanesh, N.; Millar, A. H. SUBA4: the interactive data analysis centre for Arabidopsis subcellular protein locations. *Nucleic Acids Res.* **2017**, *45* (D1), D1064−D1074.
(34) UniProt, C. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D523−D531.
(35) Durek, P.; Schmidt, R.; Heazlewood, J. L.; Jones, A.; MacLean, D.; Nagel, A.; Kersten, B.; Schulze, W. X. PhosPhAt: the Arabidopsis thaliana phosphorylation site database. An update. *Nucleic Acids Res.* **2010**, *38*, D828−34.
(36) Willems, P. Exploring Posttranslational Modifications with the Plant PTM Viewer. *Methods Mol. Biol.* **2022**, *2447*, 285−296.
(37) Cheng, C. Y.; Krishnakumar, V.; Chan, A. P.; Thibaud-Nissen, F.; Schobel, S.; Town, C. D. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *Plant J.* **2017**, *89* (4), 789−804.
(38) Lamesch, P.; Berardini, T. Z.; Li, D.; Swarbreck, D.; Wilks, C.; Sasidharan, R.; Muller, R.; Dreher, K.; Alexander, D. L.; Garcia-Hernandez, M.; Karthikeyan, A. S.; Lee, C. H.; Nelson, W. D.; Ploetz, L.; Singh, S.; Wensel, A.; Huala, E. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **2012**, *40*, D1202−10.
(39) UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D480−D489.
(40) 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.; Gonzales, N. R.; Gwadz, M.; Lanczycki, C. J.; Song, J.

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

S.; Thanki, N.; Wang, J.; Yamashita, R. A.; Yang, M.; Zheng, C.; Marchler-Bauer, A.; Thibaud-Nissen, F. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res.* **2021**, *49* (D1), 1020−1028.
(41) Hazarika, R. R.; De Coninck, B.; Yamamoto, L. R.; Martin, L. R.; Cammue, B. P.; van Noort, V. ARA-PEPs: a repository of putative sORF-encoded peptides in Arabidopsis thaliana. *BMC Bioinformatics* **2017**, *18* (1), 37.
(42) Sloan, D. B.; Wu, Z.; Sharbrough, J. Correction of Persistent Errors in Arabidopsis Reference Mitochondrial Genomes. *Plant Cell* **2018**, *30* (3), 525−527.
(43) van Wijk, K. J.; Bentolila, S.; Leppert, T.; Sun, Q.; Sun, Z.; Mendoza, L.; Li, M.; Deutsch, E. W. Detection and editing of the updated plastid- and mitochondrial-encoded proteomes for Arabidopsis with PeptideAtlas. *Plant Physiology* **2023**, kiad572.
(44) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Rompp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P. A.; Deutsch, E. W. mzML-a community standard for mass spectrometry data. *Mol. Cell Proteomics* **2011**, *10* (1), R110 000133.
(45) Hulstaert, N.; Shofstahl, J.; Sachsenberg, T.; Walzer, M.; Barsnes, H.; Martens, L.; Perez-Riverol, Y. ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. *J. Proteome Res.* **2020**, *19* (1), 537−542.
(46) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012**, *30* (10), 918−20.
(47) Keller, A.; Eng, J.; Zhang, N.; Li, X. J.; Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **2005**, *1*, 2005.0017.
(48) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Slagel, J.; Sun, Z.; Moritz, R. L. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clin Appl.* **2015**, *9* (7−8), 745−54.
(49) Deutsch, E. W.; Mendoza, L.; Shteynberg, D. D.; Hoopmann, M. R.; Sun, Z.; Eng, J. K.; Moritz, R. L. Trans-Proteomic Pipeline: Robust Mass Spectrometry-Based Proteomics Data Analysis Suite. *J. Proteome Res.* **2023**, *22* (2), 615−24.
(50) Eng, J. K.; Deutsch, E. W. Extending Comet for Global Amino Acid Variant and Post-Translational Modification Analysis Using the PSI Extended FASTA Format. *Proteomics* **2020**, *20*, No. e1900362.
(51) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **2017**, *14* (5), 513−520.
(52) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74* (20), 5383−92.
(53) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell Proteomics* **2011**, *10* (12), M111.007690.
(54) Verrastro, I.; Pasha, S.; Jensen, K. T.; Pitt, A. R.; Spickett, C. M. Mass spectrometry-based methods for identifying oxidized proteins in disease: advances and challenges. *Biomolecules* **2015**, *5* (2), 378−411.
(55) Hawkins, C. L.; Davies, M. J. Detection, identification, and quantification of oxidative protein modifications. *J. Biol. Chem.* **2019**, *294* (51), 19683−19708.
(56) 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.; Hancock, W. S.; Hermjakob, H.; Aebersold, R.; Moritz, R. L.; Omenn, G. S. Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. *J. Proteome Res.* **2016**, *15* (11), 3961−3970.
(57) Palos, K.; Nelson Dittrich, A. C.; Yu, L.; Brock, J. R.; Railey, C. E.; Wu, H. L.; Sokolowska, E.; Skirycz, A.; Hsu, P. Y.; Gregory, B. D.; Lyons, E.; Beilstein, M. A.; Nelson, A. D. L. Identification and functional annotation of long intergenic non-coding RNAs in Brassicaceae. *Plant Cell* **2022**, *34* (9), 3233−3260.
(58) Liao, Y.; Smyth, G. K.; Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **2014**, *30* (7), 923−30.
(59) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Perez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yilmaz, S.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaino, J. A. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **2019**, *47* (D1), D442−D450.
(60) Perez-Riverol, Y.; Bai, J.; Bandla, C.; Garcia-Seisdedos, D.; Hewapathirana, S.; Kamatchinathan, S.; Kundu, D. J.; Prakash, A.; Frericks-Zipper, A.; Eisenacher, M.; Walzer, M.; Wang, S.; Brazma, A.; Vizcaino, J. A. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **2022**, *50* (D1), D543−D552.
(61) Pullman, B. S.; Wertz, J.; Carver, J.; Bandeira, N. ProteinExplorer: A Repository-Scale Resource for Exploration of Protein Detection in Public Mass Spectrometry Data Sets. *J. Proteome Res.* **2018**, *17* (12), 4227−4234.
(62) Moriya, Y.; Kawano, S.; Okuda, S.; Watanabe, Y.; Matsumoto, M.; Takami, T.; Kobayashi, D.; Yamanouchi, Y.; Araki, N.; Yoshizawa, A. C.; Tabata, T.; Iwasaki, M.; Sugiyama, N.; Tanaka, S.; Goto, S.; Ishihama, Y. The jPOST environment: an integrated proteomics data repository and database. *Nucleic Acids Res.* **2019**, *47* (D1), D1218−D1224.
(63) Ma, J.; Chen, T.; Wu, S.; Yang, C.; Bai, M.; Shu, K.; Li, K.; Zhang, G.; Jin, Z.; He, F.; Hermjakob, H.; Zhu, Y. iProX: an integrated proteome resource. *Nucleic Acids Res.* **2019**, *47* (D1), D1211−D1217.
(64) Sharma, V.; Eckels, J.; Schilling, B.; Ludwig, C.; Jaffe, J. D.; MacCoss, M. J.; MacLean, B. Panorama Public: A Public Repository for Quantitative Data Sets Processed in Skyline. *Mol. Cell Proteomics* **2018**, *17* (6), 1239−1244.
(65) Makarov, A. Orbitrap journey: taming the ion rings. *Nat. Commun.* **2019**, *10* (1), 3743.
(66) Nolting, D.; Malek, R.; Makarov, A. Ion traps in modern mass spectrometry. *Mass Spectrom Rev.* **2019**, *38* (2), 150−168.
(67) Waltz, F.; Nguyen, T. T.; Arrive, M.; Bochler, A.; Chicher, J.; Hammann, P.; Kuhn, L.; Quadrado, M.; Mireau, H.; Hashem, Y.; Giege, P. Small is big in Arabidopsis mitochondrial ribosome. *Nat. Plants* **2019**, *5* (1), 106−117.
(68) Huang, A.; Tang, Y.; Shi, X.; Jia, M.; Zhu, J.; Yan, X.; Chen, H.; Gu, Y. Proximity labeling proteomics reveals critical regulators for inner nuclear membrane protein degradation in plants. *Nat. Commun.* **2020**, *11* (1), 3284.
(69) Dahhan, D. A.; Reynolds, G. D.; Cardenas, J. J.; Eeckhout, D.; Johnson, A.; Yperman, K.; Kaufmann, W. A.; Vang, N.; Yan, X.; Hwang, I.; Heese, A.; De Jaeger, G.; Friml, J.; Van Damme, D.; Pan, J.; Bednarek, S. Y. Proteomic characterization of isolated Arabidopsis clathrin-coated vesicles reveals evolutionarily conserved and plant-specific components. *Plant Cell* **2022**, *34* (6), 2150−2173.
(70) Rytz, T. C.; Miller, M. J.; McLoughlin, F.; Augustine, R. C.; Marshall, R. S.; Juan, Y. T.; Charng, Y. Y.; Scalf, M.; Smith, L. M.; Vierstra, R. D. SUMOylome Profiling Reveals a Diverse Array of Nuclear Targets Modified by the SUMO Ligase SIZ1 during Heat Stress. *Plant Cell* **2018**, *30* (5), 1077−1099.

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

(71) 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.; Reiter, L.; Rost, H. L.; Tate, S.; Ting, Y. S.; Collins, B. C.; Aebersold, R. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat. Methods* **2017**, *14* (9), 921−927.
(72) Yu, F.; Teo, G. C.; Kong, A. T.; Frohlich, K.; Li, G. X.; Demichev, V.; Nesvizhskii, A. I. Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *Nat. Commun.* **2023**, *14* (1), 4154.
(73) Rauniyar, N.; Yates, J. R., 3rd Isobaric labeling-based relative quantification in shotgun proteomics. *J. Proteome Res.* **2014**, *13* (12), 5293−309.
(74) Chen, X.; Sun, Y.; Zhang, T.; Shu, L.; Roepstorff, P.; Yang, F. Quantitative Proteomics Using Isobaric Labeling: A Practical Guide. *Genomics Proteomics Bioinformatics* **2021**, *19* (5), 689−706.
(75) Hsu, J. L.; Huang, S. Y.; Chow, N. H.; Chen, S. H. Stable-isotope dimethyl labeling for quantitative proteomics. *Anal. Chem.* **2003**, *75* (24), 6843−52.
(76) Kleifeld, O.; Doucet, A.; Prudova, A.; auf dem Keller, U.; Gioia, M.; Kizhakkedathu, J. N.; Overall, C. M. Identifying and quantifying proteolytic events and the natural N terminome by terminal amine isotopic labeling of substrates. *Nat. Protoc* **2011**, *6* (10), 1578−611.
(77) Gevaert, K.; Goethals, M.; Martens, L.; Van Damme, J.; Staes, A.; Thomas, G. R.; Vandekerckhove, J. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat. Biotechnol.* **2003**, *21* (5), 566−9.
(78) Hains, P. G.; Robinson, P. J. The Impact of Commonly Used Alkylating Agents on Artifactual Peptide Modification. *J. Proteome Res.* **2017**, *16* (9), 3443−3447.
(79) Muller, T.; Winter, D. Systematic Evaluation of Protein Reduction and Alkylation Reveals Massive Unspecific Side Effects by Iodine-containing Reagents. *Mol. Cell Proteomics* **2017**, *16* (7), 1173−1187.
(80) Niu, B.; Martinelli Ii, M.; Jiao, Y.; Wang, C.; Cao, M.; Wang, J.; Meinke, E. Nonspecific cleavages arising from reconstitution of trypsin under mildly acidic conditions. *PLoS One* **2020**, *15* (7), No. e0236740.
(81) Schittmayer, M.; Fritz, K.; Liesinger, L.; Griss, J.; Birner-Gruenberger, R. Cleaning out the Litterbox of Proteomic Scientists' Favorite Pet: Optimized Data Analysis Avoiding Trypsin Artifacts. *J. Proteome Res.* **2016**, *15* (4), 1222−9.
(82) Frankenfield, A. M.; Ni, J.; Ahmed, M.; Hao, L. Protein Contaminants Matter: Building Universal Protein Contaminant Libraries for DDA and DIA Proteomics. *J. Proteome Res.* **2022**, *21* (9), 2104−2113.
(83) Hodge, K.; Have, S. T.; Hutton, L.; Lamond, A. I. Cleaning up the masses: exclusion lists to reduce contamination with HPLC-MS/MS. *J. Proteomics* **2013**, *88*, 92−103.
(84) Shteynberg, D. D.; Deutsch, E. W.; Campbell, D. S.; Hoopmann, M. R.; Kusebauch, U.; Lee, D.; Mendoza, L.; Midha, M. K.; Sun, Z.; Whetton, A. D.; Moritz, R. L. PTMProphet: Fast and Accurate Mass Modification Localization for the Trans-Proteomic Pipeline. *J. Proteome Res.* **2019**, *18* (12), 4262−4272.
(85) Giglione, C.; Boularot, A.; Meinnel, T. Protein N-terminal methionine excision. *Cell. Mol. Life Sci.* **2004**, *61* (12), 1455−74.
(86) Ross, S.; Giglione, C.; Pierre, M.; Espagne, C.; Meinnel, T. Functional and developmental impact of cytosolic protein N-terminal methionine excision in Arabidopsis. *Plant Physiol* **2005**, *137* (2), 623−37.
(87) Meinnel, T.; Giglione, C. N-terminal modifications, the associated processing machinery, and their evolution in plastid-containing organisms. *J. Exp Bot* **2022**, *73* (18), 6013−6033.
(88) Pozoga, M.; Armbruster, L.; Wirtz, M. From Nucleus to Membrane: A Subcellular Map of the N-Acetylation Machinery in Plants. *Int. J. Mol. Sci.* **2022**, *23* (22), 14492.
(89) Willems, P.; Ndah, E.; Jonckheere, V.; Van Breusegem, F.; Van Damme, P. To New Beginnings: Riboproteogenomics Discovery of N-Terminal Proteoforms in Arabidopsis Thaliana. *Front Plant Sci.* **2021**, *12*, 778804.
(90) Rowland, E.; Kim, J.; Bhuiyan, N. H.; van Wijk, K. J. The Arabidopsis Chloroplast Stromal N-Terminome: Complexities of Amino-Terminal Protein Maturation and Stability. *Plant Physiol* **2015**, *169* (3), 1881−96.
(91) Dinh, T. V.; Bienvenut, W. V.; Linster, E.; Feldman-Salit, A.; Jung, V. A.; Meinnel, T.; Hell, R.; Giglione, C.; Wirtz, M. Molecular identification and functional characterization of the first Nalpha-acetyltransferase in plastids by global acetylome profiling. *Proteomics* **2015**, *15* (14), 2426−35.
(92) Bienvenut, W. V.; Brunje, A.; Boyer, J. B.; Muhlenbeck, J. S.; Bernal, G.; Lassowskat, I.; Dian, C.; Linster, E.; Dinh, T. V.; Koskela, M. M.; Jung, V.; Seidel, J.; Schyrba, L. K.; Ivanauskaite, A.; Eirich, J.; Hell, R.; Schwarzer, D.; Mulo, P.; Wirtz, M.; Meinnel, T.; Giglione, C.; Finkemeier, I. Dual lysine and N-terminal acetyltransferases reveal the complexity underpinning protein acetylation. *Mol. Syst. Biol.* **2020**, *16* (7), No. e9464.
(93) Huang, S.; Taylor, N. L.; Whelan, J.; Millar, A. H. Refining the definition of plant mitochondrial presequences through analysis of sorting signals, N-terminal modifications, and cleavage motifs. *Plant Physiol* **2009**, *150* (3), 1272−85.
(94) Zybailov, B.; Sun, Q.; van Wijk, K. J. Workflow for large scale detection and validation of peptide modifications by RPLC-LTQ-Orbitrap: application to the Arabidopsis thaliana leaf proteome and an online modified peptide library. *Anal. Chem.* **2009**, *81* (19), 8015−24.
(95) Kim, M. S.; Zhong, J.; Pandey, A. Common errors in mass spectrometry-based analysis of post-translational modifications. *Proteomics* **2016**, *16* (5), 700−14.
(96) Maddelein, D.; Colaert, N.; Buchanan, I.; Hulstaert, N.; Gevaert, K.; Martens, L. The iceLogo web server and SOAP service for determining protein consensus sequences. *Nucleic Acids Res.* **2015**, *43* (W1), W543−6.
(97) Tilak, P.; Kotnik, F.; Nee, G.; Seidel, J.; Sindlinger, J.; Heinkow, P.; Eirich, J.; Schwarzer, D.; Finkemeier, I. Proteome-wide lysine acetylation profiling to investigate the involvement of histone deacetylase HDA5 in the salt stress response of Arabidopsis leaves. *Plant J.* **2023**, *115* (1), 275−292.
(98) Zhang, M.; Tan, F. Q.; Fan, Y. J.; Wang, T. T.; Song, X.; Xie, K. D.; Wu, X. M.; Zhang, F.; Deng, X. X.; Grosser, J. W.; Guo, W. W. Acetylome reprograming participates in the establishment of fruit metabolism during polyploidization in citrus. *Plant Physiol* **2022**, *190* (4), 2519−2538.
(99) Fussl, M.; Konig, A. C.; Eirich, J.; Hartl, M.; Kleinknecht, L.; Bohne, A. V.; Harzen, A.; Kramer, K.; Leister, D.; Nickelsen, J.; Finkemeier, I. Dynamic light- and acetate-dependent regulation of the proteome and lysine acetylome of Chlamydomonas. *Plant J.* **2022**, *109* (1), 261−277.
(100) Balparda, M.; Elsasser, M.; Badia, M. B.; Giese, J.; Bovdilova, A.; Hudig, M.; Reinmuth, L.; Eirich, J.; Schwarzlander, M.; Finkemeier, I.; Schallenberg-Rudinger, M.; Maurino, V. G. Acetylation of conserved lysines fine-tunes mitochondrial malate dehydrogenase activity in land plants. *Plant J.* **2022**, *109* (1), 92−111.
(101) van Wijk, K. J.; Friso, G.; Walther, D.; Schulze, W. X. Meta-Analysis of Arabidopsis thaliana Phospho-Proteomics Data Reveals Compartmentalization of Phosphorylation Motifs. *Plant Cell* **2014**, *26* (6), 2367−2389.
(102) Lu, Q.; Helm, S.; Rodiger, A.; Baginsky, S. On the extent of tyrosine phosphorylation in chloroplasts. *Plant Physiol* **2015**, *169* (2), 996−1000.
(103) Guo, Y.; Xiong, L.; Ishitani, M.; Zhu, J. K. An Arabidopsis mutation in translation elongation factor 2 causes superinduction of CBF/DREB1 transcription factor genes but blocks the induction of their downstream targets under low temperatures. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (11), 7786−91.
(104) Sanderfoot, A. A.; Kovaleva, V.; Zheng, H.; Raikhel, N. V. The t-SNARE AtVAM3p resides on the prevacuolar compartment in Arabidopsis root cells. *Plant Physiol* **1999**, *121* (3), 929−38.

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

(105) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157* (1), 105−32.
(106) Silva, J.; Ferraz, R.; Dupree, P.; Showalter, A. M.; Coimbra, S. Three Decades of Advances in Arabinogalactan-Protein Biosynthesis. *Front Plant Sci.* **2020**, *11*, 610377.
(107) Medina, J.; Ballesteros, M. L.; Salinas, J. Phylogenetic and functional analysis of Arabidopsis RCI2 genes. *J. Exp Bot* **2007**, *58* (15−16), 4333−46.
(108) Ponnala, L.; Wang, Y.; Sun, Q.; van Wijk, K. J. Correlation of mRNA and protein abundance in the developing maize leaf. *Plant J.* **2014**, *78* (3), 424−40.
(109) Edfors, F.; Danielsson, F.; Hallstrom, B. M.; Kall, L.; Lundberg, E.; Ponten, F.; Forsstrom, B.; Uhlen, M. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol. Syst. Biol.* **2016**, *12* (10), 883.
(110) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25* (1), 25−9.
(111) Gene Ontology, C. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **2021**, *49* (D1), D325−D334.
(112) Ge, S. X.; Jung, D.; Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **2020**, *36* (8), 2628−2629.
(113) Stintzi, A.; Schaller, A. Biogenesis of post-translationally modified peptide signals for plant reproductive development. *Curr. Opin Plant Biol.* **2022**, *69*, 102274.
(114) Matsubayashi, Y. Posttranslationally modified small-peptide signals in plants. *Annu. Rev. Plant Biol.* **2014**, *65*, 385−413.
(115) Olsson, V.; Joos, L.; Zhu, S.; Gevaert, K.; Butenko, M. A.; De Smet, I. Look Closely, the Beautiful May Be Small: Precursor-Derived Peptides in Plants. *Annu. Rev. Plant Biol.* **2019**, *70*, 153−186.
(116) Tavormina, P.; De Coninck, B.; Nikonorova, N.; De Smet, I.; Cammue, B. P. The Plant Peptidome: An Expanding Repertoire of Structural Features and Biological Functions. *Plant Cell* **2015**, *27* (8), 2095−118.
(117) Hu, X. L.; Lu, H.; Hassan, M. M.; Zhang, J.; Yuan, G.; Abraham, P. E.; Shrestha, H. K.; Villalobos Solis, M. I.; Chen, J. G.; Tschaplinski, T. J.; Doktycz, M. J.; Tuskan, G. A.; Cheng, Z. M.; Yang, X. Advances and perspectives in discovery and functional analysis of small secreted proteins in plants. *Hortic Res.* **2021**, *8* (1), 130.
(118) Takahashi, F.; Hanada, K.; Kondo, T.; Shinozaki, K. Hormone-like peptides and small coding genes in plant stress signaling and development. *Curr. Opin Plant Biol.* **2019**, *51*, 88−95.
(119) Kaufmann, C.; Sauter, M. Sulfated plant peptide hormones. *J. Exp Bot* **2019**, *70* (16), 4267−4277.
(120) Kim, J. S.; Jeon, B. W.; Kim, J. Signaling Peptides Regulating Abiotic Stress Responses in Plants. *Front Plant Sci.* **2021**, *12*, 704490.
(121) Willoughby, A. C.; Nimchuk, Z. L. WOX going on: CLE peptides in plant development. *Curr. Opin Plant Biol.* **2021**, *63*, 102056.
(122) Yuan, B.; Wang, H. Peptide Signaling Pathways Regulate Plant Vascular Development. *Front Plant Sci.* **2021**, *12*, 719606.
(123) Zhong, S.; Liu, M.; Wang, Z.; Huang, Q.; Hou, S.; Xu, Y. C.; Ge, Z.; Song, Z.; Huang, J.; Qiu, X.; Shi, Y.; Xiao, J.; Liu, P.; Guo, Y. L.; Dong, J.; Dresselhaus, T.; Gu, H.; Qu, L. J. Cysteine-rich peptides promote interspecific genetic isolation in Arabidopsis. *Science* **2019**, *364* (6443), eaau9564.
(124) Tost, A. S.; Kristensen, A.; Olsen, L. I.; Axelsen, K. B.; Fuglsang, A. T. The PSY Peptide Family-Expression, Modification and Physiological Implications. *Genes (Basel)* **2021**, *12* (2), 218.
(125) Fujita, S. CASPARIAN STRIP INTEGRITY FACTOR (CIF) family peptides - regulator of plant extracellular barriers. *Peptides* **2021**, *143*, 170599.
(126) Bartels, S.; Lori, M.; Mbengue, M.; van Verk, M.; Klauser, D.; Hander, T.; Boni, R.; Robatzek, S.; Boller, T. The family of Peps and their precursors in Arabidopsis: differential expression and localization but similar induction of pattern-triggered immune responses. *J. Exp Bot* **2013**, *64* (17), 5309−21.
(127) Huffaker, A.; Pearce, G.; Ryan, C. A. An endogenous peptide signal in Arabidopsis activates components of the innate immune response. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (26), 10098−103.
(128) Weits, D. A.; van Dongen, J. T.; Licausi, F. Molecular oxygen as a signaling component in plant development. *New Phytol* **2021**, *229* (1), 24−35.
(129) Gibbs, D. J.; Conde, J. V.; Berckhan, S.; Prasad, G.; Mendiondo, G. M.; Holdsworth, M. J. Group VII Ethylene Response Factors Coordinate Oxygen and Nitric Oxide Signal Transduction and Stress Responses in Plants. *Plant Physiol* **2015**, *169* (1), 23−31.
(130) van Dongen, J. T.; Licausi, F. Oxygen sensing and signaling. *Annu. Rev. Plant Biol.* **2015**, *66*, 345−67.
(131) Barreto, P.; Dambire, C.; Sharma, G.; Vicente, J.; Osborne, R.; Yassitepe, J.; Gibbs, D. J.; Maia, I. G.; Holdsworth, M. J.; Arruda, P. Mitochondrial retrograde signaling through UCP1-mediated inhibition of the plant oxygen-sensing pathway. *Curr. Biol.* **2022**, *32* (6), 1403−1411.
(132) Hammarlund, E. U.; Flashman, E.; Mohlin, S.; Licausi, F. Oxygen-sensing mechanisms across eukaryotic kingdoms and their roles in complex multicellularity. *Science* **2020**, *370* (6515), eaba3512.
(133) White, M. D.; Klecker, M.; Hopkinson, R. J.; Weits, D. A.; Mueller, C.; Naumann, C.; O'Neill, R.; Wickens, J.; Yang, J.; Brooks-Bartlett, J. C.; Garman, E. F.; Grossmann, T. N.; Dissmeyer, N.; Flashman, E. Plant cysteine oxidases are dioxygenases that directly enable arginyl transferase-catalysed arginylation of N-end rule targets. *Nat. Commun.* **2017**, *8*, 14690.
(134) Abbas, M.; Sharma, G.; Dambire, C.; Marquez, J.; Alonso-Blanco, C.; Proano, K.; Holdsworth, M. J. An oxygen-sensing mechanism for angiosperm adaptation to altitude. *Nature* **2022**, *606* (7914), 565−569.
(135) Puthiyaveetil, S.; McKenzie, S. D.; Kayanja, G. E.; Ibrahim, I. M. Transcription initiation as a control point in plastid gene expression. *Biochim Biophys Acta Gene Regul Mech* **2021**, *1864* (3), 194689.
(136) Chi, W.; He, B.; Mao, J.; Jiang, J.; Zhang, L. Plastid sigma factors: Their individual functions and regulation in transcription. *Biochim. Biophys. Acta* **2015**, *1847* (9), 770−8.
(137) Wu, G. Z.; Bock, R. GUN control in retrograde signaling: How GENOMES UNCOUPLED proteins adjust nuclear gene expression to plastid biogenesis. *Plant Cell* **2021**, *33* (3), 457−474.
(138) Berger, N.; Vignols, F.; Przybyla-Toscano, J.; Roland, M.; Rofidal, V.; Touraine, B.; Zienkiewicz, K.; Couturier, J.; Feussner, I.; Santoni, V.; Rouhier, N.; Gaymard, F.; Dubos, C. Identification of client iron-sulfur proteins of the chloroplastic NFU2 transfer protein in Arabidopsis thaliana. *J. Exp Bot* **2020**, *71* (14), 4171−4187.
(139) Rodriguez, E.; Chevalier, J.; Olsen, J.; Ansbol, J.; Kapousidou, V.; Zuo, Z.; Svenning, S.; Loefke, C.; Koemeda, S.; Drozdowskyj, P. S.; Jez, J.; Durnberger, G.; Kuenzl, F.; Schutzbier, M.; Mechtler, K.; Ebstrup, E. N.; Lolle, S.; Dagdas, Y.; Petersen, M. Autophagy mediates temporary reprogramming and dedifferentiation in plant somatic cells. *EMBO J.* **2020**, *39* (4), No. e103315.
(140) Bassal, M.; Abukhalaf, M.; Majovsky, P.; Thieme, D.; Herr, T.; Ayash, M.; Tabassum, N.; Al Shweiki, M. R.; Proksch, C.; Hmedat, A.; Ziegler, J.; Lee, J.; Neumann, S.; Hoehenwarter, W. Reshaping of the Arabidopsis thaliana Proteome Landscape and Co-regulation of Proteins in Development and Immunity. *Mol. Plant* **2020**, *13* (12), 1709−1732.
(141) Tsiatsiani, L.; Heck, A. J. Proteomics beyond trypsin. *FEBS J.* **2015**, *282* (14), 2612−26.
(142) Guo, X.; Trudgian, D. C.; Lemoff, A.; Yadavalli, S.; Mirzaei, H. Confetti: a multiprotease map of the HeLa proteome for comprehensive proteomics. *Mol. Cell Proteomics* **2014**, *13* (6), 1573−84.
(143) Sinitcyn, P.; Richards, A. L.; Weatheritt, R. J.; Brademan, D. R.; Marx, H.; Shishkova, E.; Meyer, J. G.; Hebert, A. S.; Westphall, M. S.; Blencowe, B. J.; Cox, J.; Coon, J. J. Global detection of human

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214

variants and isoforms by deep proteome sequencing. *Nat. Biotechnol.* **2023** *Online ahead of print*. DOI: 10.1038/s41587-023-01714-x

(144) Kaulich, P. T.; Cassidy, L.; Bartel, J.; Schmitz, R. A.; Tholey, A. Multi-protease Approach for the Improved Identification and Molecular Characterization of Small Proteins and Short Open Reading Frame-Encoded Peptides. *J. Proteome Res.* **2021**, *20* (5), 2895−2903.

(145) Friso, G.; Giacomelli, L.; Ytterberg, A. J.; Peltier, J. B.; Rudella, A.; Sun, Q.; Wijk, K. J. In-depth analysis of the thylakoid membrane proteome of Arabidopsis thaliana chloroplasts: new proteins, new functions, and a plastid proteome database. *Plant Cell* **2004**, *16* (2), 478−99.

(146) Omenn, G. S.; Lane, L.; Overall, C. M.; Pineau, C.; Packer, N. H.; Cristea, I. M.; Lindskog, C.; Weintraub, S. T.; Orchard, S.; Roehrl, M. H. A.; Nice, E.; Liu, S.; Bandeira, N.; Chen, Y. J.; Guo, T.; Aebersold, R.; Moritz, R. L.; Deutsch, E. W. The 2022 Report on the Human Proteome from the HUPO Human Proteome Project. *J. Proteome Res.* **2023**, *22* (4), 1024−1042.

https://doi.org/10.1021/acs.jproteome.3c00536
J. Proteome Res. 2024, 23, 185−214