TITLE: PEO: Plant Expression Omnibus - a comparative transcriptomic database for 103 Archaeplastida

AUTHORS: Eugene Koh[1], William Goh[1], Irene Julca[1], Erielle Villanueva[1], Marek Mutwil[1*]

[1]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore, 637551, Singapore

*Corresponding author:
Marek Mutwil
School of Biological Sciences,
Nanyang Technological University,60 Nanyang Drive,
637551, Singapore,
Singapore
Email: mutwil@ntu.edu.sg

Abstract

The Plant Expression Omnibus (PEO) is a web application that provides biologists with access to gene expression insights across over 100 plant species, ~60,000 manually annotated RNA-seq samples, and more than four million genes. The tool allows users to explore the expression patterns of genes across different organs, identify organ-specific genes, and discover top co-expressed genes for any gene of interest. PEO also provides functional annotations for each gene, allowing for the identification of genetic modules and pathways. PEO is designed to facilitate comparative kingdom-wide gene expression analysis and provide a valuable resource for plant biology research. We provide two case studies to demonstrate the utility of PEO in identifying candidate genes in pollen coat biosynthesis and investigating the biosynthetic pathway components of capsaicin in *Capsicum annuum*. The database is freely available at https://expression.plant.tools/.

Introduction

The earliest members of the plant kingdom evolved circa one billion years ago from photosynthetic algae. The archetypical characteristics we associate with plants, such as seeds, roots, leaves, flowers, and wood, were evolutionary innovations that allowed plants to colonize diverse habitats successfully. Today, plants form an indispensable component of life on Earth as the primary food producer in the terrestrial ecosystem. Plants being mostly sessile, have also had to evolve means of adapting quickly to their local environments. These adaptations came in the form of complex regulatory and metabolic pathways, which we are now only beginning to unravel.

Studying plant biology and genetics will allow us to improve desirable traits and engineer crops that provide higher yield and/or are more tolerant to various environmental stresses (Niazian, 2019; Nowicka et al., 2018). The use of plants as a source of various medicinal components is also exemplified in the extraction of the analgesic salicins from willow bark or the antimalarial artemisinin from *Artemisia annua* (Desborough and Keeling, 2017; Krishna et al., 2008). The elucidation of biosynthetic pathways of these medicinal compounds could then be harnessed for large-scale production and modification in recombinant systems (Cravens et al., 2019; Paddon et al., 2013; Sabzehzari et al., 2020). Currently, the study of plant gene function requires the characterization of individual genes and their interactors and regulators, which are typically limited to a handful of model plants. Consequently, bioinformatic approaches that integrate genomic, transcriptomic, metabolomic, and phenotypic data have been developed to accelerate the process of gene function prediction (Hansen et al., 2018; Radivojac et al., 2013; Rhee and Mutwil, 2014).

Studying gene co-expression can effectively address this, given the “guilt-by-association” principle, where genes with similar expression are likely to have similar functions (Rhee and Mutwil, 2014). Co-expression analysis allows one to transfer functional annotation to unannotated genes, predict undiscovered members of metabolic pathways, and identify regulator-target relationships (Rao and Dixon, 2019). Plant organs have signature gene expression patterns and similar profiles between organs are correlated to the similarity in function of the organs (Ma et al., 2005). Moreover, many genes exhibiting organ-specific expression are found to be conserved across land

plants (Julca et al., 2021). Consequently, a gene's expression profile across organs can provide clues to a gene's function. For example, two genes that show similar gene expression profiles across different organs and tissues are likely to be part of the same protein complex (Persson et al., 2005) or biosynthetic pathway (Delli-Ponti et al., 2021) or to be targeted to the same subcellular compartment (Ryngajllo et al., 2011).

Over the last decade, sequencing cost has been falling faster than Moore's law (Wetterstrand, 2013), giving rise to immense volumes of plant genomic and transcriptomic data. Many online databases with diverse functionalities and plant species have emerged to provide analytical and visualization tools for the scientific community (Lim et al., 2022). However, no existing database has been designed to cover gene expression for more than 20 species. We believe that the full value of RNA-seq repositories is only realized when insights are accessible to biologists working on the genes, including non-model species. We created a user-friendly tool to help non-programmer biologists extract insights about their gene of interest from big data (Tan et al., 2020). A web application to access gene expression knowledge on any gene would be needed to provide an effortless and instantaneous experience for biologists, which requires large volumes of processed data. While the data is publicly available, the interfaces have not been specifically designed for bulk data acquisition for whole kingdoms. We addressed this by designing a parallel-processing pipeline to automatically discover, download and process RNA-seq data for the whole plant kingdom in less than a month (Goh and Mutwil, 2021), making the provision and maintenance of a kingdom-wide resource feasible.

Here, we developed the Plant Gene Expression Omnibus (PEO, https://expression.plant.tools), a web application to provide an interface for biologists to access gene expression insights across >100 plant species. Specifically, the tool addresses how a gene is expressed across different organs, whether it displays organ-specific expression patterns and identifies top organ-specific genes for each organ. The tool also automatically provides a list of the highest correlated genes based on the Pearson Correlation Coefficient for any gene of interest, negating the need for separate analysis. In addition, Pfam and Mapman annotations have also been linked to each gene, allowing for functional searches to identify genetic modules or pathways. The software design considers scalability, with the ability to include more species, other expression

conditions, and other gene annotations. The project lays the architectural foundations for a kingdom-wide gene expression analysis resource.

Methods

Data source

Supplemental Table S1 details the sources of raw files used and contributors. These were then reorganized into the upload format described in Supplemental Table S2, using custom Python scripts. A total of 103 species were included. The complete list is available in Supplemental Dataset S1. To obtain the TPM matrices, RNA-seq transcripts were obtained from public databases via our pipeline (Goh and Mutwil, 2021). Gene expression was quantified as transcript per million (TPM) via pseudoalignment to reference coding sequences (CDS) of each species, using kallisto (Bray et al., 2016). Organ annotations were done by manually mapping samples to Plant Ontology (PO) terms, which are precise and consistent references to plant anatomical entities across species (Walls et al., 2019).

Genes were functionally annotated with Mapman annotations via Mercator (Schwacke et al., 2019) for four representative species and were assigned to the remaining species via sequence similarity score from diamond (Buchfink et al., 2015). Mapman was chosen over Gene Ontology (GO) as it is specific to plant pathways and processes (Klie and Nikoloski, 2012). Genes were also annotated with protein families (Pfam) (Mistry et al., 2021). Gene identifiers were standardized to the versions in TPM matrices, and alternative isoforms were excluded.

Database schema and software architecture

The schema defines how data is modeled and stored in our document-oriented database and can be described with an entity relational diagram (ERD) (Supplemental Figure S1). The chosen database technology is MongoDB, a document-oriented database allowing flexible data modeling to optimize our specific read patterns. This is more suitable than relational databases that enforce data schema normalization, which would incur expensive join operations on every read. The database is hosted serverless on MongoDB Atlas, on AWS.

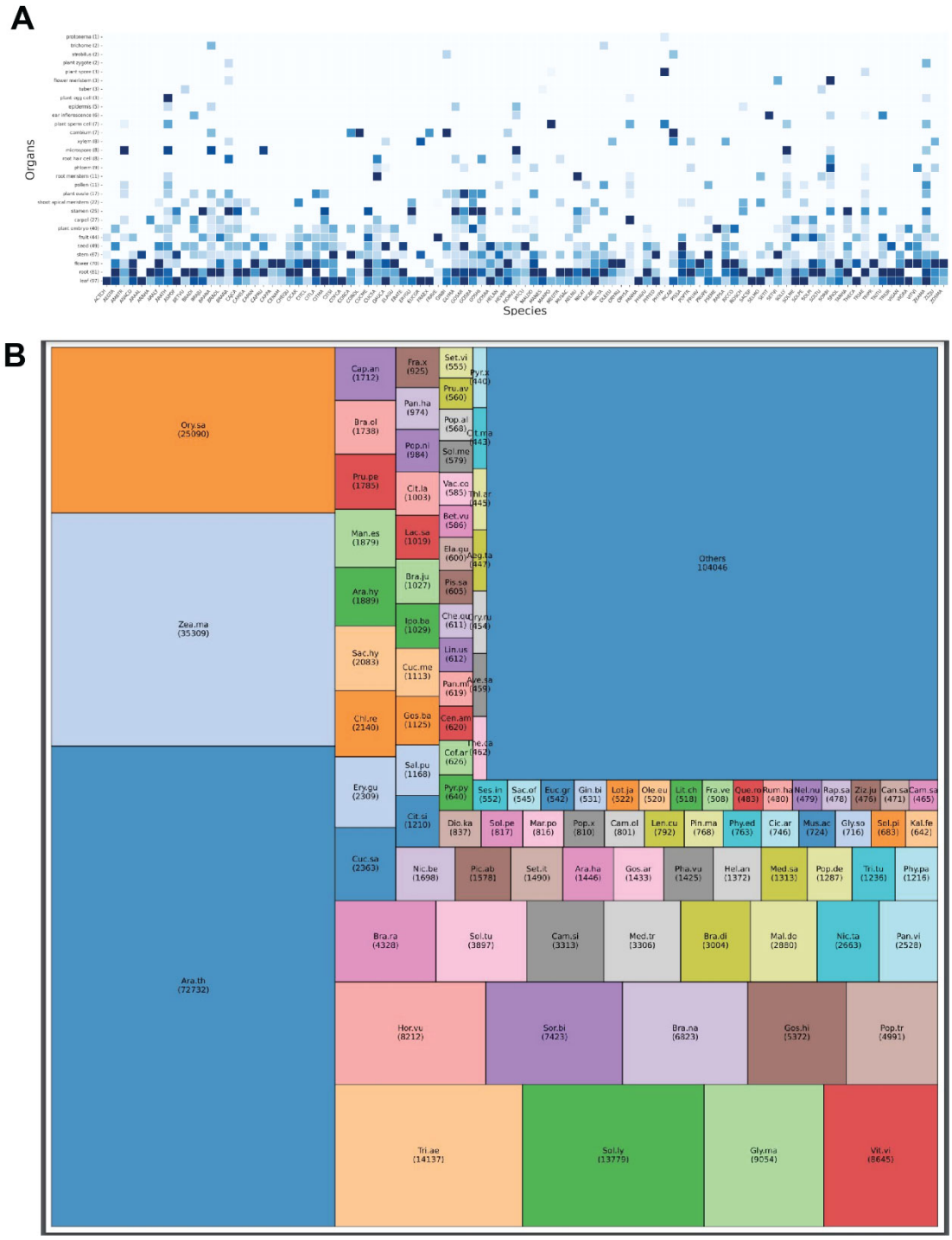
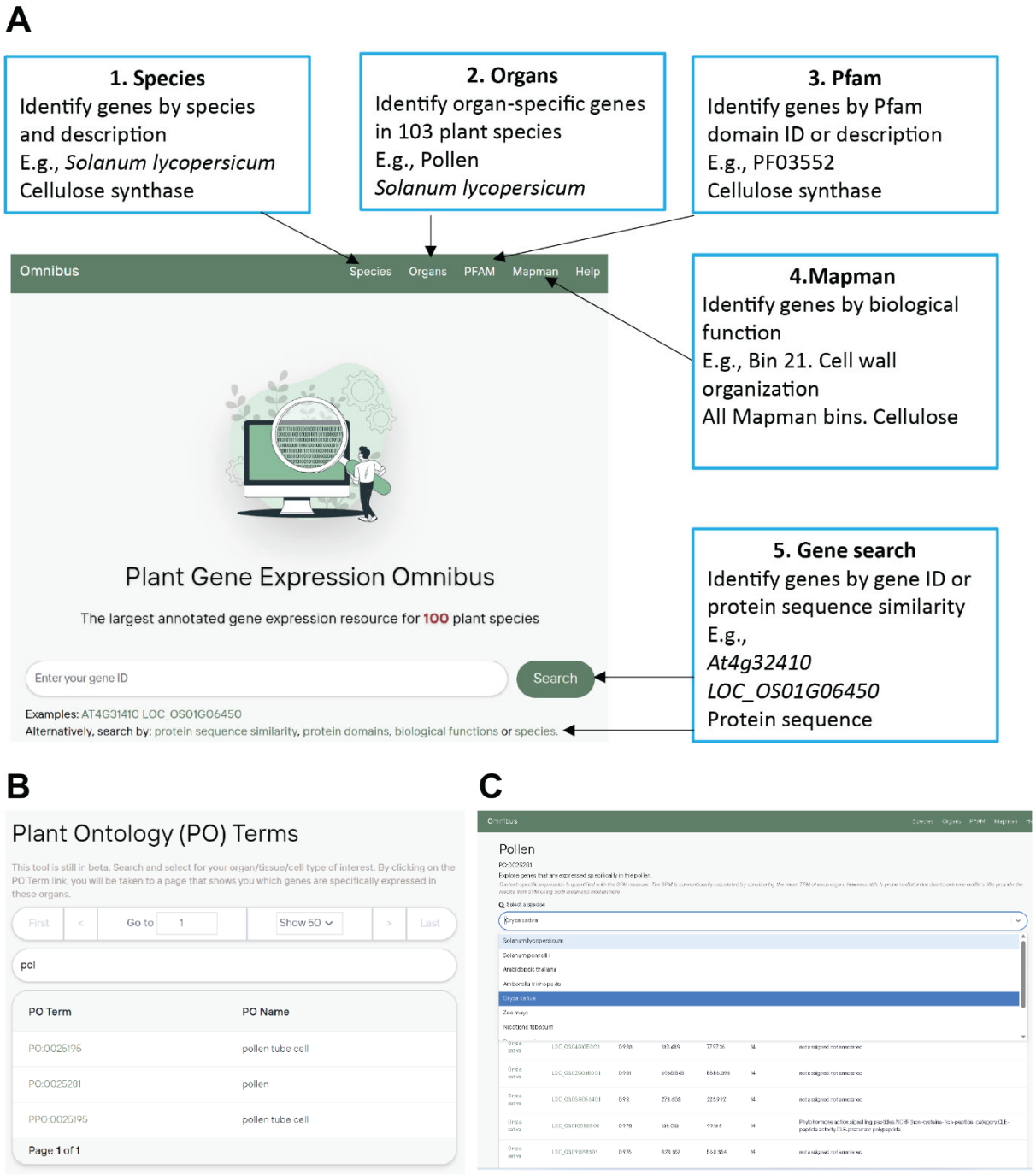


Figure 1: RNA sequencing contents of the Plant Expression Omnibus (PEO) database. A) The heatmap shows the relative abundance of RNA-seq data for the different species (columns) and organs, tissues, and cell types (rows). The numbers in paratheses in rows indicate the number of species that have data for a given plant structure. The cell opacity indicates the relative abundance of RNA-seq samples across

130 organs for each species. B) The treemap showing the relative number of RNA-seq
131 samples for the different species. The areas of the rectangles correspond to the number
132 of samples, which is given in parentheses.
133
134 **Results and discussion**
135 *Querying the Plant Expression Omnibus (PEO)*
136 PEO is an interlinked database that includes 103 species, 4.6 million genes, 5,760
137 Mapman terms, 7,566 Pfam terms and 59,856 quality-controlled samples, of which
138 41,592 are annotated with 78 PO terms. To our knowledge, no species-wide gene
139 expression database exists for many of these 103 species. The distribution of data
140 illustrating species composition and organ specificity are presented as a treemap and
141 heatmap, respectively (Figure 1). Not surprisingly, leaves were the most sampled plant
142 organs (97 species contain leaf samples), followed by roots (81), flowers (70), stems (67)
143 and seeds (49)(Figure 1A). The less represented samples represent structures not found
144 in all flowering plants (strobilus, tuber, protonema, spore) or samples that are difficult to
145 collect (trichome, zygote, egg cell). *Arabidopsis thaliana* had the highest number of RNA-
146 seq samples (Ara.th, 72,732), followed by maize (35,309), rice (25,090), wheat (14,137)
147 and other crops (Figure 1B), with the next model organism *Brachypodium distachyon* on
148 the 17th place with 3,004 samples. The annotation depth of Mapman, Pfam and Plant
149 Ontology annotations are also provided (Supplemental Figure S2). The data are sorted
150 by several modes, such as species type, organ specificity, as well as Mapman and Pfam
151 annotation. Such an arrangement allows the user to query for gene expression in specific
152 organs of interest or look for orthologous genes in non-model species. Sequence-based
153 BLAST and annotation-based Pfam searches can be used to identify homologous and
154 orthologous genes of interest across the database. Together with Mapman annotation
155 and organ specificity predictions, putative functions can be ascribed to gene orthologues.
156 Finally, PEO also provides a list of top co-expressed genes for any gene of interest for all
157 species present in the database, allowing for the discovery of specific metabolic or

158 developmental pathways that are spatiotemporally regulated.



159

160 **Figure 2: Frontend of Gene Expression Omnibus.** A) Genes of interest are searchable
161 via GeneID, protein sequence, Pfam domains, and species as query entries. Browsing of
162 various databases is available in the above tabs. B) Example search by organ type
163 (Pollen). C) The output displays a list of genes known to be present in a specific organ of
164 interest, searchable by species.

165 The home page of PEO consists of several search modes. Queries can be based
166 on specific GeneID or protein sequences. It is also possible to browse the database via
167 the 'Species', 'Organs', 'Pfam', or 'Mapman' annotations (Figure 2). For example, we can
168 select the 'Organs' tab, which provides a list of Plant Ontology (PO) terms. By selecting
169 'Pollen' as our organ of interest, we arrive at a page displaying a set of transcripts sorted
170 by SPM. This specificity measurement ranges between zero and one, where one
171 indicates the gene is exclusively expressed in the tissue (Julca et al., 2021). The page
172 also provides information on the Median and Mean Transcripts per Million (TPM) values
173 used to calculate SPM value and the number of samples the calculations are based on.
174 Mapman annotations are also listed for individual transcripts to provide information on
175 putative gene functions. This database can also be searched by species to obtain the
176 pollen-specific genes in other species present in the PEO.
177 As another example of a query strategy, we also show an example search using
178 the protein sequence of the Arabidopsis gene ABORTED MICROSPORES
179 (*AT2G16910*), a transcription factor known to be associated with the regulation of pollen
180 formation (Xu et al., 2014). The search output provides a list of genes sorted by %identity,
181 with further information on E-values, bit scores, number of mismatches, and gap
182 openings. The gene list includes matched genes from the other species present in PEO,
183 providing a means to discover gene orthologues from less common organisms
184 (Supplemental Figure S3).
185 Clicking on a gene of interest in the list (AMS, *AT2G16910*), brings us to a gene
186 page containing relevant information on this gene (Figure 3). The top of the page provides
187 information about the GeneID, Species name, and TaxonID. The next section displays
188 the putative organ specificity of the gene of interest based on SPM calculations. Details
189 on the values that the SPM calculation is based on is present in the grey subtext below
190 the SPM value. The top organs are reported with mean TPM, standard deviation, and
191 median TPM, showing the magnitude of expression. Since genes with low TPM values
192 can appear to be specifically expressed due to noisy expression, the mean TPM values
193 can be used to flag such genes (Proost and Mutwil, 2018). The sample size is also
194 reported. Together, these support user interpretation of whether the gene displays organ-
195 specific expression. The data can be visualized in a bar chart or box/scatterplot format.

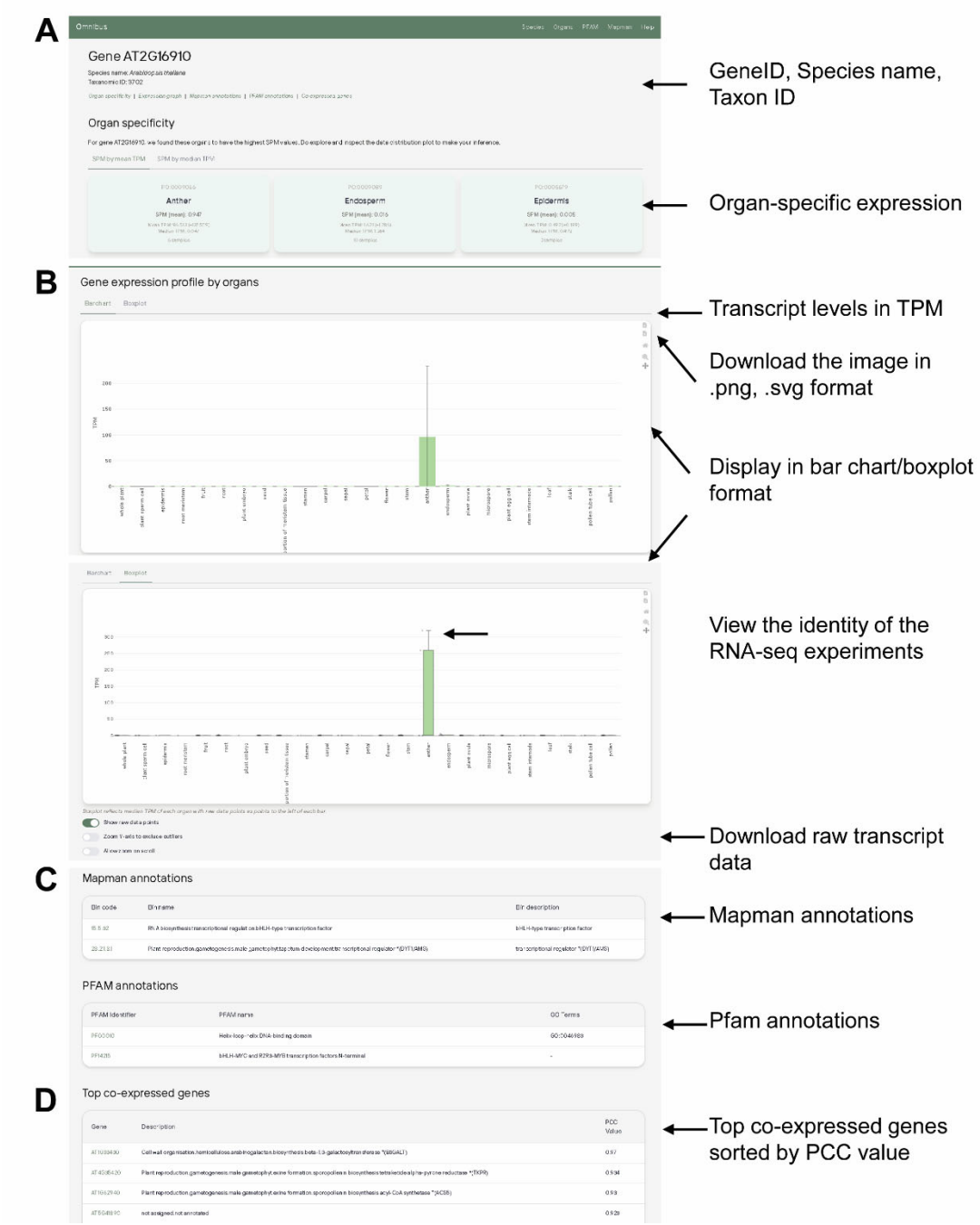
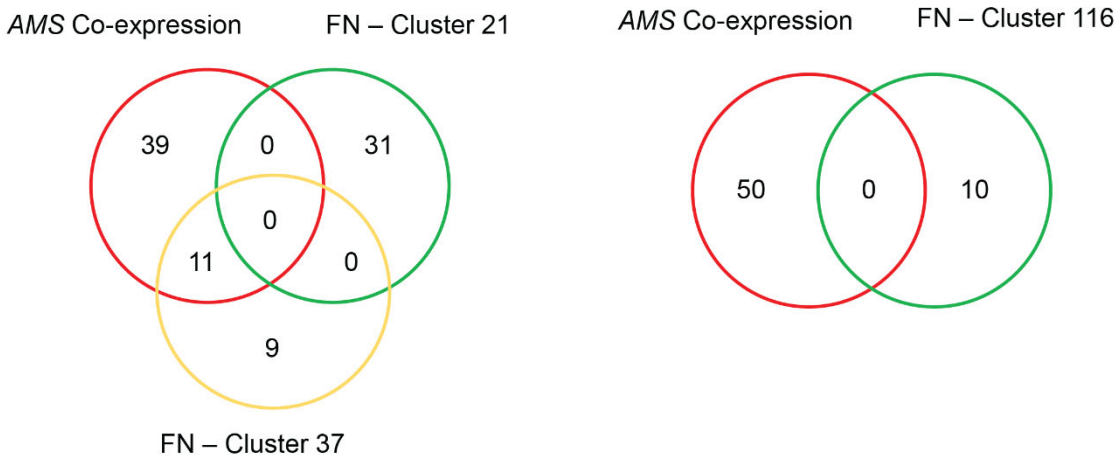


Figure 3: Example gene page of the gene of interest (*AT2G16910*). A) The gene page provides organ specificity of the gene predicted by the SPM score. B) It also displays relative transcript expression in various plant organs/tissues as a box plot or bar chart. Moving the mouse pointer over the points in the box plot shows the corresponding experiment ID. C) Mapman and Pfam annotations for the gene of interest are displayed

202 to provide information on gene function. D) A list of top co-expressed genes sorted by
203 PCC value.
204 The boxplots depict quartiles and whiskers (hover to reveal values), indicating the skew
205 of data. PEO allows for automatic rescaling to exclude the outliers for the y-axis (Figure
206 3). Manual zooming and panning is also available by the control bar at the right of each
207 chart. The charts are available to download in .png or .svg image formats, with the raw
208 data also available for download in tsv format. Associated Pfam and Mapman annotations
209 are also provided, with the relevant Mapman bin codes and Pfam names and identifiers
210 available for reference. Lastly, a list of top-coexpressed genes sorted by Pearson
211 Correlation Coefficient (PCC) is provided for the user.
212
213 *Use case 1: Identification of genes involved in pollen coat biosynthesis*
214 Here, we provide a case study on how PEO can be used to identify candidate genes in
215 pollen coat biosynthesis by using PEO and correlation with other evidence available in
216 the literature. As a starting point, we will use ABORTED MICROSPORES (AMS,
217 *AT2G16910*), a transcription factor known to regulate pollen coat biosynthesis (Xu et al.,
218 2014). The pollen coat is a multilayered structure that provides protection for the nascent
219 male gametophytes from external stresses (Scott et al., 2004). The outer layer, called
220 exine, is composed primarily of a polymer called sporopollenin, which is a composite of
221 fatty acids, phenylpropanoids, phenolics, and carotenoids (Ahlers et al., 1999). It was
222 found that the *ams* null mutant failed to produce functional pollen, and further
223 investigations led to its identification as a master regulator of various aspects of pollen
224 biosynthesis (Xu et al., 2014).
225 Using PEO, we first looked at the gene page of AMS, which showed its highly
226 specific expression in the anthers (Figure 4). Next, we obtained the list of top co-
227 expressed genes against *AMS* and used it to compare against gene clusters identified by
228 FlowerNet. FlowerNet is a flower-specific database based on microarray data focusing
229 on stamen-, pollen-, or flower-specific expression (Pearce et al., 2015). By generating a
230 correlation network, FlowerNet generates gene clusters of transcriptionally coregulated
231 genes, and based on analysis of tissue specificity and expression at different pollen
232 maturation stages, we were able to identify several clusters associated with pollen

233 development. Venn diagram analysis showed that AMS co-expressed genes overlapped
234 significantly (55%, p = 1.4E-26) with Cluster 37, representing sporopollenin biosynthesis,
235 but not Clusters 21 and 116, representing pollen exine formation and pollen tube growth,
236 respectively (Figure 4A). This suggests that *AMS* plays a key role in regulating
237 sporopollenin biosynthesis in pollen development.

A



B

Gene ID	Gene name	Function	Expression
AT1G01280	CYP703A2	cytochrome P450	Anther
AT1G02050	LAP6	Chalcone and stilbene synthase family protein	Anther
AT1G62940	ACOS5	acyl-CoA synthetase	Anther
AT1G69500	CYP704B1	cytochrome P450	Anther
AT3G11980	FAR2	fatty acid reductases	Anther
AT3G13220	ABCG26	ATP-binding cassette transporter G26	Anther
AT3G23770	N.A	O-Glycosyl hydrolases family 17 protein	Anther
AT3G42960	ATA1	homolog of TASSELSEED2	Anther
AT4G14080	MEE48	Involved in the formation of the pollen wall	Anther
AT4G34850	N.A	Chalcone and stilbene synthase family protein	Anther
AT4G35420	DRL1	Dihydroflavonol 4-reductase-like1)	Anther

238

239 **Figure 4: Example use case of gene *AMS* (*AT2G16910*), a transcription factor**
240 **known to regulate pollen coat biosynthesis.** A) Gene expression profile by organs

shows its highly specific expression in pollen. Comparison of top co-expressed genes with co-expression clusters extracted from FlowerNet shows strong overlap with Cluster 37 (Sporopollenin biosynthesis) but not Cluster 21 (Pollen exine formation) and Cluster 116 (Pollen tube formation/growth). B) Table shows gene names, functions, and specific expression of genes found in the overlap between AMS co-expression and FlowerNet Cluster 37.

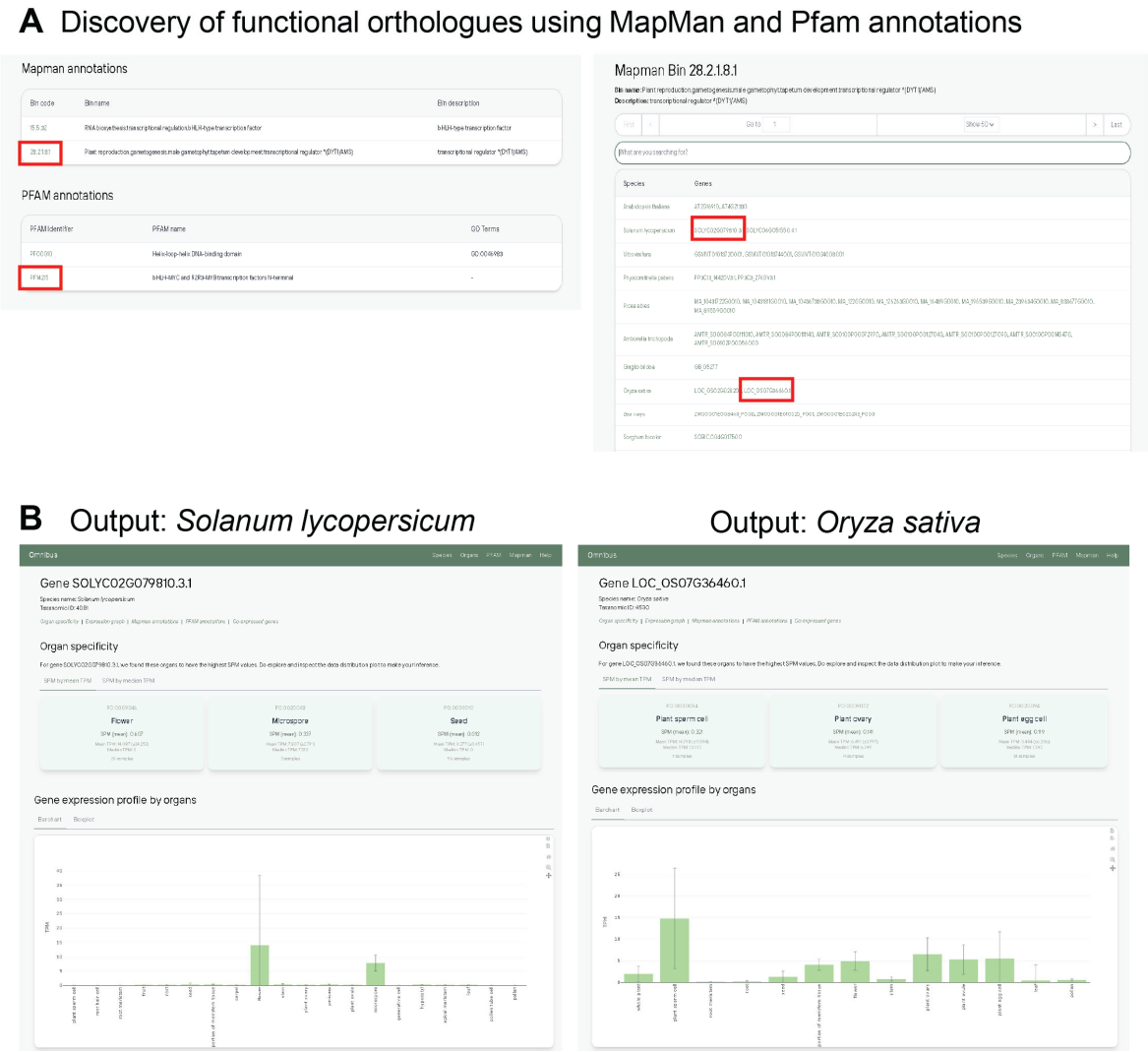
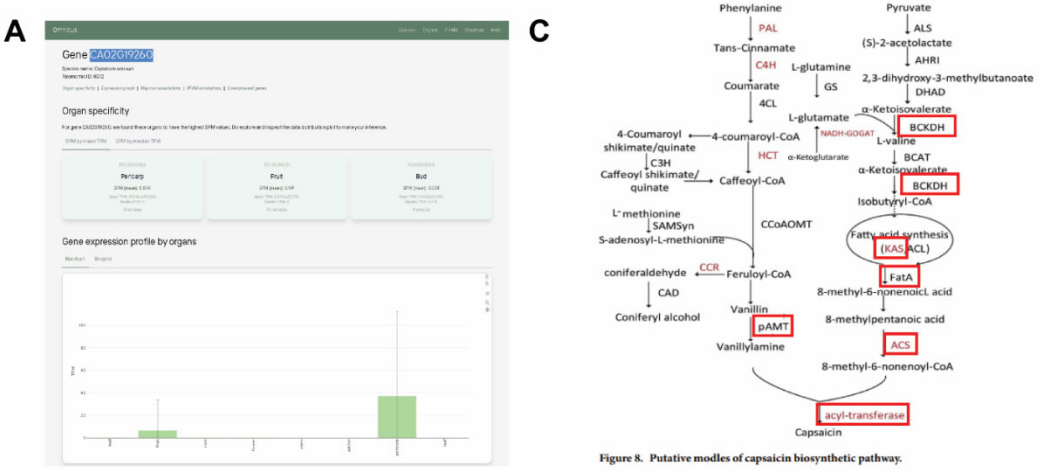
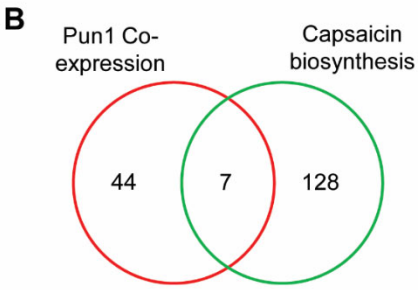


Figure 5: Discovery of functional orthologues using MapMan and Pfam annotations. A) Mapman and Pfam annotations for the gene of interest (*AT2G16910*) provide a list of genes annotated with the bins and domains in other species. B) Example output for functional orthologue discovery using MapMan annotation. The gene of

252 *Solanum lycopersicum* and *Oryza sativa* show specific expression in reproductive tissues
253 similar to *AT2G16910* in *Arabidopsis*.
254
255 Further examination of the overlapped genes (11 genes) revealed a list of genes
256 that were annotated as being involved in pollen development (*AT1G01280*, *AT1G69500*,
257 *AT4G35420*) or whose mutants were essential for male fertility (*AT1G62940*,
258 *AT3G11980*, *AT3G13220*, *AT4G35420*), and were all found to be predominantly
259 expressed in anthers (Figure 4B, Supplemental Dataset S2). This strategy of using
260 complementary databases can be used to narrow down large gene lists and identify key
261 genes involved in a pathway of interest. Finally, using the associated Mapman and Pfam
262 annotations available on the gene page, we can obtain gene orthologues from other
263 species and use that as a starting point to investigate genetic pathways in other plants
264 (Figure 5A). As an example, we examined the genes *SOLYC02G079810.3.1* and
265 *LOC_OS07G36460.1* from *Solanum lycopersicum* and *Oryza sativa*, respectively (Figure
266 5B). Observations of the respective gene loci of *Solanum lycopersicum* and *Oryza sativa*
267 show transcript enrichment in reproductive tissues similar to that in Arabidopsis.
268
269 *Use case 2: Identification of biosynthetic pathway components of capsaicin*
270 Here, we will provide an example of using PEO to investigate capsaicin biosynthesis in
271 the crop plant *Capsicum annuum*. The genus Capsicum comprises a large class of crop
272 plants commonly known as peppers (i.e. chili peppers), and its distinctive
273 pungency/spiciness is due to the alkaloid capsaicin (Suzuki et al., 1980). We looked at
274 the literature and found that *Pun1* (*CA02G19260*) is a necessary enzymatic component
275 of capsaicin biosynthesis. The *Pun1* gene encodes an acyltransferase, and the deletion
276 of the *Pun1* gene in bell peppers is the reason for its distinctive non-pungency compared
277 to pungent varieties like chili peppers (Stewart Jr et al., 2005). We searched for the *Pun1*
278 (*CA02G19260*) gene in our database and observed that it was expressed mainly in fruit
279 and pericarp tissues, consistent with the site of capsaicin accumulation in the plant
280 (Figure 6A). Next, we obtained the list of top co-expressed genes for *Pun1* and compared
281 it against a list of putative capsaicin biosynthetic pathway genes previously obtained by
282 isolation from *Capsicum annuum* fruit tissue at specific developmental stages (Mazourek

et al., 2009; Zhang et al., 2016). This comparison yielded an overlap of 13.7% (7 genes), which we examined individually for their function and organ expression (Figure 6B). We observed that all the genes were predominantly found in the pericarp tissue, consistent with the site of capsaicin accumulation (Figure 6D). We then obtained a putative map of the capsaicin biosynthetic pathway (Zhang et al., 2016) and attempted to map our gene list to it. Surprisingly, we observed that most of the genes mapped to the pyruvate-derived arm of the pathway, suggesting that *Pun1* primarily regulates this arm of capsaicin biosynthesis (Figure 6C).





D

Gene ID	Gene name	Function	Specific expression
CA06G10910	*BCKDH E1a*	Branched-chain amino acid aminotransferase	Pericarp
CA03G08530	*pAMT*	Putative aminotransferase	Pericarp
CA01G00840	*KasI*	Ketoacyl-ACP synthase I	Pericarp
CA11G05940	*AT*	acyltransferase 1	Pericarp
CA08G18150	*ACS*	Acyl-CoA synthetase 2	Flower / Pericarp
CA06G26640	*FatA*	Acyl-ACP thioesterase	Pericarp
CA01G28560	*KasIII*	Ketoacyl-ACP synthase III	Pericarp

Figure 6: Example use case of *Pun1* (*CA02G19260*) from *Capsicum annum*, an important enzymatic component of capsaicin biosynthesis. A) Gene expression profile by organs shows its highly specific expression in fruit and pericarp, which is also

295 the site of capsaicin accumulation. B) Comparison of top co-expressed genes with known
296 capsaicin biosynthetic pathway components yields an overlap of 13.7% (7 genes). C)
297 Mapping of genes to proposed capsaicin biosynthetic pathway based on putative gene
298 function (red boxes). D) Specific expression of the seven genes obtained from PEO.
299

300 **Conclusion**

301 We present here a fully annotated Gene Expression Omnibus of 103 plant species, which
302 provides information on tissue specificity and gene function information of the relevant
303 gene of interest. PEO also automatically provides a list of top-coexpressed genes for
304 better elucidation of gene functions and associated pathways. We provided two case
305 studies of how PEO can be used to study genes and pathways of interest and to narrow
306 down large putative gene lists based on information on tissue specificity, annotated Pfam
307 and Mapman functions, and comparison using complementary datasets obtained from
308 other analyses. These analyses are amenable to any species present in PEO and allow
309 researchers to build a genetic model of a pathway of interest quickly. As we expand the
310 number of species available in PEO, we will be able to more effectively empower
311 researchers working on non-model organisms and more quickly progress the pace of
312 translational research.
313

314 **Acknowledgments**

315 We would like to thank Riccardo Delli Ponti for the initial help in collecting the gene
316 expression data. M.M. is supported by the Singaporean Food Agency grant
317 SFS_RND_SUFP_001_05 and E.K. is supported by Ministry of Education Tier 3 Grant '
318 From Tough Pollen to Soft Matter'.
319

320 **Conflict of interests**

321 The authors declare no conflicts of interest.
322

323 Supplemental Tables:

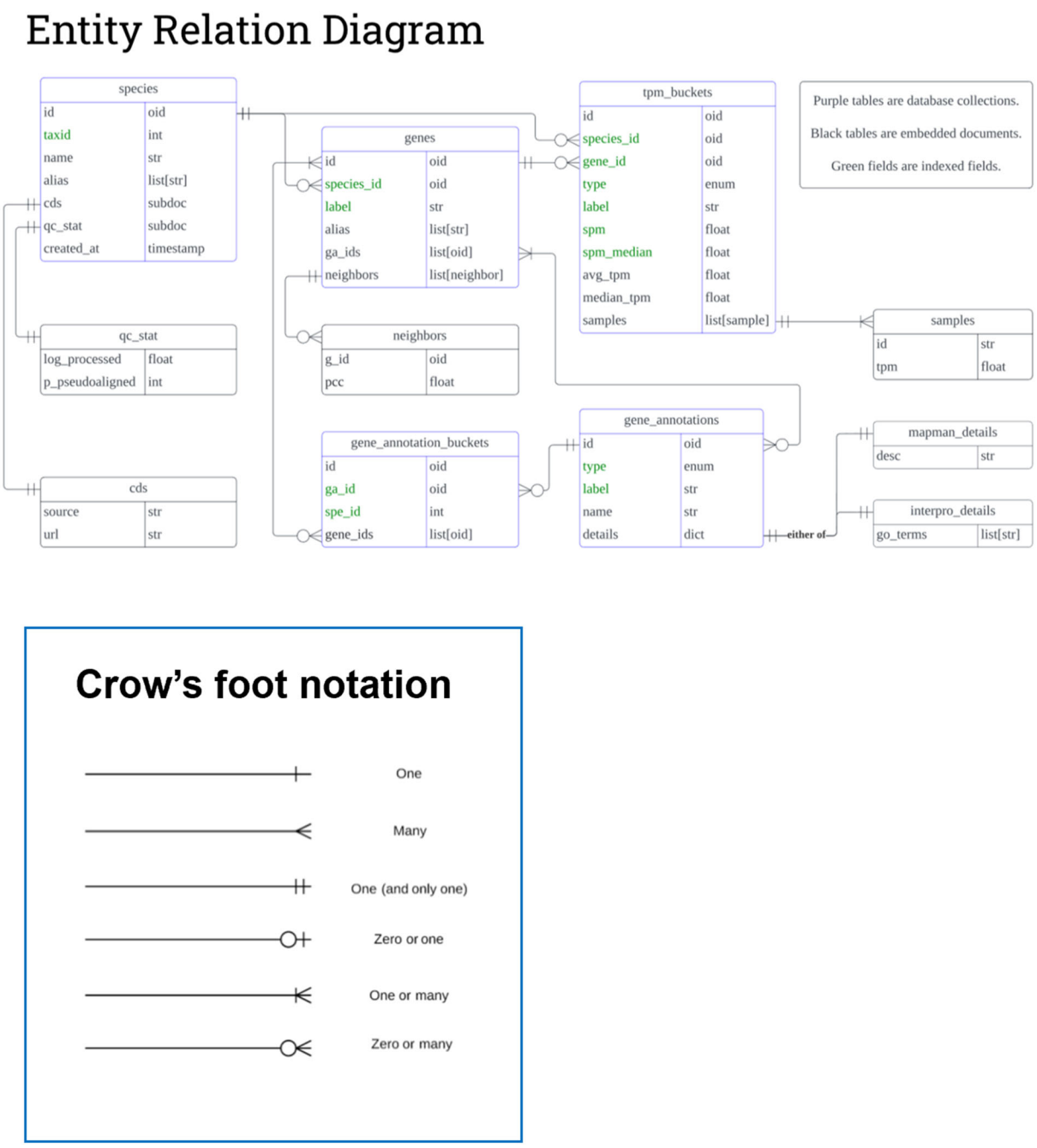
Data	Content	Prepared by
Species list	Includes list of 103 species, with the sources of their CDS.	Manual curation by Mutwil.
TPM matrices	Quantifies gene expression of each gene in each sample. Each species is represented by a 2D TPM matrix with genes along the rows and samples along the columns.	TPM matrices were processed and compiled from other projects (Goh & Mutwil, 2021; Julca et al., 2021). Villanueva and Mutwil performed quality control.
Organ annotations	Assignment of each RNA-seq sample to a PO term and name.	Manual curation and annotation by Villanueva et. al. (2022, unpublished work).
Mapman annotations	Assignment of each gene to at least one Mapman bin label and name.	Performed in this study by Goh.
Pfam annotations (Interpro)	Assignment of each gene to Pfam domains, which in turn have a name and their associated GO terms.	Manual curation and annotation by Villanueva et. al. (2022, unpublished work).
PEP files	Protein sequence files, required for building the diamond reference db. Diamond (Buchfink et al., 2015) is an alternative to BLAST for fast protein sequence search.	Raw pep files obtained by Villanueva, re-processed by Goh. Diamond protein sequence database was processed by Goh.

Supplemental Table S1: Raw data collected for PEO

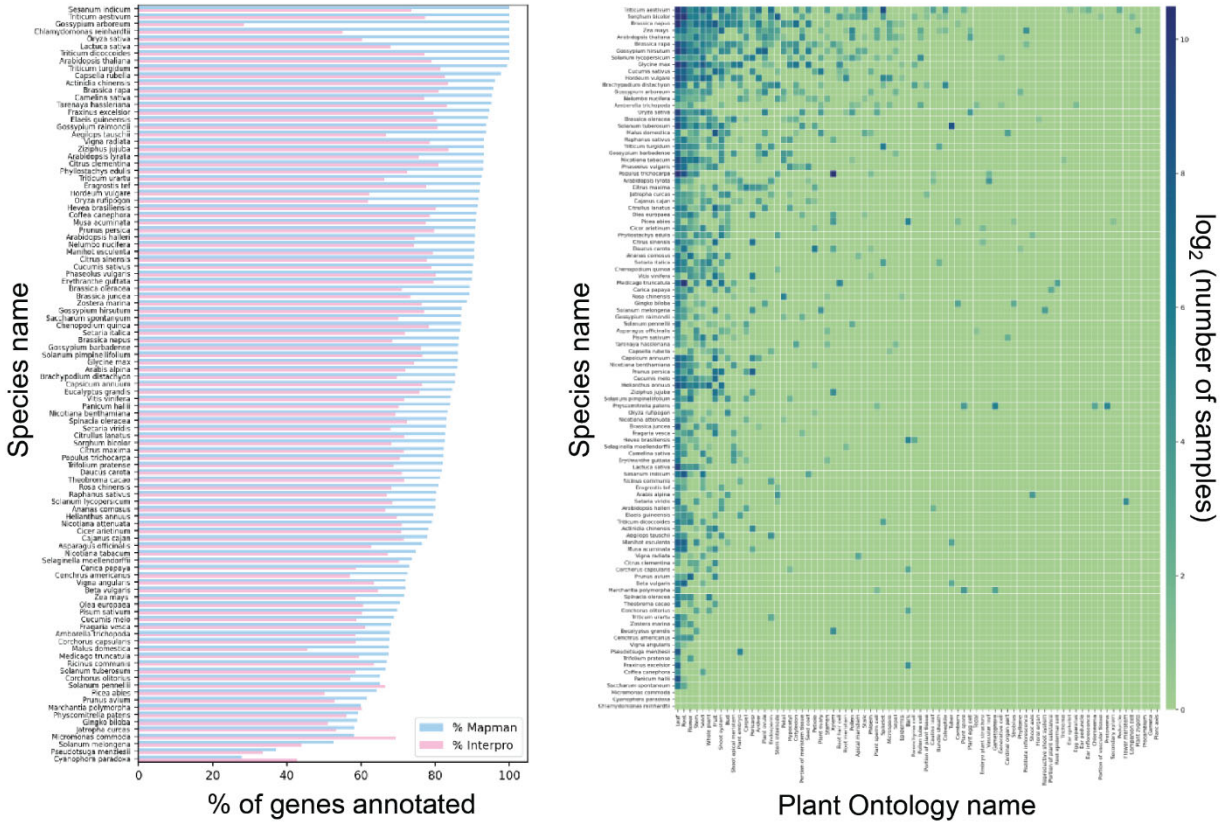
File type	Columns required	Number of files	Constraints
Species list	Taxonomic ID, scientific name, CDS source, CDS url	1 globally	-
TPM matrices	Gene IDs as row index, sample labels as column headers, TPM as values	1 per species	Species must be specified in species list
Gene annotations	Gene annotation label, name any other fields (flexible)	1 per annotation type globally (currently 2, for Mapman and Interpro)	-
Gene annotations assignment	Gene ID, Gene annotation label	1 per species, per annotation type	Genes must exist in TPM matrices, gene annotation must exist in globally defined file
Sample annotations	JSON mapping of sample annotation label (PO terms) to the name of each PO term	1 per sample annotation type globally (1 for PO currently)	-
Sample annotations assignment	Sample label, Sample annotation label (PO term)	1 per species, per sample annotation type	Samples must exist in TPM matrices, sample annotation must exist in globally defined file

Supplemental Table S2: Processed data files uploaded to PEO

Supplemental Figures

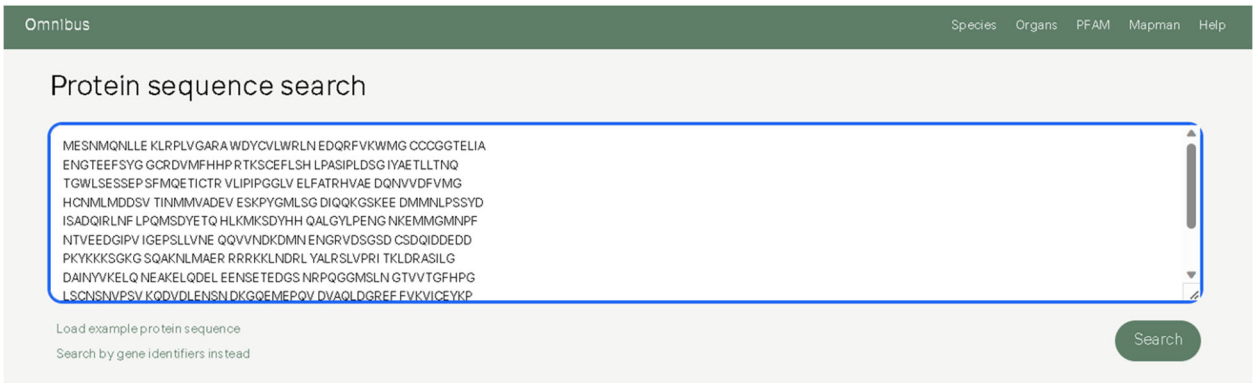


Supplemental Figure S1: Architectural design of PEO web-tool

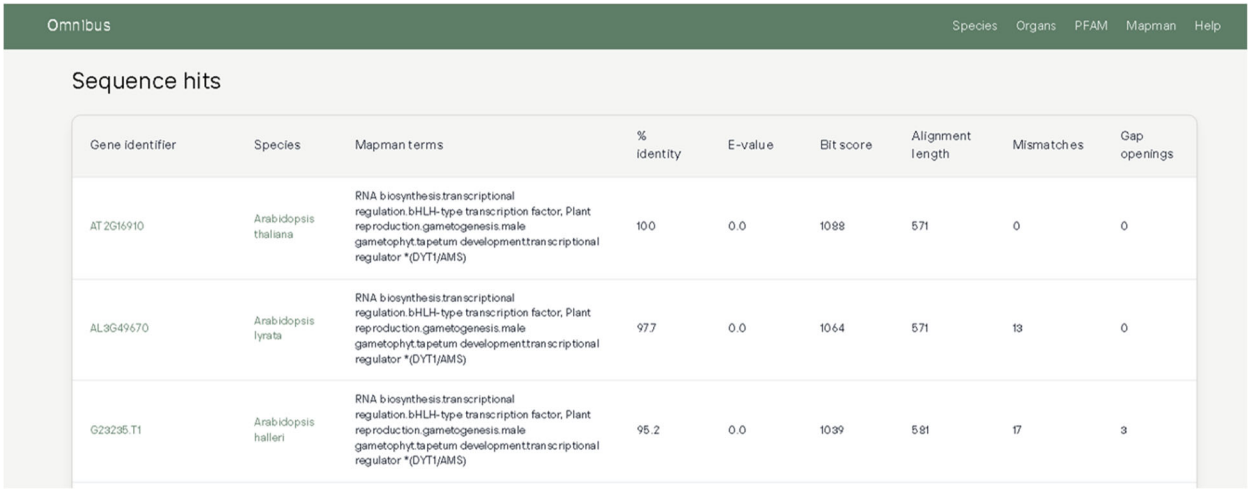


Supplemental Figure S2: Data distribution of PEO showing Mapman, Interpro and Plant Ontology annotation coverage.

Example: Input protein sequence (AT2G16910)



Output: Hits sortable by species and % identity



Gene identifier	Species	Mapman terms	% identity	E-value	Bit score	Alignment length	Mismatches	Gap openings
AT2G16910	Arabidopsis thaliana	RNA biosynthesis.transcriptional regulation.bHLH-type transcription factor, Plant reproduction.gametogenesis.male gametophyt.tapetum development.transcriptional regulator *(DYT1/AMS)	100	0.0	1088	571	0	0
AL3G49670	Arabidopsis lyrata	RNA biosynthesis.transcriptional regulation.bHLH-type transcription factor, Plant reproduction.gametogenesis.male gametophyt.tapetum development.transcriptional regulator *(DYT1/AMS)	97.7	0.0	1064	571	13	0
G23235.T1	Arabidopsis halleri	RNA biosynthesis.transcriptional regulation.bHLH-type transcription factor, Plant reproduction.gametogenesis.male gametophyt.tapetum development.transcriptional regulator *(DYT1/AMS)	95.2	0.0	1039	581	17	3

Supplemental Figure S3: Example search output for protein sequence. Search output retrieves and sorts candidate orthologous genes from other species available in the database.

References

Ahlers, F., Lambert, J., Wiermann, R., 1999. Structural Elements of Sporopollenin from the Pollen of Torreya californica Torr. (Gymnospermae): Using the 1H-NMR Technique. Zeitschrift für Naturforschung C 54, 492–495. https://doi.org/10.1515/znc-1999-7-806

Bray, N.L., Pimentel, H., Melsted, P., Pachter, L., 2016. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527. https://doi.org/10.1038/nbt.3519

348 Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using
349 DIAMOND. Nat. Methods 12, 59–60. https://doi.org/10.1038/nmeth.3176
350
351 Cravens, A., Payne, J., Smolke, C.D., 2019. Synthetic biology strategies for microbial
352 biosynthesis of plant natural products. Nat. Commun. 10, 2142.
353
354 Delli-Ponti, R., Shivhare, D., Mutwil, M., 2021. Using Gene Expression to Study
355 Specialized Metabolism—A Practical Guide. Front. Plant Sci. 11, 2074.
356 https://doi.org/10.3389/fpls.2020.625035
357
358 Desborough, M.J., Keeling, D.M., 2017. The aspirin story–from willow to wonder drug. Br.
359 J. Haematol. 177, 674–683.
360
361 Goh, W., Mutwil, M., 2021. LSTrAP-Kingdom: an automated pipeline to generate
362 annotated gene expression atlases for kingdoms of life. Bioinforma. Oxf. Engl. btab168.
363 https://doi.org/10.1093/bioinformatics/btab168
364
365 Hansen, B.O., Meyer, E.H., Ferrari, C., Vaid, N., Movahedi, S., Vandepoele, K., Nikoloski,
366 Z., Mutwil, M., 2018. Ensemble gene function prediction database reveals genes
367 important for complex I formation in Arabidopsis thaliana. New Phytol. 217, 1521–1534.
368 https://doi.org/10.1111/nph.14921
369
370 Julca, I., Ferrari, C., Flores-Tornero, M., Proost, S., Lindner, A.-C., Hackenberg, D.,
371 Steinbachová, L., Michaelidis, C., Gomes Pereira, S., Misra, C.S., Kawashima, T., Borg,
372 M., Berger, F., Goldberg, J., Johnson, M., Honys, D., Twell, D., Sprunck, S., Dresselhaus,
373 T., Becker, J.D., Mutwil, M., 2021. Comparative transcriptomic analysis reveals
374 conserved programmes underpinning organogenesis and reproduction in land plants.
375 Nat. Plants 7, 1143–1159. https://doi.org/10.1038/s41477-021-00958-2
376

377 Klie, S., Nikoloski, Z., 2012. The choice between MapMan and Gene ontology for
378 automated gene function prediction in plant science. Front. Genet. 3, 1–14.
379 https://doi.org/10.3389/fgene.2012.00115
380 Krishna, S., Bustamante, L., Haynes, R.K., Staines, H.M., 2008. Artemisinins: their
381 growing importance in medicine. Trends Pharmacol. Sci. 29, 520–527.
382
383 Lim, P.K., Zheng, X., Goh, J.C., Mutwil, M., 2022. Exploiting plant transcriptomic
384 databases: Resources, tools, and approaches. Plant Commun. 100323.
385
386 Ma, L., Sun, N., Liu, X., Jiao, Y., Zhao, H., Deng, X.W., 2005. Organ-specific expression
387 of Arabidopsis genome during development. Plant Physiol. 138, 80–91.
388
389 Mazourek, M., Pujar, A., Borovsky, Y., Paran, I., Mueller, L., Jahn, M.M., 2009. A Dynamic
390 Interface for Capsaicinoid Systems Biology. Plant Physiol. 150, 1806–1821.
391 https://doi.org/10.1104/pp.109.136549
392
393 Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer,
394 E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D., Bateman, A.,
395 2021. Pfam: The protein families database in 2021. Nucleic Acids Res. 49, D412–D419.
396 https://doi.org/10.1093/nar/gkaa913
397
398 Niazian, M., 2019. Application of genetics and biotechnology for improving medicinal
399 plants. Planta 249, 953–973.
400
401 Nowicka, B., Ciura, J., Szymańska, R., Kruk, J., 2018. Improving photosynthesis, plant
402 productivity and abiotic stress tolerance–current trends and future perspectives. J. Plant
403 Physiol. 231, 415–433.
404
405 Paddon, C.J., Westfall, P.J., Pitera, D.J., Benjamin, K., Fisher, K., McPhee, D., Leavell,
406 M., Tai, A., Main, A., Eng, D., 2013. High-level semi-synthetic production of the potent
407 antimalarial artemisinin. Nature 496, 528–532.

408

409 Pearce, S., Ferguson, A., King, J., Wilson, Z.A., 2015. FlowerNet: A Gene Expression
410 Correlation Network for Anther and Pollen Development. Plant Physiol. 167, 1717–1730.
411 https://doi.org/10.1104/pp.114.253807

412

413 Proost, S., Mutwil, M., 2018. CoNekT: An open-source framework for comparative
414 genomic and transcriptomic network analyses. Nucleic Acids Res. 46, W133–W140.
415 https://doi.org/10.1093/nar/gky336

416

417 Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., Wittkop, T., Sokolov, A., Graim,
418 K., Funk, C., Verspoor, K., Ben-Hur, A., Pandey, G., Yunes, J.M., Talwalkar, A.S., Repo,
419 S., Souza, M.L., Piovesan, D., Casadio, R., Wang, Z., Cheng, J., Fang, H., Gough, J.,
420 Koskinen, P., Törönen, P., Nokso-Koivisto, J., Holm, L., Cozzetto, D., Buchan, D.W.A.,
421 Bryson, K., Jones, D.T., Limaye, B., Inamdar, H., Datta, A., Manjari, S.K., Joshi, R.,
422 Chitale, M., Kihara, D., Lisewski, A.M., Erdin, S., Venner, E., Lichtarge, O., Rentzsch, R.,
423 Yang, H., Romero, A.E., Bhat, P., Paccanaro, A., Hamp, T., Kaßner, R., Seemayer, S.,
424 Vicedo, E., Schaefer, C., Achten, D., Auer, F., Boehm, A., Braun, T., Hecht, M., Heron,
425 M., Hönigschmid, P., Hopf, T.A., Kaufmann, S., Kiening, M., Krompass, D., Landerer, C.,
426 Mahlich, Y., Roos, M., Björne, J., Salakoski, T., Wong, A., Shatkay, H., Gatzmann, F.,
427 Sommer, I., Wass, M.N., Sternberg, M.J.E., Škunca, N., Supek, F., Bošnjak, M., Panov,
428 P., Džeroski, S., Šmuc, T., Kourmpetis, Y.A.I., Van Dijk, A.D.J., Ter Braak, C.J.F., Zhou,
429 Y., Gong, Q., Dong, X., Tian, W., Falda, M., Fontana, P., Lavezzo, E., Di Camillo, B.,
430 Toppo, S., Lan, L., Djuric, N., Guo, Y., Vucetic, S., Bairoch, A., Linial, M., Babbitt, P.C.,
431 Brenner, S.E., Orengo, C., Rost, B., Mooney, S.D., Friedberg, I., 2013. A large-scale
432 evaluation of computational protein function prediction. Nat. Methods 10, 221–227.
433 https://doi.org/10.1038/nmeth.2340

434 Rao, X., Dixon, R.A., 2019. Co-expression networks for plant biology: why and how. Acta
435 Biochim. Biophys. Sin. 51, 981–988. https://doi.org/10.1093/abbs/gmz080

436

437 Rhee, S.Y., Mutwil, M., 2014. Towards revealing the functions of all genes in plants.
438 Trends Plant Sci. 19, 212–221. https://doi.org/10.1016/j.tplants.2013.10.006

Ryngajllo, M., Childs, L., Lohse, M., Giorgi, F.M., Lude, A., Selbig, J., Usadel, B., 2011. SLocX: predicting subcellular localization of Arabidopsis proteins leveraging gene expression data. Front. Plant Sci. 43.

Sabzehzari, M., Zeinali, M., Naghavi, M.R., 2020. Alternative sources and metabolic engineering of Taxol: Advances and future perspectives. Biotechnol. Adv. 43, 107569.

Schwacke, R., Ponce-Soto, G.Y., Krause, K., Bolger, A.M., Arsova, B., Hallab, A., Gruden, K., Stitt, M., Bolger, M.E., Usadel, B., 2019. MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. Mol. Plant 12, 879–892. https://doi.org/10.1016/j.molp.2019.01.003

Scott, R.J., Spielman, M., Dickinson, H.G., 2004. Stamen Structure and Function. Plant Cell 16, S46–S60. https://doi.org/10.1105/tpc.017012

Stewart Jr, C., Kang, B.-C., Liu, K., Mazourek, M., Moore, S.L., Yoo, E.Y., Kim, B.-D., Paran, I., Jahn, M.M., 2005. The Pun1 gene for pungency in pepper encodes a putative acyltransferase. Plant J. 42, 675–688. https://doi.org/10.1111/j.1365-313X.2005.02410.x

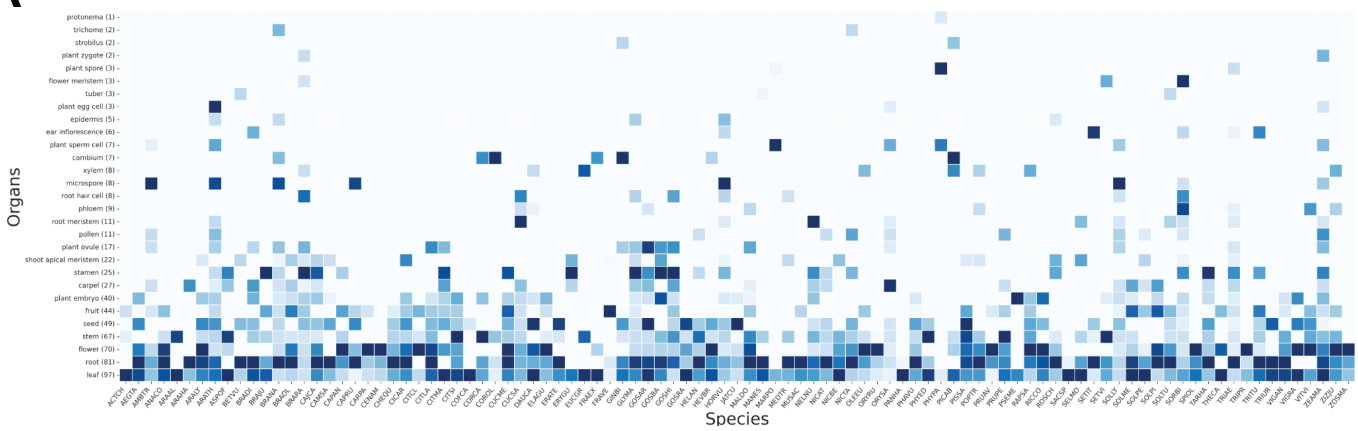
Suzuki, T., Fujiwake, H., Iwai, K., 1980. Intracellular localization of capsaicin and its analogues, capsaicinoid, in Capsicum fruit 1. Microscopic investigation of the structure of the placenta of Capsicum annuum var. annuum cv. Karayatsubusa1. Plant Cell Physiol. 21, 839–853. https://doi.org/10.1093/oxfordjournals.pcp.a076058

Tan, Q.W., Goh, W., Mutwil, M., 2020. LSTrAP-Cloud: A User-Friendly Cloud Computing Pipeline to Infer Co-expression Networks. Genes 11, 428. https://doi.org/10.3390/genes11040428

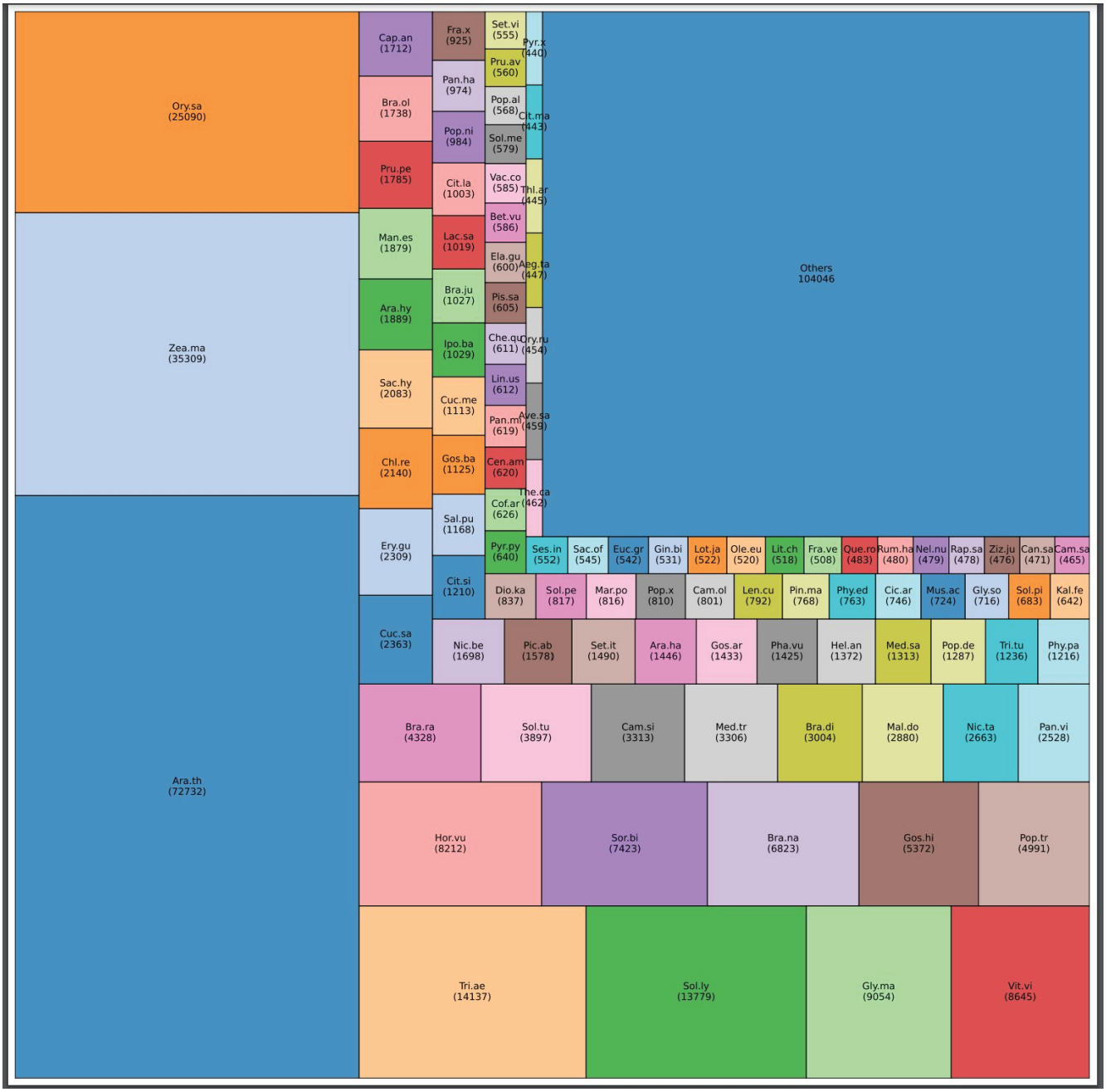
Walls, R.L., Cooper, L., Elser, J., Gandolfo, M.A., Mungall, C.J., Smith, B., Stevenson, D.W., Jaiswal, P., 2019. The Plant Ontology Facilitates Comparisons of Plant

470 Development Stages Across Species. Front. Plant Sci. 10.
471 https://doi.org/10.3389/fpls.2019.00631
472
473 Wetterstrand, K.A., 2013. DNA sequencing costs: data from the NHGRI Genome
474 Sequencing Program (GSP). Natl. Hum. Genome Res. Inst.
475
476 Xu, J., Ding, Z., Vizcay-Barrena, G., Shi, J., Liang, W., Yuan, Z., Werck-Reichhart, D.,
477 Schreiber, L., Wilson, Z.A., Zhang, D., 2014. ABORTED MICROSPORES Acts as a
478 Master Regulator of Pollen Wall Formation in Arabidopsis. Plant Cell 26, 1544–1556.
479 https://doi.org/10.1105/tpc.114.122986
480
481 Zhang, Z.-X., Zhao, S.-N., Liu, G.-F., Huang, Z.-M., Cao, Z.-M., Cheng, S.-H., Lin, S.-S.,
482 2016. Discovery of putative capsaicin biosynthetic genes by RNA-Seq and digital gene
483 expression analysis of pepper. Sci. Rep. 6, 34121. https://doi.org/10.1038/srep34121
484
485

A


Organs
protonema (1)
trichome (2)
strobilus (2)
plant zygote (2)
plant spore (3)
flower meristem (3)
tuber (3)
plant egg cell (3)
epidermis (5)
ear inflorescence (6)
plant sperm cell (7)
cambium (7)
xylem (8)
microspore (8)
root hair cell (8)
phloem (9)
root meristem (11)
pollen (11)
plant ovule (17)
shoot apical meristem (22)
stamen (25)
carpel (27)
plant embryo (40)
fruit (44)
seed (49)
stem (67)
flower (70)
root (81)
leaf (97)
Species

B

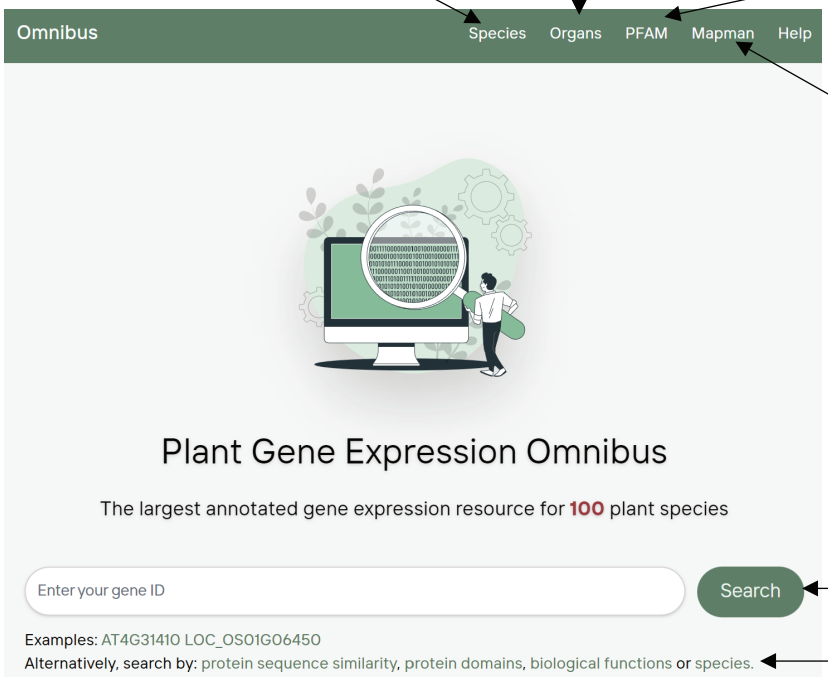

Ory.sa (25090)
Zea.ma (35309)
Ara.th (72732)
Cap.an (1712)
Bra.ol (1738)
Pru.pe (1785)
Man.es (1879)
Ara.hy (1889)
Sac.hy (2083)
Chl.re (2140)
Ery.gu (2309)
Cuc.sa (2363)
Fra.x (925)
Pan.ha (974)
Pop.ni (984)
Cit.la (1003)
Lac.sa (1019)
Bra.ju (1027)
Ipo.ba (1029)
Cuc.me (1113)
Gos.ba (1125)
Sal.pu (1168)
Cit.si (1210)
Set.vi (555)
Pru.av (560)
Pop.al (568)
Sol.me (579)
Vac.co (585)
Bet.vu (586)
Ela.gu (600)
Pis.sa (605)
Che.qu (611)
Lin.us (612)
Pan.mi (619)
Cen.am (620)
Cof.ar (626)
Pyr.x (440)
Cit.ma (443)
Thl.ar (445)
Aeg.ta (447)
Ory.ru (454)
Ave.sa (459)
The.ca (462)
Others 104046
Pyr.py (640)
Ses.in (552)
Sac.of (545)
Euc.gr (542)
Gin.bi (531)
Lot.ja (522)
Ole.eu (520)
Lit.ch (518)
Fra.ve (508)
Que.ro (483)
Rum.ha (480)
Nel.nu (479)
Rap.sa (478)
Ziz.ju (476)
Can.sa (471)
Cam.sa (465)
Dio.ka (837)
Sol.pe (817)
Mar.po (816)
Pop.x (810)
Cam.ol (801)
Len.cu (792)
Pin.ma (768)
Phy.ed (763)
Cic.ar (746)
Mus.ac (724)
Gly.so (716)
Sol.pi (683)
Kal.fe (642)
Nic.be (1698)
Pic.ab (1578)
Set.it (1490)
Ara.ha (1446)
Gos.ar (1433)
Pha.vu (1425)
Hel.an (1372)
Med.sa (1313)
Pop.de (1287)
Tri.tu (1236)
Phy.pa (1216)
Bra.ra (4328)
Sol.tu (3897)
Cam.si (3313)
Med.tr (3306)
Bra.di (3004)
Mal.do (2880)
Nic.ta (2663)
Pan.vi (2528)
Hor.vu (8212)
Sor.bi (7423)
Bra.na (6823)
Gos.hi (5372)
Pop.tr (4991)
Tri.ae (14137)
Sol.ly (13779)
Gly.ma (9054)
Vit.vi (8645)

A

1. Species
Identify genes by species and description
E.g., *Solanum lycopersicum*
Cellulose synthase

2. Organs
Identify organ-specific genes in 103 plant species
E.g., Pollen
Solanum lycopersicum

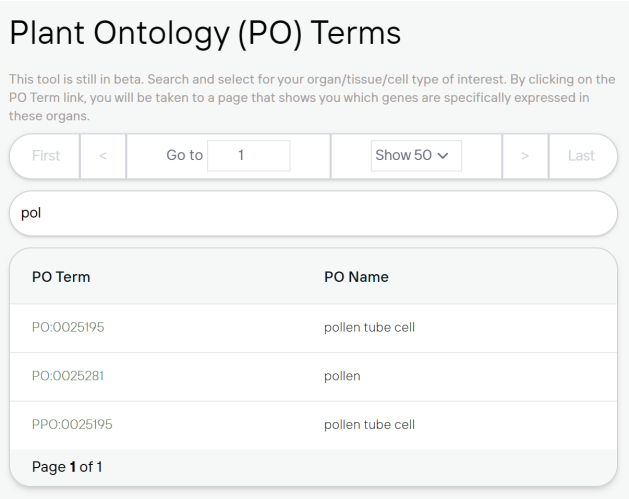
3. Pfam
Identify genes by Pfam domain ID or description
E.g., PF03552
Cellulose synthase



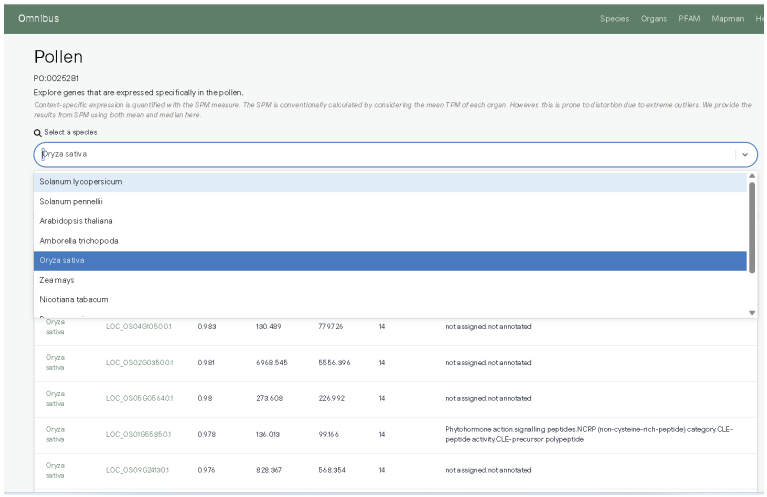
4.Mapman
Identify genes by biological function
E.g., Bin 21. Cell wall organization
All Mapman bins. Cellulose

5. Gene search
Identify genes by gene ID or protein sequence similarity
E.g.,
At4g32410
LOC_OS01G06450
Protein sequence

B



C



A

Omnibus — Species | Organs | PFAM | Mapman | Help

Gene AT2G16910

Species name: *Arabidopsis thaliana*
Taxanomic ID: 3702

Organ specificity | Expression graph | Mapman annotations | PFAM annotations | Co-expressed genes

Organ specificity

For gene AT2G16910, we found these organs have the highest SPM values. Do explore and inspect the data distribution plot to make your inference.

SPM by mean TPM | SPM by median TPM

PO:0009066	PO:0009089	PO:0005679
Anther	**Endosperm**	**Epidermis**
SPM (mean): 0.947	SPM (mean): 0.016	SPM (mean): 0.005
Mean TPM: 96.513 (±137.559) Median TPM: 0.047	Mean TPM: 1.621 (±1.286) Median TPM: 1.364	Mean TPM: 0.492 (±0.199) Median TPM: 0.492
6 samples	10 samples	2 samples

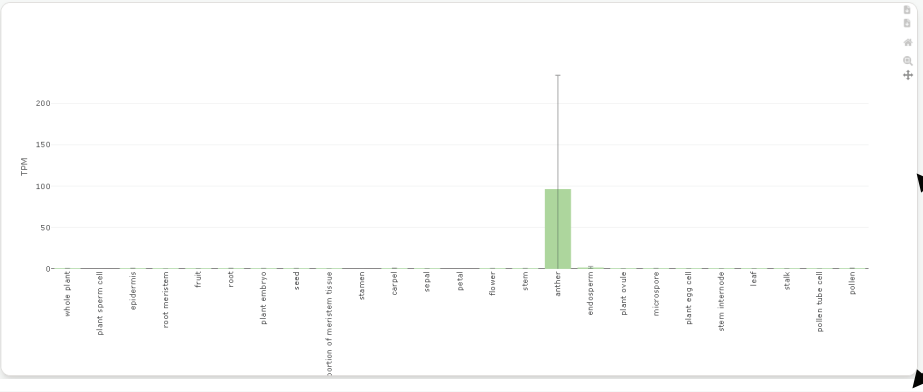
GeneID, Species name, Taxon ID

Organ-specific expression

B

Gene expression profile by organs

Barchart | Boxplot

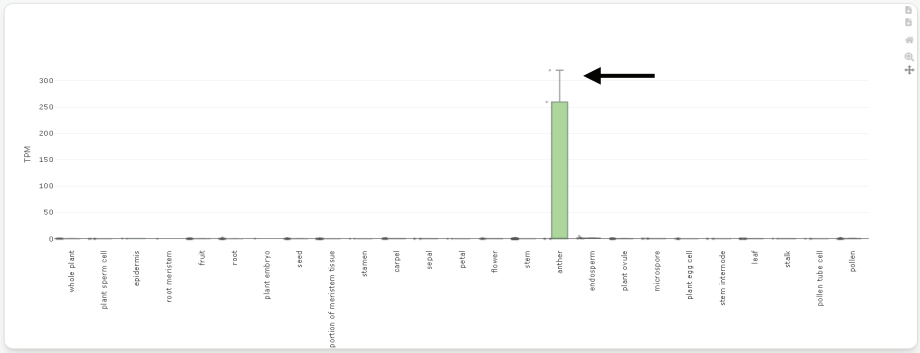


Transcript levels in TPM

Download the image in .png, .svg format

Display in bar chart/boxplot format

Barchart | Boxplot



Boxplot reflects median TPM of each organ with raw data points as points to the left of each bar.

- [x] Show raw data points
- [ ] Zoom Y-axis to exclude outliers
- [ ] Allow zoom on scroll

View the identity of the RNA-seq experiments

Download raw transcript data

C

Mapman annotations

Bin code	Bin name	Bin description
15.5.32	RNA biosynthesis.transcriptional regulation.bHLH-type transcription factor	bHLH-type transcription factor
29.2.1.8.1	Plant reproduction.gametogenesis.male gametophyt.tapetum development.transcriptional regulator *(DYT1/AMS)	transcriptional regulator *(DYT1/AMS)

Mapman annotations

PFAM annotations

PFAM Identifier	PFAM name	GO Terms
PF00010	Helix-loop-helix DNA-binding domain	GO:0046983
PF14215	bHLH-MYC and R2R3-MYB transcription factors N-terminal	-

Pfam annotations

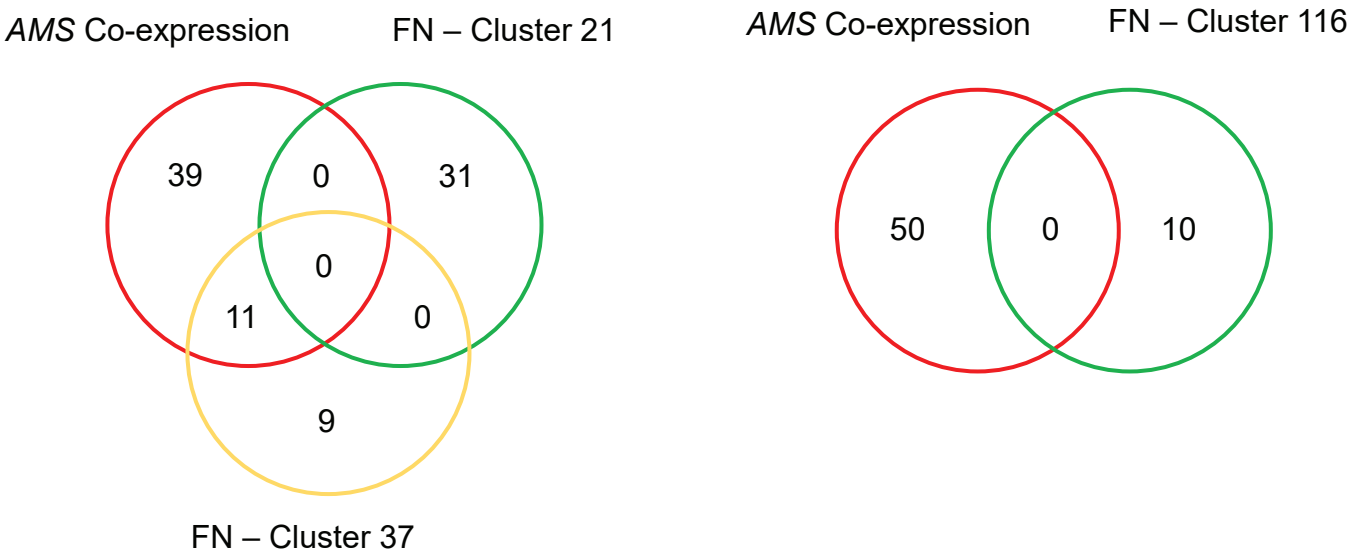
D

Top co-expressed genes

Gene	Description	PCC Value
AT1G33430	Cell wall organisation.hemicellulose.arabinogalactan.biosynthesis.beta-1,3-galactosyltransferase *(B3GALT)	0.97
AT4G35420	Plant reproduction.gametogenesis.male gametophyt.exine formation.sporopollenin biosynthesis.tetraketide alpha-pyrone reductase *(TKPR)	0.934
AT1G62940	Plant reproduction.gametogenesis.male gametophyt.exine formation.sporopollenin biosynthesis.acyl-CoA synthetase *(ACS5)	0.93
AT5G41890	not assigned.not annotated	0.923

Top co-expressed genes sorted by PCC value

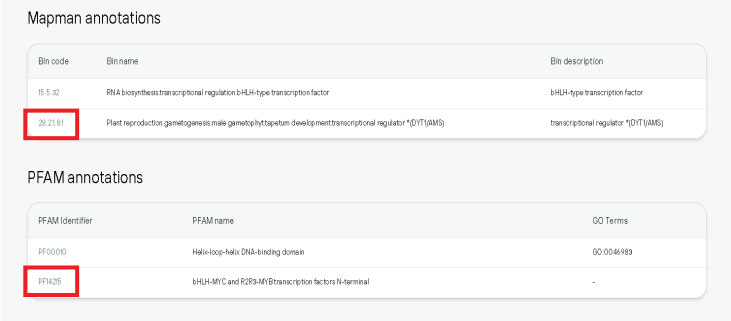
A

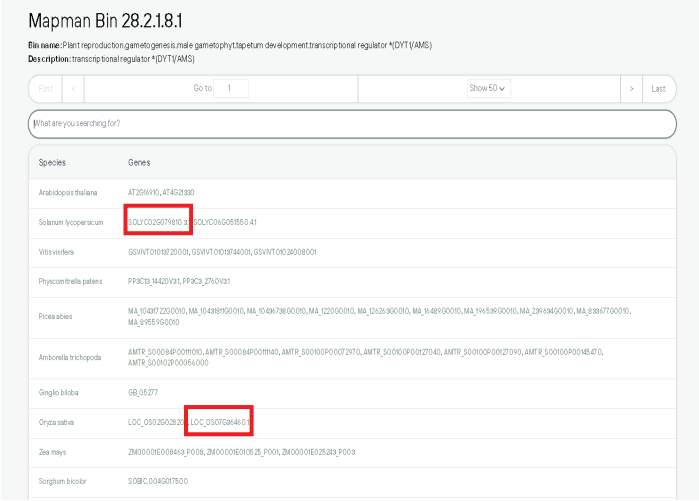


B

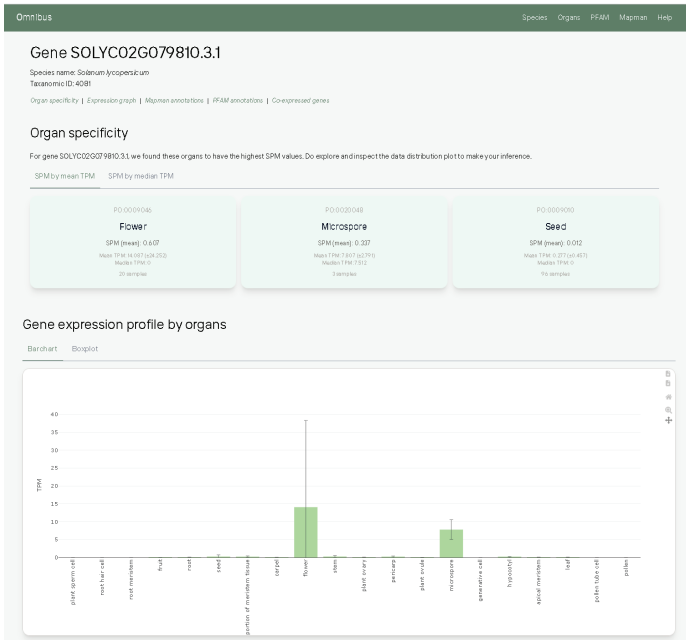
Gene ID	Gene name	Function	Expression
AT1G01280	CYP703A2	cytochrome P450	Anther
AT1G02050	LAP6	Chalcone and stilbene synthase family protein	Anther
AT1G62940	ACOS5	acyl-CoA synthetase	Anther
AT1G69500	CYP704B1	cytochrome P450	Anther
AT3G11980	FAR2	fatty acid reductases	Anther
AT3G13220	ABCG26	ATP-binding cassette transporter G26	Anther
AT3G23770	N.A	O-Glycosyl hydrolases family 17 protein	Anther
AT3G42960	ATA1	homolog of TASSELSEED2	Anther
AT4G14080	MEE48	Involved in the formation of the pollen wall	Anther
AT4G34850	N.A	Chalcone and stilbene synthase family protein	Anther
AT4G35420	DRL1	Dihydroflavonol 4-reductase-like1)	Anther

A Discovery of functional orthologues using MapMan and Pfam annotations

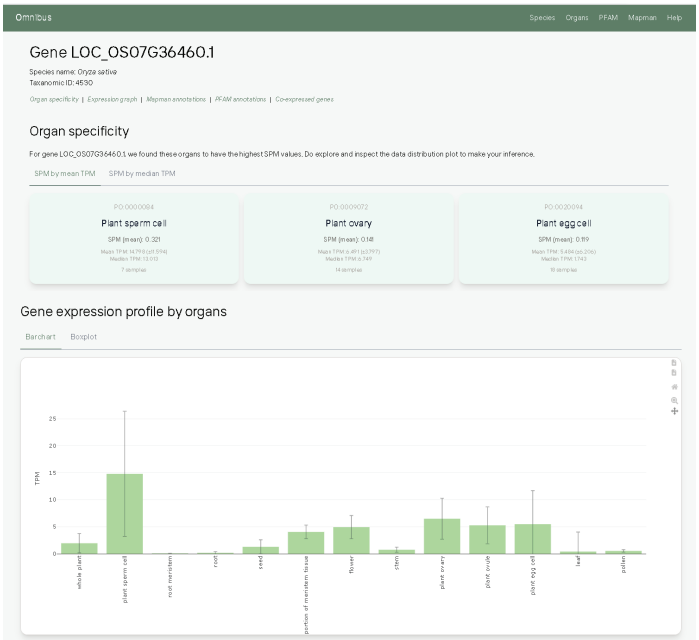




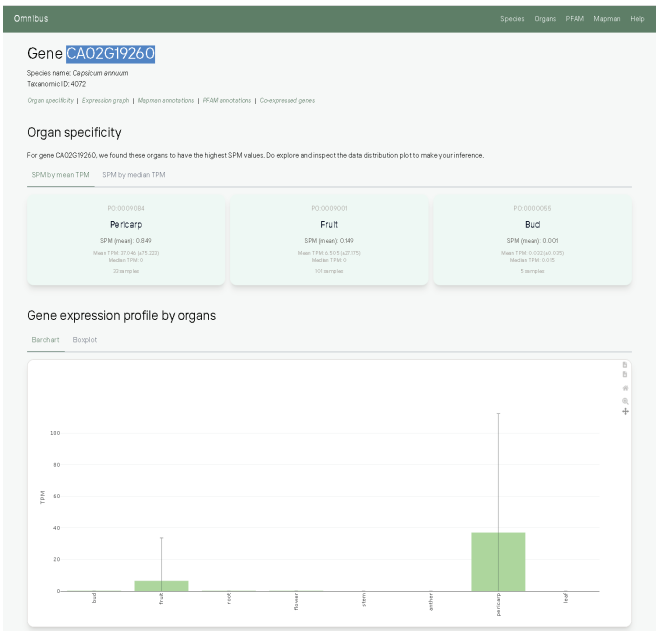
B Output: *Solanum lycopersicum*



Output: *Oryza sativa*



A



C

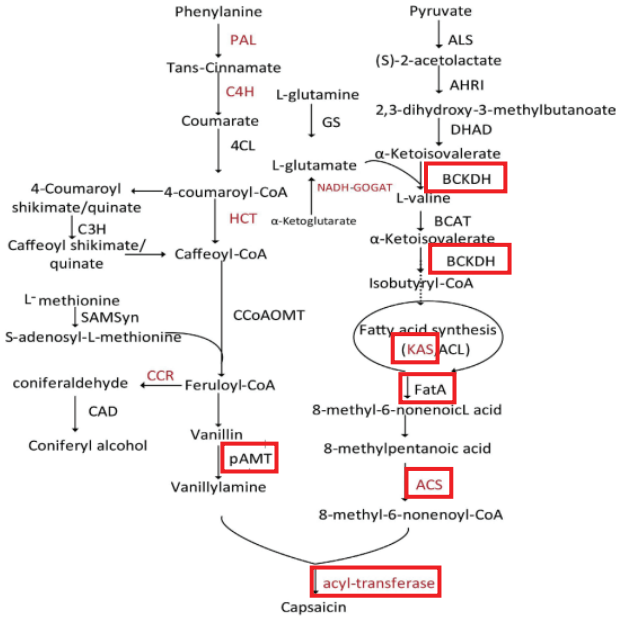
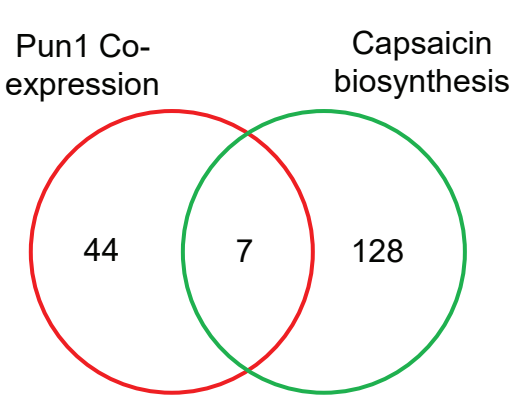


Figure 8. Putative modles of capsaicin biosynthetic pathway.

B



D

Gene ID	Gene name	Function	Specific expression
CA06G10910	*BCKDH E1a*	Branched-chain amino acid aminotransferase	Pericarp
CA03G08530	*pAMT*	Putative aminotransferase	Pericarp
CA01G00840	*KasI*	Ketoacyl-ACP synthase I	Pericarp
CA11G05940	*AT*	acyltransferase 1	Pericarp
CA08G18150	*ACS*	Acyl-CoA synthetase 2	Flower / Pericarp
CA06G26640	*FatA*	Acyl-ACP thioesterase	Pericarp
CA01G28560	*KasIII*	Ketoacyl-ACP synthase III	Pericarp