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; Liu et al., 2023; Robinson et al., 2018; Yu et al., 2022). However, no existing database has been designed to cover well-annotated gene expression data. 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. Gene expression matrices were obtained from a previous study from public databases via our pipeline using recommended thresholds of $\log_{10}$-normalized number of processed reads and percentage of reads pseudo-aligned to the reference CDS (Goh and Mutwil, 2021), which removed samples with low number and percentage of mapping reads. 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). The expression data and annotations are available on 10.6084/m9.figshare.24257125.

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. Co-expression analysis was performed using Pearson Correlation Coefficient (PCC).

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.

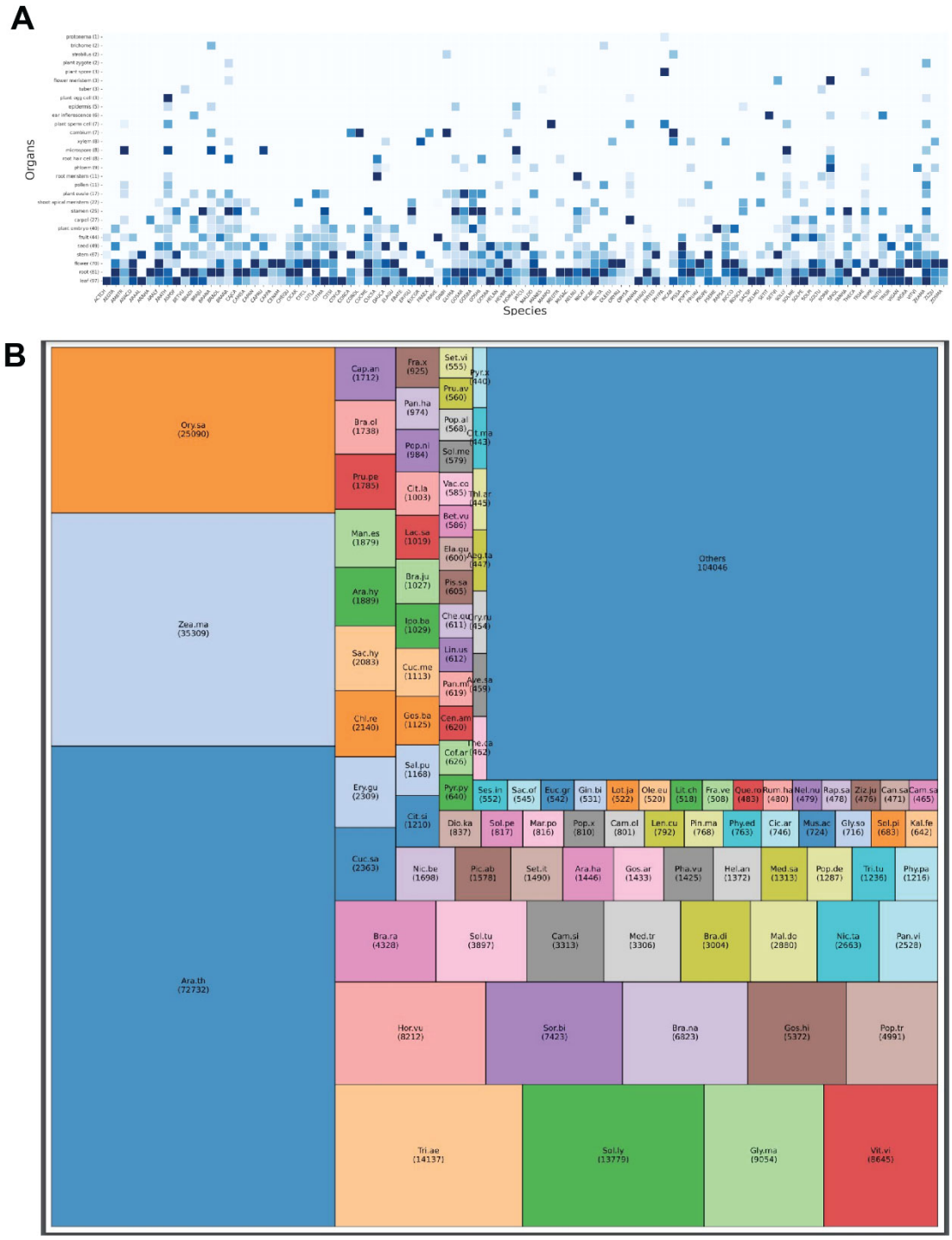
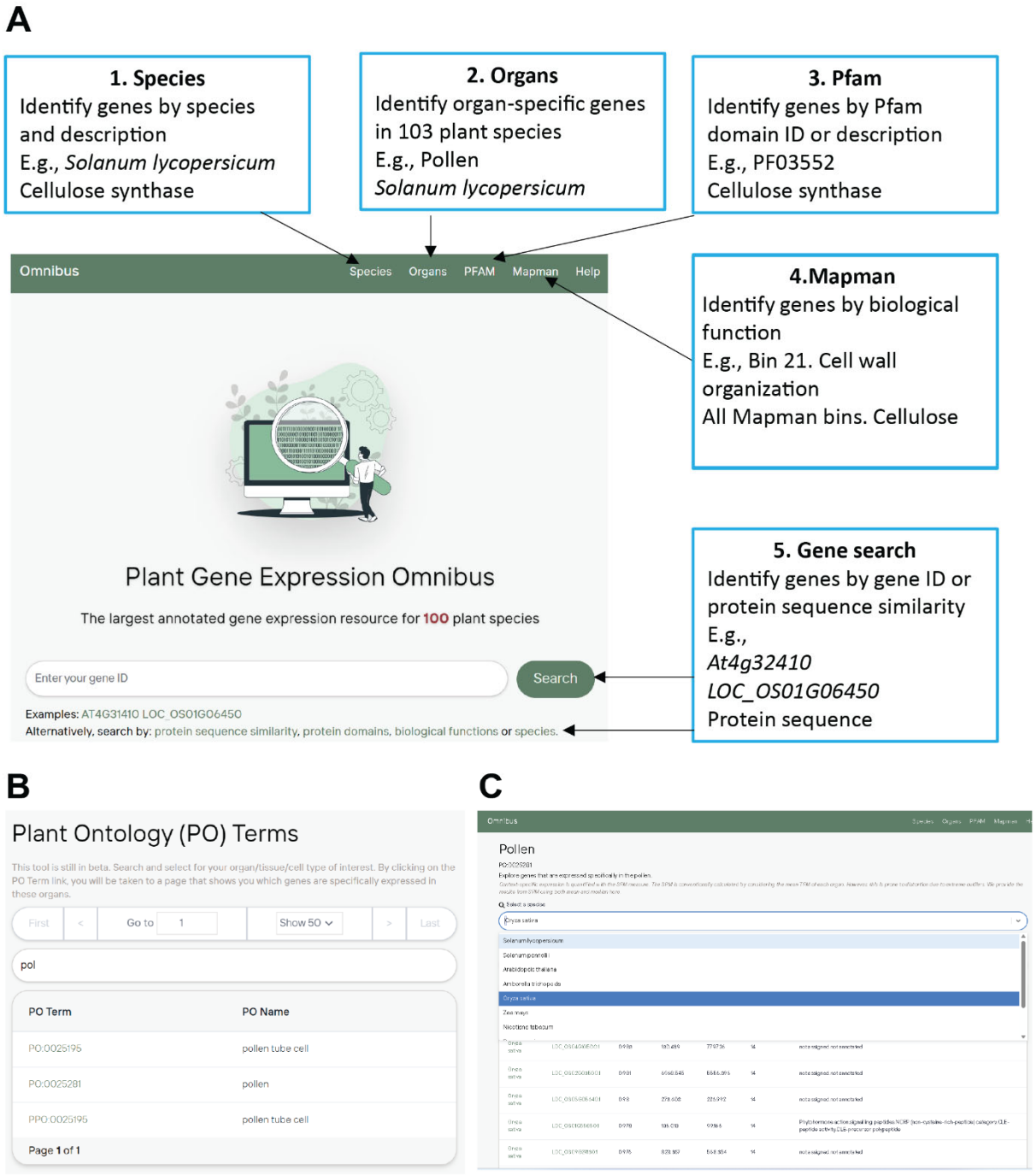


Figure 1: RNA sequencing contents of the Plant Expression Omnibus (PEO) database. A) The heatmap shows the relative abundance of RNA-seq data for the different species (columns) and organs, tissues, and cell types (rows). The numbers in paratheses in rows indicate the number of species that have data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

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

164 developmental pathways that are spatiotemporally regulated.



165

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

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

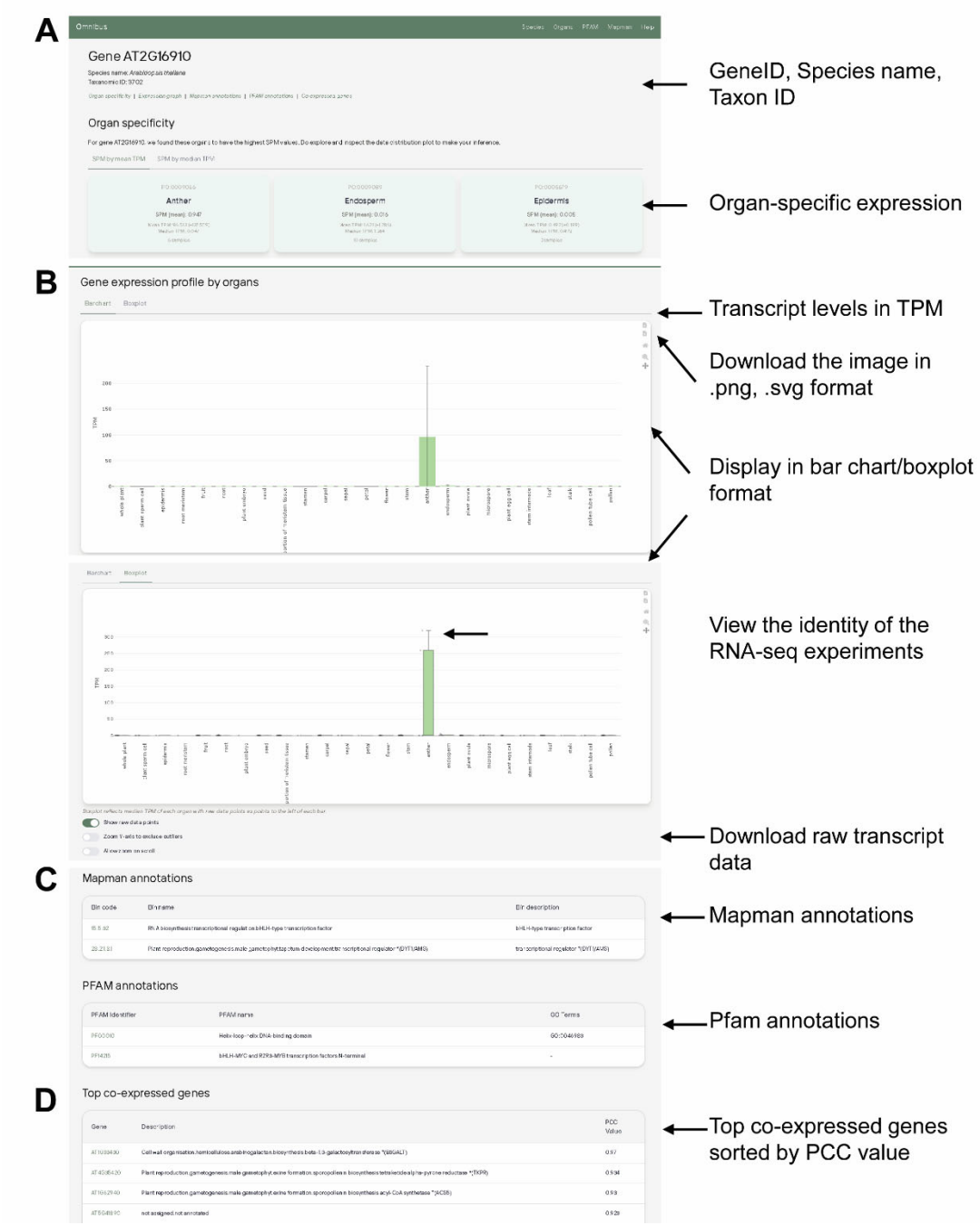
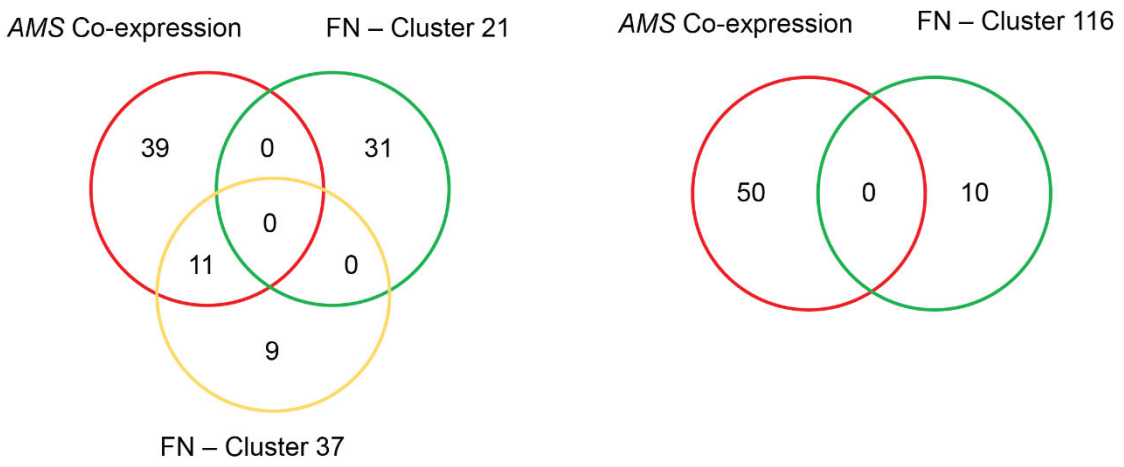


Figure 3: Example gene page of the gene of interest (*AT2G16910*). A) The gene page provides organ specificity of the gene predicted by the SPM score. B) It also displays relative transcript expression in various plant organs/tissues as a box plot or bar chart. Moving the mouse pointer over the points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

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

239 pollen-, or flower-specific expression (Pearce et al., 2015). By generating a correlation
240 network, FlowerNet generates gene clusters of transcriptionally coregulated genes, and
241 based on analysis of tissue specificity and expression at different pollen maturation
242 stages, we were able to identify several clusters associated with pollen development.
243 Venn diagram analysis showed that AMS co-expressed genes overlapped significantly
244 (55%, p = 1.4E-26) with Cluster 37, representing sporopollenin biosynthesis, but not
245 Clusters 21 and 116, representing pollen exine formation and pollen tube growth,
246 respectively (Figure 4A). This suggests that *AMS* plays a key role in regulating
247 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

Figure 4: Example use case of gene *AMS* (*AT2G16910*), a transcription factor 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.

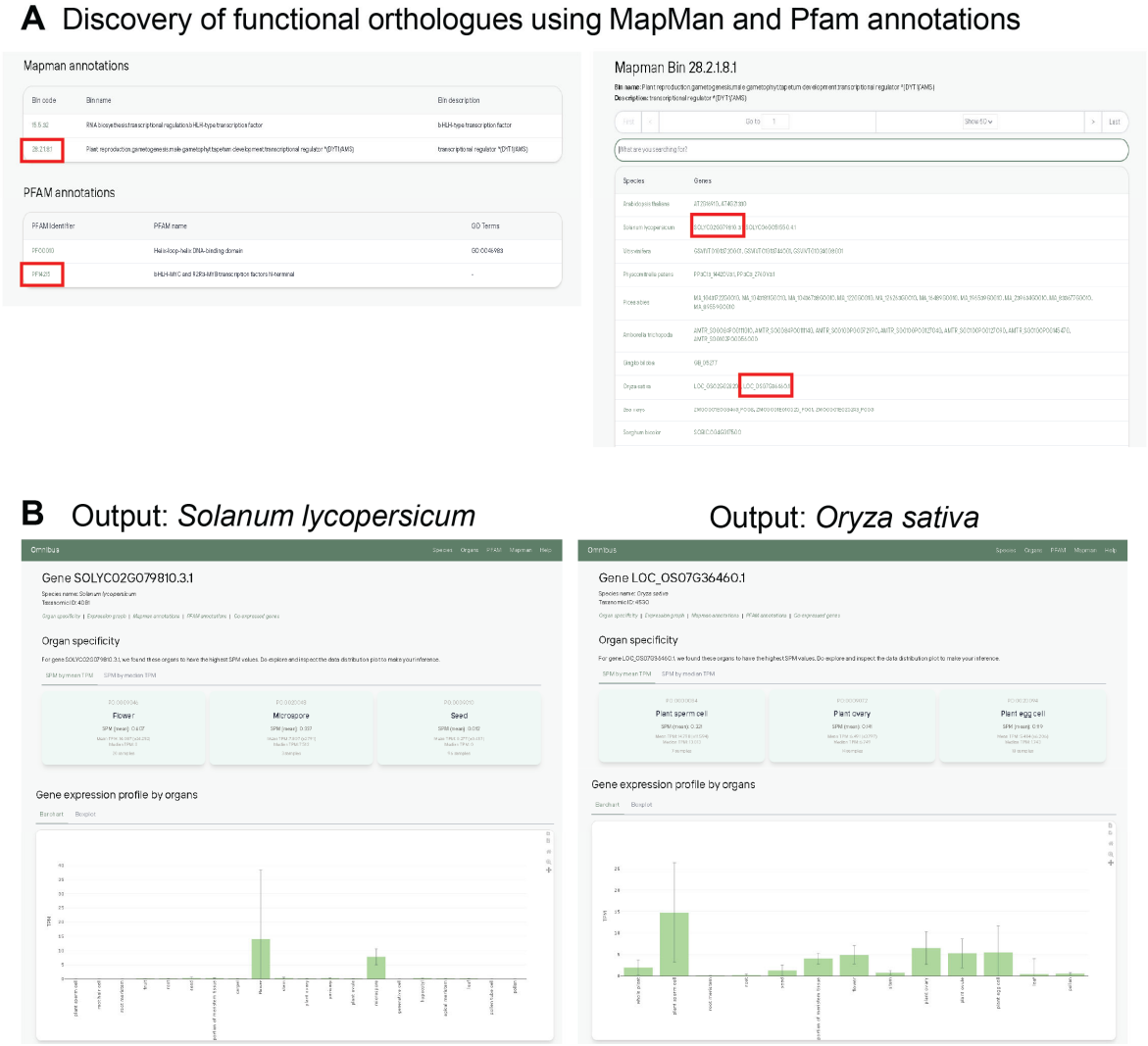
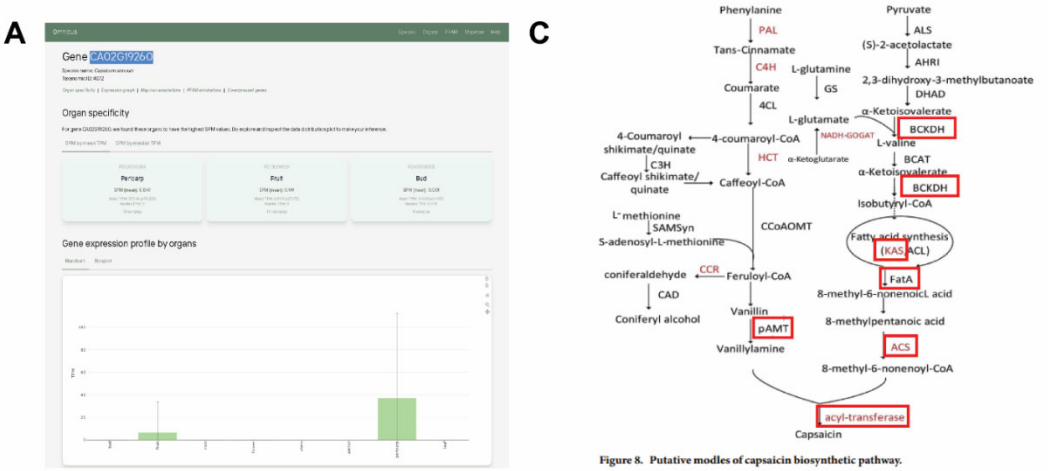


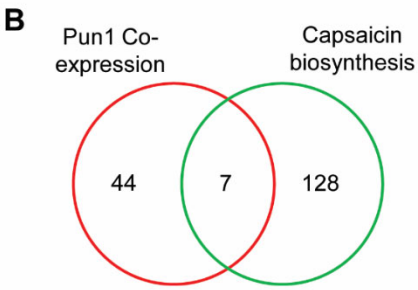
Figure 5: Discovery of functional orthologues using MapMan and Pfam annotations. A) Mapman and Pfam annotations for the gene of interest (*AT2G16910*) provide a list of genes annotated with the bins and domains in other species. B) Example output for functional orthologue discovery using MapMan annotation. The gene of *Solanum lycopersicum* and *Oryza sativa* show specific expression in reproductive tissues similar to *AT2G16910* in *Arabidopsis*.

Further examination of the overlapped genes (11 genes) revealed a list of genes that were annotated as being involved in pollen development (*AT1G01280*, *AT1G69500*,

267 *AT4G35420*) or whose mutants were essential for male fertility (*AT1G62940*,
268 *AT3G11980*, *AT3G13220*, *AT4G35420*), and were all found to be predominantly
269 expressed in anthers (Figure 4B, Supplemental Dataset S2). This strategy of using
270 complementary databases can be used to narrow down large gene lists and identify key
271 genes involved in a pathway of interest. Finally, using the associated Mapman and Pfam
272 annotations available on the gene page, we can obtain gene orthologues from other
273 species and use that as a starting point to investigate genetic pathways in other plants
274 (Figure 5A). As an example, we examined the genes *SOLYC02G079810.3.1* and
275 *LOC_OS07G36460.1* from *Solanum lycopersicum* and *Oryza sativa*, respectively (Figure
276 5B). Observations of the respective gene loci of *Solanum lycopersicum* and *Oryza sativa*
277 show transcript enrichment in reproductive tissues similar to that in Arabidopsis.
278
279 *Use case 2: Identification of biosynthetic pathway components of capsaicin*
280 Here, we will provide an example of using PEO to investigate capsaicin biosynthesis in
281 the crop plant *Capsicum annuum*. The genus Capsicum comprises a large class of crop
282 plants commonly known as peppers (i.e. chili peppers), and its distinctive
283 pungency/spiciness is due to the alkaloid capsaicin (Suzuki et al., 1980). We looked at
284 the literature and found that *Pun1* (*CA02G19260*) is a necessary enzymatic component
285 of capsaicin biosynthesis. The *Pun1* gene encodes an acyltransferase, and the deletion
286 of the *Pun1* gene in bell peppers is the reason for its distinctive non-pungency compared
287 to pungent varieties like chili peppers (Stewart Jr et al., 2005). We searched for the *Pun1*
288 (*CA02G19260*) gene in our database and observed that it was expressed mainly in fruit
289 and pericarp tissues, consistent with the site of capsaicin accumulation in the plant
290 (Figure 6A). Next, we obtained the list of top co-expressed genes for *Pun1* and compared
291 it against a list of putative capsaicin biosynthetic pathway genes previously obtained by
292 isolation from *Capsicum annuum* fruit tissue at specific developmental stages (Mazourek
293 et al., 2009; Zhang et al., 2016). This comparison yielded an overlap of 13.7% (7 genes),
294 which we examined individually for their function and organ expression (Figure 6B). We
295 observed that all the genes were predominantly found in the pericarp tissue, consistent
296 with the site of capsaicin accumulation (Figure 6D). We then obtained a putative map of
297 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* is primarily co-regulated with this arm of capsaicin biosynthesis (Figure 6C).





D

Gene ID	Gene name	Function	Specific expression
CA06G1 0910	*BCKDH E1a*	Branched-chain amino acid aminotransferase	Pericarp
CA03G0 8530	*pAMT*	Putative aminotransferase	Pericarp
CA01G0 0840	*KasI*	Ketoacyl-ACP synthase I	Pericarp
CA11G0 5940	*AT*	acyltransferase 1	Pericarp
CA08G1 8150	*ACS*	Acyl-CoA synthetase 2	Flower / Pericarp
CA06G2 6640	*FatA*	Acyl-ACP thioesterase	Pericarp
CA01G2 8560	*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 the site of capsaicin accumulation. B) Comparison of top co-expressed genes with known capsaicin biosynthetic pathway components yields an overlap of 13.7% (7 genes). C) Mapping of genes to proposed capsaicin biosynthetic pathway based on putative gene function (red boxes). D) Specific expression of the seven genes obtained from PEO.

310 Finally, we also provide several other options for general gene and pathway discovery
311 using either the organ, Pfam or Mapman specific annotation modes. For instance, by
312 selecting the 'Organs' tab, we are directed to a page listing all specific organ / tissue types
313 sorted by Plant Ontology (PO) number. Selecting the PO number (PO:0009005)
314 corresponding to root directs us to a page containing genes specific to the root, sorted by
315 SPM values (Supplemental Figure S4). Here, we can select the species of interest, in this
316 case, *Solanum lycopersicum*, and ask PEO to provide a heatmap of the top 50 genes. As
317 observed, the heatmap displays a set of genes showing strong enrichment in root tissue
318 (Supplemental Figure S4).
319
320 Next, we show an example whereby Pfam IDs are used as a starting point by selecting
321 the Pfam tab. Here we chose Pfam ID (PF00182) which corresponds to Chitinase class
322 I, which is important for defense against fungal pathogens. We are directed to a page
323 which displays a list of genes sorted by species that are associated with this Pfam ID
324 (Supplemental Figure S5). Selecting *Physcomitrella patens* as our species of interest, we
325 can ask PEO to provide a heat map of gene expression in various tissues. We observe
326 that different genes are associated with different tissue types, providing hypothesis
327 generation capabilities for the investigation of these genes in specific tissues.
328
329 Lastly, we also show how Mapman bins can be used as a starting point by selecting the
330 Mapman tab. We select the Mapman bincode (12.1.1.1 Chromatin
331 organisation.chromatin structure. DNA wrapping.histone *(H2A)) as an example. PEO
332 directs us to a page where we are provided with a list of genes associated with this
333 bincode sorted by species. In this case, we generate a heatmap of the species Oryza
334 sativa, where we observe that the majority of the genes are associated with meristematic
335 or reproductive tissue (Supplemental Figure 6). Again, this simple analysis can provide
336 useful insights into gene function and any tissue specific properties it may possess as a
337 starting point for experimental investigations.
338
339 **Conclusion**

340 We present here a fully annotated Gene Expression Omnibus of 103 plant species, which
341 provides information on tissue specificity and gene function information of the relevant
342 gene of interest. PEO also automatically provides a list of top-coexpressed genes for
343 better elucidation of gene functions and associated pathways. We provided two case
344 studies of how PEO can be used to study genes and pathways of interest and to narrow
345 down large putative gene lists based on information on tissue specificity, annotated Pfam
346 and Mapman functions, and comparison using complementary datasets obtained from
347 other analyses. These analyses are amenable to any species present in PEO and allow
348 researchers to build a genetic model of a pathway of interest quickly. As we expand the
349 number of species available in PEO, we will be able to more effectively empower
350 researchers working on non-model organisms and more quickly progress the pace of
351 translational research.
352

353 **Acknowledgments**

354 We would like to thank Riccardo Delli Ponti for the initial help in collecting the gene
355 expression data. M.M. is supported by the Singaporean Food Agency grant
356 SFS_RND_SUFP_001_05 and E.K. is supported by Ministry of Education Tier 3 Grant '
357 From Tough Pollen to Soft Matter'.
358

359 **Conflict of interests**

360 The authors declare no conflicts of interest.
361

362 **References**

363 Ahlers, F., Lambert, J., Wiermann, R., 1999. Structural Elements of Sporopollenin from the
364 Pollen of Torreya californica Torr. (Gymnospermae): Using the 1H-NMR Technique. Zeitschrift
365 für Naturforschung C 54, 492–495. https://doi.org/10.1515/znc-1999-7-806
366 Bray, N.L., Pimentel, H., Melsted, P., Pachter, L., 2016. Near-optimal probabilistic RNA-seq
367 quantification. Nat. Biotechnol. 34, 525–527. https://doi.org/10.1038/nbt.3519
368 Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND.
369 Nat. Methods 12, 59–60. https://doi.org/10.1038/nmeth.3176
370 Cravens, A., Payne, J., Smolke, C.D., 2019. Synthetic biology strategies for microbial
371 biosynthesis of plant natural products. Nat. Commun. 10, 2142.

372 Delli-Ponti, R., Shivhare, D., Mutwil, M., 2021. Using Gene Expression to Study Specialized
373 Metabolism—A Practical Guide. Front. Plant Sci. 11, 2074.
374 https://doi.org/10.3389/fpls.2020.625035
375 Desborough, M.J., Keeling, D.M., 2017. The aspirin story–from willow to wonder drug. Br. J.
376 Haematol. 177, 674–683.
377 Goh, W., Mutwil, M., 2021. LSTrAP-Kingdom: an automated pipeline to generate annotated
378 gene expression atlases for kingdoms of life. Bioinforma. Oxf. Engl. btab168.
379 https://doi.org/10.1093/bioinformatics/btab168
380 Hansen, B.O., Meyer, E.H., Ferrari, C., Vaid, N., Movahedi, S., Vandepoele, K., Nikoloski, Z.,
381 Mutwil, M., 2018. Ensemble gene function prediction database reveals genes important for
382 complex I formation in Arabidopsis thaliana. New Phytol. 217, 1521–1534.
383 https://doi.org/10.1111/nph.14921
384 Julca, I., Ferrari, C., Flores-Tornero, M., Proost, S., Lindner, A.-C., Hackenberg, D.,
385 Steinbachová, L., Michaelidis, C., Gomes Pereira, S., Misra, C.S., Kawashima, T., Borg, M.,
386 Berger, F., Goldberg, J., Johnson, M., Honys, D., Twell, D., Sprunck, S., Dresselhaus, T.,
387 Becker, J.D., Mutwil, M., 2021. Comparative transcriptomic analysis reveals conserved
388 programmes underpinning organogenesis and reproduction in land plants. Nat. Plants 7, 1143–
389 1159. https://doi.org/10.1038/s41477-021-00958-2
390 Klie, S., Nikoloski, Z., 2012. The choice between MapMan and Gene ontology for automated
391 gene function prediction in plant science. Front. Genet. 3, 1–14.
392 https://doi.org/10.3389/fgene.2012.00115
393 Krishna, S., Bustamante, L., Haynes, R.K., Staines, H.M., 2008. Artemisinins: their growing
394 importance in medicine. Trends Pharmacol. Sci. 29, 520–527.
395 Lim, P.K., Zheng, X., Goh, J.C., Mutwil, M., 2022. Exploiting plant transcriptomic databases:
396 Resources, tools, and approaches. Plant Commun. 100323.
397 Liu, J., Zhang, Y., Zheng, Y., Zhu, Y., Shi, Y., Guan, Z., Lang, K., Shen, D., Huang, W., Dou, D.,
398 2023. PlantExp: a platform for exploration of gene expression and alternative splicing based on
399 public plant RNA-seq samples. Nucleic Acids Res. 51, D1483–D1491.
400 Ma, L., Sun, N., Liu, X., Jiao, Y., Zhao, H., Deng, X.W., 2005. Organ-specific expression of
401 Arabidopsis genome during development. Plant Physiol. 138, 80–91.
402 Mazourek, M., Pujar, A., Borovsky, Y., Paran, I., Mueller, L., Jahn, M.M., 2009. A Dynamic
403 Interface for Capsaicinoid Systems Biology. Plant Physiol. 150, 1806–1821.
404 https://doi.org/10.1104/pp.109.136549
405 Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L.,
406 Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D., Bateman, A., 2021. Pfam: The
407 protein families database in 2021. Nucleic Acids Res. 49, D412–D419.
408 https://doi.org/10.1093/nar/gkaa913
409 Niazian, M., 2019. Application of genetics and biotechnology for improving medicinal plants.
410 Planta 249, 953–973.
411 Nowicka, B., Ciura, J., Szymańska, R., Kruk, J., 2018. Improving photosynthesis, plant
412 productivity and abiotic stress tolerance–current trends and future perspectives. J. Plant
413 Physiol. 231, 415–433.

414 Paddon, C.J., Westfall, P.J., Pitera, D.J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M., Tai,
415 A., Main, A., Eng, D., 2013. High-level semi-synthetic production of the potent antimalarial
416 artemisinin. Nature 496, 528–532.
417 Pearce, S., Ferguson, A., King, J., Wilson, Z.A., 2015. FlowerNet: A Gene Expression
418 Correlation Network for Anther and Pollen Development. Plant Physiol. 167, 1717–1730.
419 https://doi.org/10.1104/pp.114.253807
420 Proost, S., Mutwil, M., 2018. CoNekT: An open-source framework for comparative genomic and
421 transcriptomic network analyses. Nucleic Acids Res. 46, W133–W140.
422 https://doi.org/10.1093/nar/gky336
423 Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., Wittkop, T., Sokolov, A., Graim, K., Funk,
424 C., Verspoor, K., Ben-Hur, A., Pandey, G., Yunes, J.M., Talwalkar, A.S., Repo, S., Souza, M.L.,
425 Piovesan, D., Casadio, R., Wang, Z., Cheng, J., Fang, H., Gough, J., Koskinen, P., Törönen, P.,
426 Nokso-Koivisto, J., Holm, L., Cozzetto, D., Buchan, D.W.A., Bryson, K., Jones, D.T., Limaye, B.,
427 Inamdar, H., Datta, A., Manjari, S.K., Joshi, R., Chitale, M., Kihara, D., Lisewski, A.M., Erdin, S.,
428 Venner, E., Lichtarge, O., Rentzsch, R., Yang, H., Romero, A.E., Bhat, P., Paccanaro, A.,
429 Hamp, T., Kaßner, R., Seemayer, S., Vicedo, E., Schaefer, C., Achten, D., Auer, F., Boehm, A.,
430 Braun, T., Hecht, M., Heron, M., Hönigschmid, P., Hopf, T.A., Kaufmann, S., Kiening, M.,
431 Krompass, D., Landerer, C., Mahlich, Y., Roos, M., Björne, J., Salakoski, T., Wong, A., Shatkay,
432 H., Gatzmann, F., Sommer, I., Wass, M.N., Sternberg, M.J.E., Škunca, N., Supek, F., Bošnjak,
433 M., Panov, P., Džeroski, S., Šmuc, T., Kourmpetis, Y.A.I., Van Dijk, A.D.J., Ter Braak, C.J.F.,
434 Zhou, Y., Gong, Q., Dong, X., Tian, W., Falda, M., Fontana, P., Lavezzo, E., Di Camillo, B.,
435 Toppo, S., Lan, L., Djuric, N., Guo, Y., Vucetic, S., Bairoch, A., Linial, M., Babbitt, P.C., Brenner,
436 S.E., Orengo, C., Rost, B., Mooney, S.D., Friedberg, I., 2013. A large-scale evaluation of
437 computational protein function prediction. Nat. Methods 10, 221–227.
438 https://doi.org/10.1038/nmeth.2340
439 Rao, X., Dixon, R.A., 2019. Co-expression networks for plant biology: why and how. Acta
440 Biochim. Biophys. Sin. 51, 981–988. https://doi.org/10.1093/abbs/gmz080
441 Rhee, S.Y., Mutwil, M., 2014. Towards revealing the functions of all genes in plants. Trends
442 Plant Sci. 19, 212–221. https://doi.org/10.1016/j.tplants.2013.10.006
443 Robinson, A.J., Tamiru, M., Salby, R., Bolitho, C., Williams, A., Huggard, S., Fisch, E.,
444 Unsworth, K., Whelan, J., Lewsey, M.G., 2018. AgriSeqDB: an online RNA-Seq database for
445 functional studies of agriculturally relevant plant species. BMC Plant Biol. 18, 1–8.
446 Ryngajllo, M., Childs, L., Lohse, M., Giorgi, F.M., Lude, A., Selbig, J., Usadel, B., 2011. SLocX:
447 predicting subcellular localization of Arabidopsis proteins leveraging gene expression data.
448 Front. Plant Sci. 43.
449 Sabzehzari, M., Zeinali, M., Naghavi, M.R., 2020. Alternative sources and metabolic
450 engineering of Taxol: Advances and future perspectives. Biotechnol. Adv. 43, 107569.
451 Schwacke, R., Ponce-Soto, G.Y., Krause, K., Bolger, A.M., Arsova, B., Hallab, A., Gruden, K.,
452 Stitt, M., Bolger, M.E., Usadel, B., 2019. MapMan4: A Refined Protein Classification and
453 Annotation Framework Applicable to Multi-Omics Data Analysis. Mol. Plant 12, 879–892.
454 https://doi.org/10.1016/j.molp.2019.01.003
455 Scott, R.J., Spielman, M., Dickinson, H.G., 2004. Stamen Structure and Function. Plant Cell 16,
456 S46–S60. https://doi.org/10.1105/tpc.017012

457 Stewart Jr, C., Kang, B.-C., Liu, K., Mazourek, M., Moore, S.L., Yoo, E.Y., Kim, B.-D., Paran, I.,
458 Jahn, M.M., 2005. The Pun1 gene for pungency in pepper encodes a putative acyltransferase.
459 Plant J. 42, 675–688. https://doi.org/10.1111/j.1365-313X.2005.02410.x
460 Suzuki, T., Fujiwake, H., Iwai, K., 1980. Intracellular localization of capsaicin and its analogues,
461 capsaicinoid, in Capsicum fruit 1. Microscopic investigation of the structure of the placenta of
462 Capsicum annuum var. annuum cv. Karayatsubusa1. Plant Cell Physiol. 21, 839–853.
463 https://doi.org/10.1093/oxfordjournals.pcp.a076058
464 Tan, Q.W., Goh, W., Mutwil, M., 2020. LSTrAP-Cloud: A User-Friendly Cloud Computing
465 Pipeline to Infer Coexpression Networks. Genes 11, 428.
466 https://doi.org/10.3390/genes11040428
467 Walls, R.L., Cooper, L., Elser, J., Gandolfo, M.A., Mungall, C.J., Smith, B., Stevenson, D.W.,
468 Jaiswal, P., 2019. The Plant Ontology Facilitates Comparisons of Plant Development Stages
469 Across Species. Front. Plant Sci. 10. https://doi.org/10.3389/fpls.2019.00631
470 Wetterstrand, K.A., 2013. DNA sequencing costs: data from the NHGRI Genome Sequencing
471 Program (GSP). Natl. Hum. Genome Res. Inst.
472 Xu, J., Ding, Z., Vizcay-Barrena, G., Shi, J., Liang, W., Yuan, Z., Werck-Reichhart, D.,
473 Schreiber, L., Wilson, Z.A., Zhang, D., 2014. ABORTED MICROSPORES Acts as a Master
474 Regulator of Pollen Wall Formation in Arabidopsis. Plant Cell 26, 1544–1556.
475 https://doi.org/10.1105/tpc.114.122986
476 Yu, Y., Zhang, H., Long, Y., Shu, Y., Zhai, J., 2022. Plant public RNA-seq database: a
477 comprehensive online database for expression analysis of~ 45 000 plant public RNA-seq
478 libraries. Plant Biotechnol. J. 20, 806.
479 Zhang, Z.-X., Zhao, S.-N., Liu, G.-F., Huang, Z.-M., Cao, Z.-M., Cheng, S.-H., Lin, S.-S., 2016.
480 Discovery of putative capsaicin biosynthetic genes by RNA-Seq and digital gene expression
481 analysis of pepper. Sci. Rep. 6, 34121. https://doi.org/10.1038/srep34121
482

483