

Supplemental Data

Pep2pro database

- **Modified ambiguity filter**

In the pep2pro database peptides matching to several different protein sequences are excluded from further analysis. In the TAIR7 protein database, 302 loci share the same protein sequence with at least one other locus (in TAIR8 316 and in TAIR9 318) and 126 protein sequences are assigned to more than one locus (in TAIR8 133 and in TAIR9 135). Peptides matching to different loci with identical sequence are also accepted. The rationale for this is that in the case of loci sharing the same sequence the protein was identified, no matter on which position in the genome it is encoded. The change of the ambiguity filter resulted in 129 more identified loci for the AtProteome TAIR7 dataset.

- **Loci, protein sequences and sequence databases**

MS/MS spectrum interpretation in high-throughput proteomics is usually done with database-dependent spectrum assignment. Dealing with different sequence databases or different releases of the same database both of genome and protein sequences is therefore central to proteome data analysis databases. In Figure 1 the tables dealing with loci, proteins and sequence databases are labeled in dark blue. In table *species* the chromosome names, the NCBI taxon ID and other information relevant for the species is stored. The sequence databases are inserted in table *grelease* together with additional information by linking them to the corresponding species. Each species and different versions of the sequence databases can have different ways of how proteins and loci are called. Table *regex* therefore contains for each database stored in *grelease* the database-specific regular expressions that are needed to distinguish protein, whole genome or decoy database hits, and for handling them appropriately.

Protein sequence databases generally contain a list of proteins in FASTA format, where each description line is followed by the protein sequence. The description line usually contains a locus protein accession number, the protein name and synonyms, as well as a short description of the protein. As mentioned above, different loci in a protein database can share the same protein sequence and different sequences

can have more than one locus. All the loci (protein accession numbers without splice information) in the different databases are stored with the respective grelease_ids in table *locus*. In this table also the number of theoretical tryptic peptides is inserted, which is the sum of all distinct tryptic peptides that all the splice variants of that locus can produce in theory. In table *protein* all the distinct sequences contained in the sequence databases are inserted, linked to the respective grelease_id. Table *gene_model* now assigns each locus, splice variant and alternative name of the protein to its specific sequence. To give an example (see also Supplemental Table S1): the *Arabidopsis thaliana* protein *rubisco activase* has locus AT2G39730 and alternative name RCA. It has the splice variants AT2G39730.1, AT2G39730.2 and AT2G39730.3, each with a different sequence. The locus, the alternative name and the splice variants are inserted in table *gene_model* and the three distinct protein sequences in table *protein*. In table *gene_model*, each protein sequence is assigned to its according splice variant. The locus and the alternative protein name are assigned to all three of these sequences. Like this, all information contained in the protein sequence files is stored in the database in a way that allows querying the database for loci, splice variants, protein sequences, or even alternative protein names, all in relation to the specific sequence database.

- **Which proteins were identified?**

This is one of the most important questions querying proteomics results. For answering this question, a link between the information contained in the sequence database files and the experimental results has to be created. In pep2pro, this link is provided by tables *peptide*, *peptide_locus* and *peptide_protein*, painted in red in Figure 1. Upon loading peptide spectrum assignments into the database, the peptide sequence is stored in table *peptide*. The peptide sequence is then linked to matching protein sequences through table *peptide_protein*, and to matching accessions through table *peptide_locus*. This structure allows for assigning one peptide to several loci and protein sequences. The ambiguity filter implemented in pep2pro only allows assignment of one peptide to different loci if the protein sequences of those loci are identical, but the database structure would also allow for other ambiguity filters or no ambiguity filter at all.

- **Where were the proteins identified?**

Depending on the specific question it is important to know in which project, condition, experiment, organ, sample, fraction or measurement the peptide or protein was identified. The tables painted in green in Table 1 are storing this information. For this, table *hit* is linked to table *measurement* in which the name of the mass spectrometry run is stored. By linking the measurement to table *experiment*, each mass spectrometry run is annotated with the corresponding organ, sample and fraction from which the injected material was retrieved. To deal with different users with different projects, tables *project*, *user_project* and *user* were set in place. This setup allows for assigning all spectrum assignments with metadata concerning the sample, experiment, project and user, which is important for interpreting the data and for drawing biological conclusions from mass spectrometry measurements.

- **Target-decoy approach**

The target-decoy search strategy allows for estimating the number of false positives that are associated with an entire data set. It permits estimation of the likelihood that a peptide spectrum assignment is correct given that it came from a dataset with a determined false discovery rate.¹ Furthermore, this approach allows for the estimation of local false discovery rates by counting the number of false positives in a subset of the data.² Given our database setup in which outputs from several search engines are combined, it is crucial to have an estimate about the local false discovery rates of the individual datasets to evaluate the score cut-offs for each search algorithm. In addition, it is useful if the effect of different filtering criteria can be assessed and if the false discovery rate of the final dataset can be calculated and given as a quality measure. This is feasible when the reverse hits are imported into the database and are retained upon combining the outputs of different search algorithms. One of the assumptions taken when creating a target-decoy database is that the target and the decoy databases do not overlap. In the study published by Gygi and Elia they mentioned that this assumption is reasonable, as only 0.02% peptides with lengths longer than eight amino acids were in common between the target and decoy databases.¹ In the TAIR8 protein sequence database 544 peptides with minimum length 6 are both reverse and forward. Upon loading search results into the pep2pro database these few peptides represented a major obstacle. For applying the target-decoy approach, one has to be able to assess how many peptides are hits against

the reverse database and each peptide can only be either a forward or a reverse hit. If the peptide exists in both databases, it can either be a forward or reverse hit depending on the output of the search algorithm. This leads to unstable false discovery estimates, as the number of decoy hits varies depending on which results were loaded first and might be different when the same data are loaded again. To solve this problem, the forward/decoy peptides are stored in the lookup table *peptide_truepositive*. Whenever a new peptide gets loaded into the database, the loader script checks, whether its sequence exists in the *peptide_truepositive* table and if yes, the peptide is set as a forward database hit. This ensures that no forward database hits are counted as false positives and that the calculated false discovery rates are stable.

Calculation of true tryptic peptides (missed cleavages and non-tryptic ends)

The number of theoretical tryptic peptides of a protein is an important measure for normalizing spectrum data in quantitative analyses. For quantification it is therefore important to have a stable definition of a true tryptic peptide, which is applied to the whole dataset and does not depend on the search algorithm. Yet depending on the search algorithm applied, the definition of a true tryptic peptide varies. PeptideProphet for example considers peptides at the C-terminus of a protein as tryptic, whereas for PepSplice they are semi-tryptic. For this reason, criteria were set into place in the pep2pro database that define a true tryptic peptide. For this, the protein sequence database is digested *in silico* applying the rule that trypsin cuts after Arg or Lys. If Arg or Lys are followed by Pro (KP/RP) site, the site is both cut and not cut (resulting in 3 true tryptic peptides). The resulting peptides between 400 and 6000 Da and minimum length 6 amino acids are then entered into the lookup table *peptide_truetryptic*. Whenever a peptide is loaded into the database, the loader script checks, whether its sequence exists in table *peptide_truetryptic* and if yes, the peptide is flagged to be true tryptic in table *peptide*.

- Post-translational modifications

Post-translational modification of proteins is an important regulatory mechanism of many cellular processes, and many proteomics studies are aimed at identifying them. For a proteome data analysis database this is inasmuch a challenge as the same peptide can be identified in its post-translationally modified and/or unmodified form, with different modifications at different positions, or even with the

information that the peptide bears a post-translational modification, without knowing the position. This scenario is especially true for phosphopeptides, for which the exact position of phosphorylation is often difficult to obtain. The tables in pep2pro dealing with this issue are painted in yellow in Figure 1. Table *peptide* containing the peptide sequence without modification is linked to table *peptide_mod*. Here, the modification is specified inside the peptide sequence for those variable modifications for which the position is known (e.g. EEFD^PRY^PS166RQYELIK is phosphorylated at Ser in position 7). Linked to table *peptide_mod* is table *peptide_mod_modification* in which the position and the kind of modification are given. In table *modification* the different modifications are listed. One field in there is *phosphorylation*. This value will be set to 1 for all peptides that are phosphorylated, also for those cases in which the phosphorylation position is not known and the peptide sequence in *peptide_mod* will be identical to the one in *peptide*.

A special case in this scenario are stable modifications. Stable modifications change the molecular mass of the peptide and its fragment masses and it is therefore necessary to specify these modifications if the data are to be displayed in a Spectrum Viewer or exported to a data repository. Upon loading new data into the pep2pro database the set of stable modifications included in the search are specified. As the stable modifications are mainly dependent on the experimental procedure applied (e.g. iodoacetamide treatment or labeling with isotope tags), the table *experiment* contains a link to the set of stable modifications described in table *mod_fix_set*, which links to table *mod_fix_set_modification*, in which the modifications and the modified amino acids are stored.

- **Different search algorithms and integration of the results into assemblies**

Depending on the aim of the proteomics study, the mass spectrometer used, the computation facilities available and personal preferences, different search algorithm are used for the peptide spectrum assignment, and in many cases more than one algorithm is used to search the same data. For a proteome data analysis database it is therefore important to integrate results from different search algorithms. In Figure 1 the tables in green are dealing with this issue. In pep2pro the four search algorithms Sequest/PeptideProphet^{3,4}, PepSplice⁵, Mascot (Matrix Science) and Inspect⁶ have been integrated so far.

For each search algorithm, the tables *xx_search* and *xx_hit* are created (*xx* = pp (PeptideProphet), ps (PepSplice), ma (Mascot), in (Inspect)) into which the spectrum assignments are inserted. Additional tables are dealing with issues specific for the individual search algorithms; e.g. in table *ma_datfile* the names of the Mascot .dat-files are inserted.

After searching the data with the different algorithms, the results have to be combined in order to get the final dataset. In pep2pro this is done by creating so-called ‘assemblies’ and the tables dealing with this assembling are painted in orange in Figure 1. For each algorithm, a threshold is defined above or below which data should be entered into the assembly (e.g. minimum probability of 0.9 for PeptideProphet results). The threshold and the *assembly_id* are stored for each algorithm in table *assembly_xx_searchstack* and the *search_ids* of those spectrum assignments fulfilling the criteria are stored in table *xx_searchstack*. Upon building the assembly, the peptide spectrum matches from the different search algorithms whose ids are in the respective *xx_searchstack* tables enter table *hit* and are linked with the *assembly_id*. When searching the same data with different algorithms, there will be an overlap where the algorithms correspond with each other, but there will also be cases in which they contradict each other. In those cases in which they contradict each other, the same spectrum gets assigned to different peptides or to peptides with different modifications. Upon creating an assembly in pep2pro, all instances of contradictions are flagged. The spectrum assignments in the assembly that are not hits against the decoy database and are not flagged, constitute the final dataset that can be analyzed further, exported to other resources and displayed on the public web interface.

- **Job control**

Table *control* serves to orchestrate the various processes run by the pep2pro database server. These processes include running Inspect and PepSplice searches, loading of search results and build-up of assemblies. For each process running on the server, the time, the date and the *user_id* are stored. Like this it can be tracked what was done when by whom and in which order. This setup also controls that those processes that should not run in parallel, like for example loading of search results, are run sequentially.

- **Coverage and proteotypic peptides**

A quality measure often given for protein identifications is the coverage, being the % of amino acids of a protein that are covered with identified peptide sequences. In pep2pro this value is calculated for each assembly and the value is stored in table *assembly_locus*. In this table also the number of proteotypic peptides is stored. The definition of a proteotypic peptide varies. In pep2pro a peptide is considered proteotypic when it unambiguously identifies a protein and in case it was detected with at least 3 different spectrum assignments to fully tryptic peptides in at least one fraction.

- **Quantification with normalized spectral counting**

In pep2pro, quantification is done by calculating the expected contribution of each individual protein to the samples total peptide tool.⁷ For normalization, the number of measured spectra of tryptic peptides and the sum of the theoretical tryptic peptides of the identified proteins have to be calculated for each organ or sample. These values are then stored in tables *assembly_apex_organ* or *assembly_apex_sample* in relation to the *assembly_id*. With this information, the abundance can be calculated for each identified protein in each assembly.

- **Handling whole genome hits**

Whole genome hits are peptides that were identified searching the genome databases. As they match to genome regions without annotated protein coding capacity, they can not be linked to a protein locus or sequence. In addition, not all spectrum assignments against the genome database are accepted as whole genome hit, but only those fulfilling certain criteria as specified above. To deal with the whole genome hits, tables *alithia* and *dna* were therefore set in place. In table *peptide* all spectrum assignments against the whole genome database get a *dna_id*, which links to table *dna*. In this table, the peptide's nucleotide sequence, its start and end positions as well as the reading direction are stored. Those peptides in table *dna* that fulfill the selection criteria and hence define a new locus get an *alithia_id*, which links the peptide to table *alithia*. In this table, the new loci are stored with an identifier, the description and additional information, e.g. whether it was identified using a gene prediction algorithm.

- **Export of the data to a public data repository**

The amount of large-scale data being generated in proteomics has increased exponentially over the last few years. It is therefore increasingly important that scientists are able to exchange, compare and retrieve datasets and to disseminate their data to the scientific community. To properly interpret a proteomics study and for comparison of different studies, it is important that the datasets are accompanied with metadata, describing the sample and the analysis method applied. To this end the Proteomics Standards Initiative (PSI) develops the Minimum Information about a Proteomics Experiment (MIAPE) checklist, specifying the data and metadata that should be collected from various proteomics workflows.⁸ In addition, data format standards and controlled vocabularies are important for annotating data with all the information required. In order to make full use of the data, the data and accompanying metadata have to be made accessible to the scientific community, which is best done by centralized and standard compliant data repositories. The database of choice for export of data from the pep2pro database is the PRoteomics IDentification database (PRIDE).⁹ The tables dealing with this export are *pride_experiment* and *pride*. In table *pride* the various metadata values are stored, for example the NEWT accession number and name for the organism, or the PSI accession number and name of the mass spectrometer instrument. Table *pride* is linked to table *assembly*, as only the final datasets are to be exported. Table *pride_experiment* links the PRIDE dataset to table *experiment* so that individual experiments, for which an output file is to be created, can be specified. With this setup experiments representing a 'logical bundle' can be combined in one export file. The data format for the export files is PRIDE 2.1 XML fitting to the PRIDE XSD schema. In the pep2pro database, the workflow is such, that the PRIDE.XML files can be directly put onto the PRIDE FTP server.

- **Proteogenomic mapping of the peptides**

A recent development is to use proteomics data in genome annotation, a process often referred to as proteogenomics.¹⁰ Proteogenomic mapping of peptides onto the genome for efficient visualization is of growing importance, yet not easy to achieve. The complication arises from the fact that the genome sequence coding for a peptide can be distributed over several different exons and that the peptide cannot be matched onto a contiguous stretch of genomic DNA. As a further complication, intron lengths are

independent of the reading frame, thus the number of genomic introns that are a multiple of 3 basepairs ($3n$) conserving the reading frame is similar to the number that are a multiple of 3 plus 1 bases ($3n+1$) or plus 2 bases ($3n+2$) changing the reading frame.¹¹ Therefore it can be expected that the triplet of nucleotides encoding an amino acid may be spread over different exons, being separated by an intron sequence. These complications make the mapping of peptides, which derive from spliced mRNA impossible with standard tools that are typically used for mapping peptide sequences onto DNA sequences. Mapping the peptides onto the genome is important for the integration of proteomics data into genome browsers, however, there is no simple protocol to uncouple the mapping process from platforms like Ensembl¹² and PeptideAtlas¹³ where this has been implemented, and to make use of the underlying algorithms in individualized mapping procedures. We have therefore developed the Pep2Pro2DNA algorithm, which enables the mapping of peptides onto genomes by calculating the exact start and stop positions on the DNA, overcoming the difficulties discussed above by attributing the genome position to each coding nucleotide.

The tables in the pep2pro database storing the information needed for proteogenomic mapping are indicated in light blue. Table *peptide_locus_location* stores the sequence of each gene_model in table *gene_model* and table *exon* contains the start and stop positions of all coding sequences for each gene model, together with the frame. Tables *transsplice* and *transsplice_sequence* deal with the special case of trans-spliced sequences. Upon trans-splicing, different primary RNA transcripts coded on different genomic locations and with different accessions are joined together for form one protein. Table *transsplice* contains the name of the trans-spliced sequence (e.g. ATMG00513,ATMG00665,ATMG00060:Mitochondrial NADH dehydrogenase subunit 5. The gene is trans-spliced from three different pre-cursors, NAD5a, NAD5b and NAD5c). In table *transsplice_sequence*, the trans-spliced sequence is linked to the sequences of the corresponding gene models, together with the information, in which order the precursor RNAs of the gene models are spliced together.

The information in table *exon* was originally taken from a GFF3 (Generic Feature Format 3) file provided by www.arabidopsis.org. As we consider the pep2Pro2DNA algorithm to be useful also independently of the pep2pro database we provide it as a stand-alone application for which the information concerning the start and stop positions of the coding sequences are taken from the GFF3 file. Additional prerequisites for mapping the peptides onto the genome using Pep2Pro2DNA are the peptide sequences and the accession IDs of the corresponding proteins, as well as the genomic DNA sequence spanning the coding sequences of each protein in FASTA format. The Pep2ProDNA algorithm pseudocode and perl program are available at www.pb.ethz.ch/downloads/Pep2Pro2DNA, together with a perl implementation for which sample input files are provided. For running the program, installation of Perl and BioPerl¹⁴ is required (current implementation with Perl v. 5.8.5 and BioPerl v. 1.26.4.4).

The principle of the mapping procedure is demonstrated in Figure S1. In a first step, the start and stop positions of all coding sequences for each protein identified by the peptides to me matched are retrieved and ordered by their position on the DNA (Figure S1B). With this information, a two-dimensional array is created, containing the position of every amino acid in the protein as the first dimension, the triplet position of every encoding nucleotide as the second dimension, and its genome position as value. In addition, the nucleotides of the individual coding sequences are concatenated to generate the complete coding sequence of each splice variant of the protein. In the second step the complete coding sequence for each splice variant is first translated into its amino acid sequence, onto which the peptide is mapped to retrieve its start and end positions in the protein. The algorithm now loops through the array created in step one commencing with the peptide start position in the protein, and assigns the positions of the coding nucleotides to the peptide (Figure S1C). In this way, the peptide coding sequence is located on the genome, which is the basis for the visualization of the identified peptides on the genome (Figure S1D). Figure S1E shows the result of the mapping of 4 peptides of locus AT4G00490. For peptide 4, whose mRNA had been spliced, two start and stop positions are given for each coding sequence, with the first stop position being the end of the second last exon and the second start position the start of the last exon.

References

- 1 Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry*. Nat Methods, 2007. **4**(3): p. 207-14.
- 2 Castellana, N.E., S.H. Payne, Z. Shen, M. Stanke, V. Bafna and S.P. Briggs, *Discovery and revision of Arabidopsis genes by proteogenomics*. Proc Natl Acad Sci U S A, 2008. **105**(52): p. 21034-8.
- 3 Eng, J.K., A.L. McCormack and J.R. Yates, *An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database*. Journal of the American Society for Mass Spectrometry, 1994. **5**(11): p. 976-989.
- 4 Keller, A., A.I. Nesvizhskii, E. Kolker and R. Aebersold, *Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search*. Anal Chem, 2002. **74**(20): p. 5383-92.
- 5 Roos, F.F., R. Jacob, J. Grossmann, B. Fischer, J.M. Buhmann, W. Gruissem, S. Baginsky and P. Widmayer, *PepSplice: cache-efficient search algorithms for comprehensive identification of tandem mass spectra*. Bioinformatics, 2007. **23**(22): p. 3016-23.
- 6 Payne, S.H., M. Yau, M.B. Smolka, S. Tanner, H. Zhou and V. Bafna, *Phosphorylation-specific MS/MS scoring for rapid and accurate phosphoproteome analysis*. J Proteome Res, 2008. **7**(8): p. 3373-81.
- 7 Baerenfaller, K., J. Grossmann, M.A. Grobei, R. Hull, M. Hirsch-Hoffmann, S. Yalovsky, P. Zimmermann, U. Grossniklaus, W. Gruissem and S. Baginsky, *Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics*. Science, 2008. **320**(5878): p. 938-41.
- 8 Taylor, C.F., N.W. Paton, K.S. Lilley, P.A. Binz, R.K. Julian, Jr., A.R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E.W. Deutsch, M.J. Dunn, A.J. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T.A. Neubert, S.D. Patterson, P. Ping, S.L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, T.M. Vondriska, J.P. Whitelegge, M.R. Wilkins, I. Xenarios, J.R. Yates, 3rd and H. Hermjakob, *The minimum information about a proteomics experiment (MIAPE)*. Nat Biotechnol, 2007. **25**(8): p. 887-93.
- 9 Vizcaino, J.A., R. Cote, F. Reisinger, H. Barsnes, J.M. Foster, J. Rameseder, H. Hermjakob and L. Martens, *The Proteomics Identifications database: 2010 update*. Nucleic Acids Res, 2010. **38**(Database issue): p. D736-42.
- 10 Ansong, C., S.O. Purvine, J.N. Adkins, M.S. Lipton and R.D. Smith, *Proteogenomics: needs and roles to be filled by proteomics in genome annotation*. Brief Funct Genomic Proteomic, 2008. **7**(1): p. 50-62.
- 11 Roy, S.W. and D. Penny, *Intron length distributions and gene prediction*. Nucleic Acids Res, 2007. **35**(14): p. 4737-42.
- 12 Hubbard, T.J., B.L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S.C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider,

- M. Hammond, J. Herrero, R. Holland, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X.M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal and E. Birney, *Ensembl 2007*. Nucleic Acids Res, 2007. **35**(Database issue): p. D610-7.
- 13 Deutsch, E.W., *The PeptideAtlas Project*. Methods Mol Biol, 2010. **604**: p. 285-96.
- 14 Stajich, J.E., D. Block, K. Boulez, S.E. Brenner, S.A. Chervitz, C. Dagdigian, G. Fuellen, J.G. Gilbert, I. Korf, H. Lapp, H. Lehvaslaiho, C. Matsalla, C.J. Mungall, B.I. Osborne, M.R. Pocock, P. Schattner, M. Senger, L.D. Stein, E. Stupka, M.D. Wilkinson and E. Birney, *The Bioperl toolkit: Perl modules for the life sciences*. Genome Res, 2002. **12**(10): p. 1611-8.

Supplemental Table S1: Relation between accession, gene model and protein sequence, exemplified on rubisco activase.

accession	gene_model	protein_id
AT2G39730	AT2G39730	289982
AT2G39730	AT2G39730.1	289982
AT2G39730	RCA	289982
AT2G39730	AT2G39730	290608
AT2G39730	AT2G39730.2	290608
AT2G39730	RCA	290608
AT2G39730	AT2G39730	290673
AT2G39730	AT2G39730.3	290673
AT2G39730	RCA	290673

Supplemental Table S2: Number of assigned spectra, distinct peptides, and proteins in different samples and organs of the pep2pro TAIR9 dataset (= pep2pro TAIR9 without single hits).

Plant tissue	Spectra	Distinct peptides	Proteins
<i>Differentiated organs</i>	2,331,761	127,463	14,302
Roots	994,495	69,098	11,369
Roots 10 days	38,506	20,291	5174
Roots 23 days	33,220	16,993	4470
Roots Castellana	922,769	58,402	10,996
Leaves	515,635	50,510	8722
Cotyledons	88,516	22,435	5063
Juvenile leaves	95,346	24,932	5580
Leaves Castellana	331,773	32,553	7123
Flowers	363,696	53,288	9383
Flower buds	90,040	27,992	6559
Open flowers	95,305	29,507	6758
Flowers Castellana	142,426	17,239	4970
Carpels	35,925	13,514	4005
Siliques	147,081	29,081	6698
Siliques AtProteome	80,165	23,127	5747
Siliques Castellana	66,916	8362	2961
Seeds	91,145	14,006	3764
Pollen Grobei	132,268	25,693	4914
Seedling shoots	87,441	23,077	5317
Shoots 3 weeks	41,784	16,131	4313
Shoots 5 weeks	45,657	16,622	4339
<i>Cell culture</i>	329,091	50,167	8358
Dark	151,075	35,036	6486
Light	146,032	32,926	6393
Light; small	31,984	15,324	4482
Total	2,660,852	141,235	14,522
TAIR9	27,379		

Supplemental Table S4: P-values for the functional classification of the identified organ proteomes.

Shown are over-represented GO categories with p-value < 10⁻⁶ from aspect *biological process* in the organ proteomes as compared to all proteins in the TAIR9 database.

GO ID	GO Term	Flowers	Leaves	Roots	Seeds	Siliques	Pollen	Cell culture
GO:0046686	response to cadmium ion	< 1e-30	< 1e-30	< 1e-30	< 1e-30	< 1e-30	< 1e-30	< 1e-30
GO:0006412	translation	< 1e-30	< 1e-30	2.80E-24	1.20E-26	< 1e-30	4.60E-21	< 1e-30
GO:0009793	embryonic development ending in seed dor...	8.10E-21	6.90E-18	2.20E-13	1.80E-16	1.90E-14	2.60E-10	1.10E-23
GO:0006457	protein folding	7.40E-20	1.20E-18	1.40E-13	1.50E-11	6.10E-17	1.60E-14	1.70E-18
GO:0006886	intracellular protein transport	3.00E-19	2.00E-15	4.50E-15	3.30E-20	1.10E-17	1.20E-25	1.30E-14
GO:0009651	response to salt stress	6.80E-19	1.30E-25	2.60E-23	5.10E-26	8.10E-26	2.60E-20	4.80E-17
GO:0009409	response to cold	1.10E-15	1.50E-19	2.00E-13	3.70E-21	8.80E-23	1.50E-08	2.10E-15
GO:0006888	ER to Golgi vesicle-mediated transport	7.40E-09	2.90E-09	4.60E-08	7.00E-14	4.80E-07	1.30E-11	3.90E-09
GO:0006096	glycolysis	1.30E-08	3.20E-09	2.30E-08	4.80E-14	8.30E-14	1.10E-11	6.20E-11
GO:0006414	translational elongation	6.90E-10	2.40E-08	2.00E-07	4.60E-10	5.20E-11	8.30E-11	1.30E-09
GO:0006418	tRNA aminoacylation for protein translat...	2.10E-08	4.20E-09		6.10E-10	3.00E-11	4.20E-10	4.00E-11
GO:0042742	defense response to bacterium	2.70E-16	3.20E-16	1.60E-13	8.10E-16	8.20E-20		7.40E-14
GO:0042254	ribosome biogenesis	1.80E-15	6.50E-11	2.50E-08	9.00E-07	1.20E-11		1.50E-18
GO:0006635	fatty acid beta-oxidation	3.80E-07		8.40E-07	7.60E-08		6.50E-07	7.80E-11
GO:0000059	protein import into nucleus, docking	1.10E-07	5.50E-08		3.90E-07		4.70E-07	8.70E-09
GO:0006633	fatty acid biosynthetic process	1.20E-09	5.40E-08		1.00E-08	1.50E-09		5.70E-07
GO:0016192	vesicle-mediated transport	2.60E-13	6.80E-08	2.10E-11		6.40E-11	9.40E-16	
GO:0006508	proteolysis	8.80E-12	5.10E-10	3.30E-09	5.90E-10	1.80E-12		
GO:0055114	oxidation reduction	1.70E-08	1.70E-09	1.10E-08	7.50E-10	4.10E-10		
GO:0006413	translational initiation		9.70E-07			6.70E-07	7.50E-11	2.50E-10
GO:0015031	protein transport	1.50E-08		3.20E-07		1.00E-09	6.30E-09	
GO:0009408	response to heat	1.00E-07	4.80E-09		6.50E-13			3.70E-08
GO:0018119	peptidyl-cysteine S-nitrosylation	3.70E-08	1.80E-08		3.70E-09	1.40E-08		
GO:0009853	photorespiration	2.10E-07	9.50E-09	9.30E-09			1.70E-08	
GO:0006979	response to oxidative stress			5.10E-11	8.00E-08	8.00E-10		
GO:0015979	photosynthesis	1.40E-10	5.00E-14			1.20E-08		
GO:0006098	pentose-phosphate shunt	1.60E-08	7.40E-08			4.00E-07		
GO:0015995	chlorophyll biosynthetic process		6.00E-07			5.70E-07		
GO:0009073	aromatic amino acid family biosynthetic ...		7.80E-07					3.50E-07
GO:0006099	tricarboxylic acid cycle				2.20E-07		1.10E-09	
GO:0009658	chloroplast organization		6.70E-08					
GO:0043623	cellular protein complex assembly		5.40E-07					
GO:0018130	heterocycle biosynthetic process			7.60E-09				
GO:0044271	cellular nitrogen compound biosynthetic ...			8.00E-08				
GO:0042364	water-soluble vitamin biosynthetic proce...			2.90E-07				
GO:0009407	toxin catabolic process			9.90E-07				
GO:0019915	lipid storage				2.50E-09			
GO:0042542	response to hydrogen peroxide				2.60E-07			
GO:0009846	pollen germination						6.30E-13	
GO:0009860	pollen tube growth						5.20E-12	
GO:0006511	ubiquitin-dependent protein catabolic pr...						2.50E-08	
GO:0009225	nucleotide-sugar metabolic process						8.80E-08	
GO:0006396	RNA processing							2.60E-09
GO:0006996	organelle organization							4.90E-07
GO:0000398	nuclear mRNA splicing, via spliceosome							6.50E-07

Supplemental Table S5: P-values for the functional classification of the identified organ biomarkers.

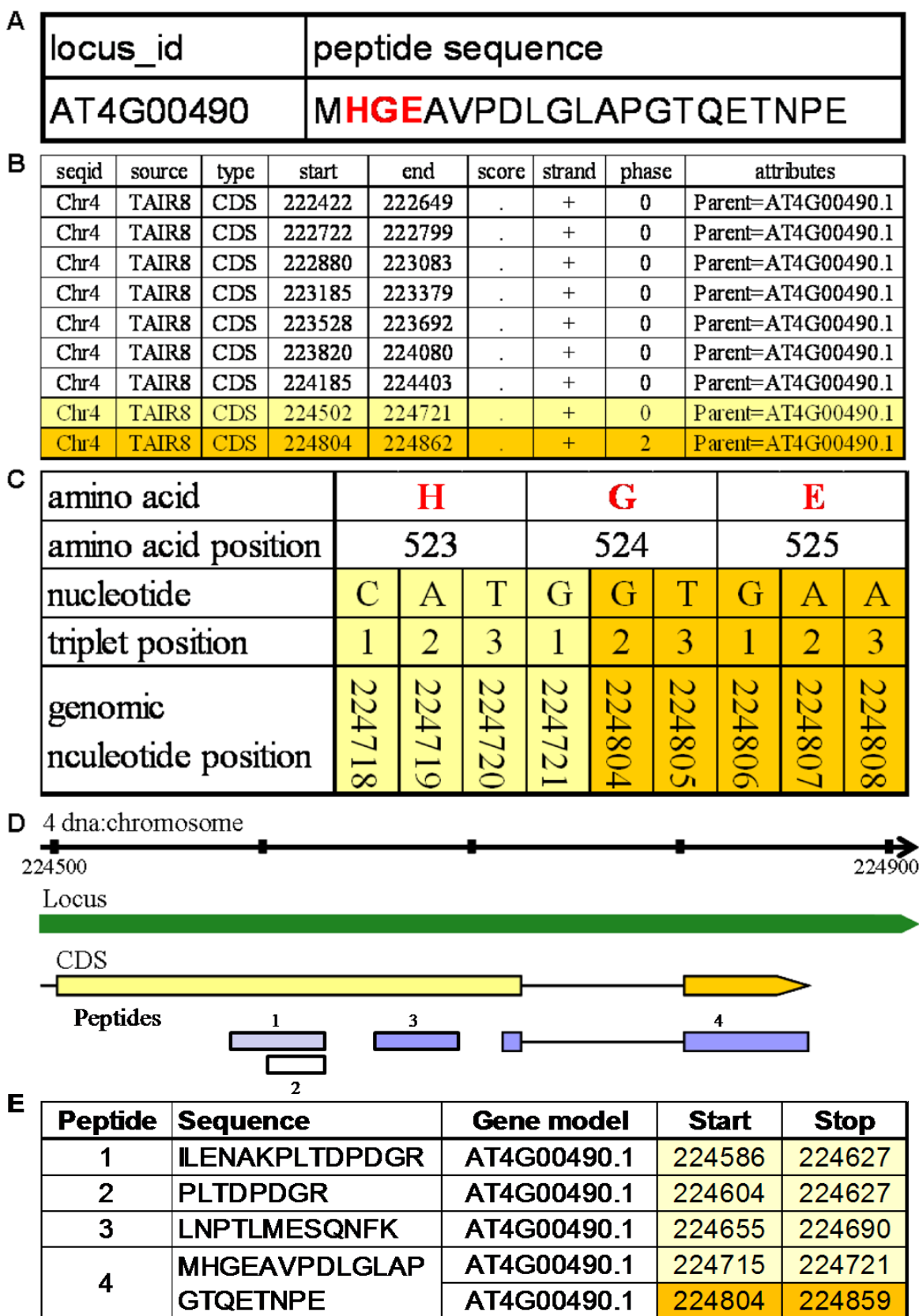
Shown are over-represented GO categories with p-value < 0.01 from aspect *biological process* in the organ biomarkers as compared to all identified proteins.

GO ID	GO Term	Flowers	Roots	Seeds	Siliques	Pollen
GO:0010584	pollen exine formation	4.60E-07				
GO:0006367	transcription initiation from RNA polyme...	0.0012				
GO:0009556	microsporogenesis	0.0017				
GO:0009793	embryonic development ending in seed dor...	0.0035		0.00041		
GO:0008299	isoprenoid biosynthetic process	0.0072				
GO:0006869	lipid transport		0.00031	0.00762		
GO:0032957	inositol trisphosphate metabolic process		0.0008			
GO:0009407	toxin catabolic process		0.00143			
GO:0009735	response to cytokinin stimulus		0.00247			
GO:0006829	zinc ion transport		0.00408			
GO:0019760	glucosinolate metabolic process		0.00776			
GO:0009834	secondary cell wall biogenesis		0.00818			
GO:0010162	seed dormancy			1.70E-05		
GO:0019915	lipid storage			6.80E-05		
GO:0006950	response to stress			0.00766		
GO:0043193	positive regulation of gene-specific tra...				0.00034	
GO:0009718	anthocyanin biosynthetic process				0.00089	
GO:0042545	cell wall modification				0.00634	0.0036
GO:0009860	pollen tube growth					4.60E-05
GO:0006468	protein amino acid phosphorylation					0.003
GO:0006865	amino acid transport					0.0053
GO:0007264	small GTPase mediated signal transductio...					0.0093

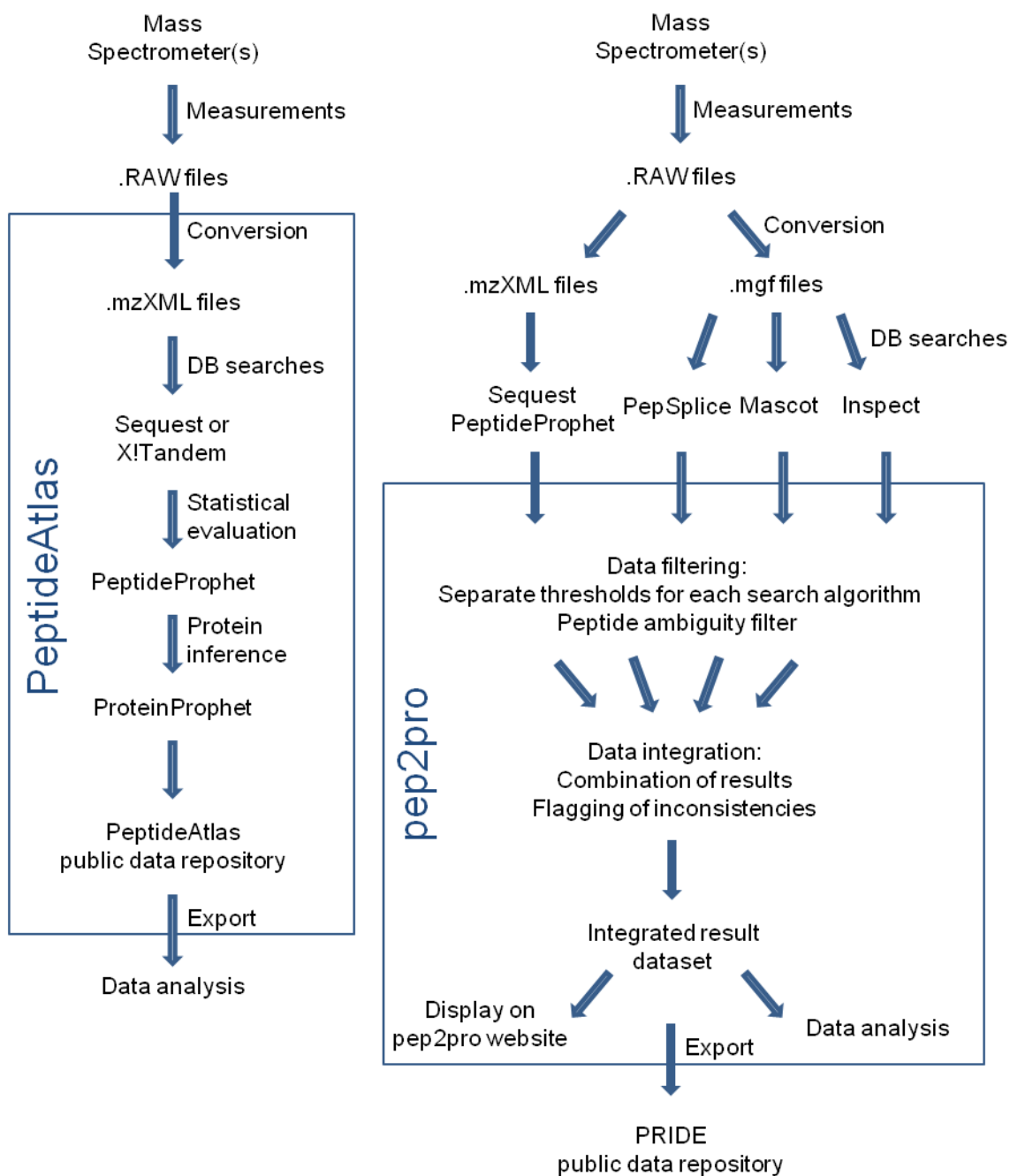
Supplemental Table S8: Primers for the new gene or refined gene models

Primer	Sequence
AT00078 forward	5' CACGCGAGGAGCATTACAA 3'
AT00078 reverse	5' CTCACAAACTCTGACGCTTC 3'
AT00065 forward	5' ATGGCTAGGGTTTATAGTAATTGGG 3'
AT00065 reverse	5' TACCATCACTGTTAGGGTCC 3'
AT00073 forward	5' TATGGCTATGCTGTTGATTCCC 3'
AT00073 reverse	5' GAACTCCTGAAAGATGTGCAGC 3'
AT00086 forward	5' GGTTGCTGACACGGAAAAGC 3'
AT00086 reverse	5' TCAACCAAAGTAGTAGAACCTT 3'
AT00088 forward	5' GATTGACACTCAGGTGCACAGTAG 3'
AT00088 reverse	5' GTCTCAGAGGATAAGTTAGGAAG 3'
AT00067 forward	5' CGTACTCGTAGAGATTCCTTAGC 3'
AT00067 reverse	5' GAAAAGAGCCTCGCCTAAGAC 3'
AT00080 forward	5' CTTGACAACAAAATACAAAGGCGG 3'
AT00080 reverse	5' GAAGGGAACCTGGGTTTCTTG 3'

Supplemental Figure S1: Principle of mapping peptides onto genomes using the Pep2Pro2DNA algorithm. The mapping procedure is exemplified for four peptides identified from gene model AT4G00490. **A** Sequence of the peptide to be mapped. **B** The information in the GFF3 file provides the start and stop position of every coding sequence (CDS) predicted for this protein. **C** Principle of how the genomic location for each coding nucleotide of the peptide is calculated. **D** Visualization of the mapping with four peptides that map to this region. **E** The gene model, and the start and stop positions for each peptide as given by the Pep2Pro2DNA algorithm.



Supplemental Figure S2: Comparison of PeptideAtlas and pep2pro highlighting the fundamental differences between the analysis workflows of the two systems.



Supplemental Figure S3: Gene model AT00067 A Primer design for the sequencing reactions with the gene model as predicted by the TAIR database (blue), the identified peptide (magenta) and the primers (red). **B** Alignment of the primer sequences, the genomic sequence and the sequencing results. **C** Translation of the sequencing results into amino acid sequence with the sequence of the predicted gene model in blue and the peptide sequence in magenta.

