1. Supplementary Information:

A. Flat Files

Item	Present?	Filename This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. The extension must be .pdf	A brief, numerical description of file contents. i.e.: *Supplementary Figures 1-4, Supplementary Discussion, and Supplementary Tables 1-4.*
Supplementary Information	Yes	Supplementary_figures+methods.pdf	Supplementary Figures 1-11, Supplementary Methods
Reporting Summary	Yes	ReportingSummary.pdf	

B. Additional Supplementary Files

Type	Number If there are multiple files of the same type this should be the numerical	Filename This should be the name the file is saved as when it is uploaded to our	Legend or Descriptive Caption Describe the contents of the file

	indicator. i.e. “1” for Video 1, “2” for Video 2, etc.	system, and should include the file extension. i.e.: *Smith_ Supplementary_Video_1.mov*	
Supplementary Data	1	Supplementary_Tables_v2.xls	Supplementary Table 1-22

Comparative transcriptomic analysis reveals conserved programs underpinning organogenesis and reproduction in land plants

Authors: Irene Julca[1], Camilla Ferrari[2], María Flores-Tornero[3], Sebastian Proost[2,4,5], Ann-Cathrin Lindner[6], Dieter Hackenberg[7,8], Lenka Steinbachová[9], Christos Michaelidis[9], Sónia Gomes Pereira[6], Chandra Shekhar Misra[6,13], Tomokazu Kawashima[10,11], Michael Borg[10], Frédéric Berger[10], Jacob Goldberg[12], Mark Johnson[12], David Honys[9], David Twell[7], Stefanie Sprunck[3], Thomas Dresselhaus[3], Jörg D. Becker[6,13*], Marek Mutwil[1*]

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

2) Max Planck Institute for Molecular Plant Physiology, Am Muehlenberg 1, 14476 Potsdam-Golm, Germany

3) Cell Biology and Plant Biochemistry, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany

4) Laboratory of Molecular Bacteriology, Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium

5) VIB, Center for Microbiology, Kasteelpark Arenberg 31, 3000 Leuven, Belgium

6) Instituto Gulbenkian de Ciência, Rua da Quinta Grande 6, 2780-156 Oeiras, Portugal

7) Department of Genetics and Genome Biology, University of Leicester, University Road, Leicester, LE1 7RH, UK.

8) School of Life Sciences, Gibbet Hill Campus, The University of Warwick, Coventry, CV4 7AL, UK

9) Laboratory of Pollen Biology, Institute of Experimental Botany of the Czech Academy of Sciences, Rozvojová 263, 165 02, Prague, Czech Republic

10) Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna, BioCenter (VBC), Dr. Bohr-Gasse 3, 1030 Vienna, Austria

11) Dept. of Plant and Soil Sciences, University of Kentucky, 321 Plant Science Building, 1405 Veterans Dr., Lexington, KY 40546-0312

12) Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University, Providence, RI, 02912, USA

13) Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Av. da República, 2780-157 Oeiras, Portugal

*Corresponding authors:

Marek Mutwil (mutwil@ntu.edu.sg)

Jörg D. Becker (jbecker@igc.gulbenkian.pt)

Abstract

The appearance of plant organs mediated the explosive radiation of land plants, which shaped the biosphere and allowed the establishment of terrestrial animal life. The evolution of organs and immobile gametes required the coordinated acquisition of novel gene functions, the co-option of existing genes, and the development of novel regulatory programs. However, no large-scale analyses of genomic and transcriptomic data have been performed for land plants. To remedy this, we have generated gene expression atlases for various organs and gametes of 10 plant species comprising bryophytes, vascular plants, gymnosperms, and flowering plants. Comparative analysis of the atlases identified hundreds of organ- and gamete-specific orthogroups and revealed that most of the specific transcriptomes are significantly conserved. Interestingly, our results suggest that co-option of existing genes is the main mechanism for evolving new organs. In contrast to female gametes, male gametes showed a high number and conservation of specific genes, indicating that male reproduction is highly specialized. The expression atlas capturing pollen development revealed numerous transcription factors and kinases essential for pollen biogenesis and function.

Introduction

The evolution of land plants has completely changed the appearance of our planet. In contrast to most of their algal relatives, land plants are characterized by three-dimensional growth and the development of complex and specialized organs[1]. They possess a host of biochemical adaptations, including those necessary for tolerating desiccation and UV stress encountered on land, allowing them to colonize most terrestrial surfaces. The earliest land plants were likely not equipped with these adaptations, and many of these adaptations were likely gained on land[2]. The earliest land plants which arose ~470 million years ago[3], possessed tiny fertile axes or an axis terminated by a sporangium[1,4]. The innovation of shoots and leaves mediated the 10-fold expansion in the diversification of vascular plants[5] and an 8–20-fold atmospheric CO2 drawdown[6], which significantly shaped the Earth's geosphere and biosphere[7]. To

enable soil attachment and nutrient uptake, the first land plants only had rhizoids, filamentous structures homologous to root hairs[8]. Roots later evolved to provide increased anchorage (and thus increased height), nutrient uptake, and enable survival in more arid environments. Parallel with innovations of vegetative cell types, land plants evolved new reproductive structures such as spores, pollen, embryo sacs, and seeds together with the gradual reduction of the haploid phase. In contrast to algae, bryophytes, and ferns that require moist habitats, the male and female gametophytes of gymnosperms and angiosperms are strongly reduced, consisting of only a few cells, including the gametes[9,10]. Moreover, sperm cells have lost their mobility (with the exception of the gymnosperm Ginkgo and the cycads[11]) and use pollen grains as a protective vehicle for long-distance transport and a pollen tube for their delivery deep into maternal reproductive tissues[12]. The precise interaction of plant male and female gametes, leading to cell fusion, karyogamy, and development of both the embryo and endosperm after double fertilization has just begun to be deciphered at the molecular level[13]. These anatomical innovations are mediated by coordinated changes in gene expression and the appearance of novel genes and/or repurposing of existing genetic material. Genes that are specifically expressed in these organs often play a major role in their establishment and function[14,15], but the identity and conservation of these specifically-expressed genes have not been extensively studied.

Nowadays, flowering plants comprise 90% of all land plants and serve as the basis for the terrestrial food chain, either directly or indirectly. The use of model plants like *Arabidopsis thaliana* and maize and technical advances allowing live-cell imaging of double fertilization have been instrumental for several major discoveries[16]. When assessing current knowledge of male and female gamete development in plants, it is evident that the male germline has been studied to a greater extent [9,10]. This is mainly due to its accessibility and the development of methods to separate the sperm cells from the surrounding vegetative cell of pollen, e.g. by FACS[17]. Analysis of male germline differentiation, for example, has led to the identification of *Arabidopsis DUO POLLEN 1* (*DUO1*) and the network of genes it controls, which include the fertilization factors, *HAP2/GCS1* and *GEX2*[18]. However, as novel genes are still being discovered that control the development of male and female gametes[9,10] or their

functions[19], it is clear that our knowledge of the molecular basis of gamete formation and function is far from complete.

Current approaches to study evolution and gene function mainly use genomic data to reveal which orthogroups are gained, expanded, contracted, or lost. Comparison of 208 genomes revealed two bursts of genomic novelties in the ancestors of streptophytes and land plants, which were most likely required for the establishment of multicellularity and terrestrialization[20]. While invaluable, genomic approaches alone might not reveal the function of genes that show no sequence similarity to known genes[21]. To our knowledge, no comprehensive comparisons of organ- and tissue-specific transcriptomes in land plants have been done. To remedy this, we combined comparative genomic approaches with newly established, comprehensive gene expression atlases of two bryophytes, a lycophyte , two gymnosperms , a sister to all angiosperms , two eudicots and two monocots. We then compared these organ-, tissue- and cell-specific genes to identify novel and missing components involved in organogenesis and gamete development.

We show that transcriptomes of most organs are conserved across land plants and report the identity of hundreds of organ-specific orthogroups. We demonstrate that the age of orthogroups is positively correlated with organ-specific expression and the appearance of organ-specific orthogroups does not coincide with the appearance of the corresponding organ. We observed a high number of male-specific orthogroups and strong conservation of male-specific transcriptomes, while female-specific transcriptomes showed fewer specific orthogroups with less conservation. Our detailed analysis of gene expression data capturing the development of pollen revealed numerous transcription factors and kinases potentially important for pollen biogenesis and function. Finally, we present a user-friendly, online database www.evorepro.plant.tools, which allows the browsing and comparative analysis of the genomic and transcriptomic data derived from sporophytic and gametophytic samples across 13 members of the plant kingdom.

Results

Constructing gene expression atlases and identifying organ-specific genes

We constructed gene expression atlases for ten phylogenetically representative species (Table 1). These include the bryophytes *Physcomitrium patens* (*Physcomitrella*) (Fig. 1a) and *Marchantia polymorpha* (Fig. 1b), the lycophyte *Selaginella moellendorffii*, the gymnosperms *Ginkgo biloba* and *Picea abies*, the sister lineage of all other angiosperms *Amborella trichopoda*, the monocots *Oryza sativa* and *Zea mays*, and the eudicots *Arabidopsis thaliana* and *Solanum lycopersicum* (Fig. 1c). The atlases were constructed by combining publicly available RNA sequencing (RNA-seq) data with 134 fastq files generated by the EVOREPRO consortium, which after quality control captured 18 different organs, tissues or cell types in ten land plants (see Supplementary Table 1). For each species, we generated an expression matrix that contains transcript-level abundances captured by transcript per million (TPM) values[22]. The expression matrices capture gene expression values from the main anatomical sample types (from now on called organs), which we grouped into ten classes: flower (comprising whole flowers, or floral tissues with absent or small proportion of gametes), female, male, seeds, spore, leaf, stem, apical meristem, root meristem, and root (Fig. 1a-c). Furthermore, the expression data was used to construct co-expression networks and to create an online EVOREPRO database allowing further analysis of the data (www.evorepro.plant.tools).

To identify genes expressed in the different organs, we included only those with an average TPM >2 (see methods). For all ten species, approximately 71% of their genes were expressed in at least one structure (Supplementary Table 2). Interestingly, the male sample has a lower percentage (38%) followed by root meristems (46%), while the other organs have between 50-60% expressed genes (Fig. 1d).

Organ- and cell-specific genes can often play a major role in the establishment and function of the organ and cell type[14,15]. To identify such genes, we calculated the specificity measure (SPM) of each gene, which ranges from 0 (not expressed in an organ) to 1 (expressed only in the organ). A threshold capturing top 5% of the SPM values was used to identify the organ-specific genes for all species (Supplementary Fig. 1, Supplementary Table 3). To examine the organ-specific gene expression

profiles, we plotted the scaled TPM values of these genes for *A. thaliana*. Visual inspection shows that the TPM values of the organ-specific genes are in all cases highest in the organs that the genes are specific to (Fig. 1e, Supplementary Fig. 2). We then used the Plant Ontology (PO) annotations of *Arabidopsis* to test whether the experimentally verified organ-specific function of genes defined by PO corresponds to our predictions. We divided PO annotations in 11 groups: 10 corresponding to the organs we studied, and one named 'others', which were the annotations that could either correspond to more than one organ (i.e guard mother cell could correspond to Leaf or Stem), or represent organs and tissues not analyzed in this study (e.g., hypocotyl, coleoptile). From the total of genes classified as organ-specific in *Arabidopsis* (9,798 genes), only an average of 11.4% had PO annotation (Flower - 11.4%, Female - 6.9%, Male - 8.5%, Seeds - 9.4%, Leaf - 11.3%, Stem - 16.6%, Apical meristem - 17.6%, Root meristem - 9.4%, and Root - 11.4%). In general, the PO annotation of those genes show correspondence with the organ at which were assigned (i.e. the higher percentage of flower-specific genes have PO annotations related to flowers, Fig. 1f) except for leaf-specific genes, where most genes belong to the 'Others' category.

For the ten species, an average of 21% of the genes were identified as organ-specific (Supplementary Table 2). The lowest percentage of organ-specific genes was found in *P. abies* (5%), followed by *M. polymorpha* (11%) and *P. patens* (11%), while the highest percentage was found in *A. thaliana,* where 35% of the transcripts showed organ-specific expression (Supplementary Table 2). These differences can be partially explained by the number of organs and cell types that we analyzed, and the availability of data for each species, with *Arabidopsis* having most data (Supplementary Table 1). Interestingly, we observed that the male (5.3%) and root (5.0%) samples typically contained the highest percentage of specific genes in the studied species (Fig. 1g, Supplementary Table 2). In *A. thaliana*, the higher percentage of male-specific genes was in agreement with previous studies that showed a high specialization of the male transcriptome[23]. Conversely, stem, spore, apical meristem, root meristem, flower, and female show values lower than 3% (Fig. 1g, Supplementary Table 2). This is in line with the previous studies that also showed a low number of genes specific to the female gametophyte[24].

To summarize, these results show that organ-specific genes represent a significant part of the transcriptome, with male and root samples possessing the most specialized transcriptomes.

Are the transcriptomes of organs conserved across species?

Our above analysis suggests that organ-specific gene expression is widespread, and we set out to investigate whether these patterns are conserved across species. To this end, we investigated which organs specifically expressed similar sets of orthogroups by employing a Jaccard distance that ranges from 0 (two samples express an identical set of organ-specific orthogroups) to 1 (none of the organ-specific orthogroups are the same in the two samples). We expected that if, e.g., the root-specific transcriptome is conserved across angiosperms, then Jaccard distance of root vs. root transcriptomes (e.g., *Arabidopsis* root vs. rice root) should be lower than when comparing root vs. non-root transcriptomes (e.g., *Arabidopsis* root vs. rice leaf).

The analysis revealed that *Arabidopsis* flower-, male-, seeds-, stem- and root-specific transcriptomes were significantly more similar to the corresponding organ in the other species (Wilcoxon rank-sum test p-value < 0.05, Fig. 2a). When performing the analysis for all ten species, we observed that root, male, and seeds expressed specifically similar orthogroups in all species with the samples (7 species for root, 7 for male, and 5 for seeds) and for other organs, some species show significance, flowers (5 out of 7 species with flower samples), female (2 out of 6), leaf (7 out of 10), stem (5 out of 7), apical meristem (4 out of 5), root meristem (4 out of 5) (Fig. 2b, Supplementary Fig. 3). Conversely, spore (0 out of 2) samples did not show similar transcriptomes across *Marchantia* and *Physcomitrium* (Fig. 2b, Supplementary Fig. 3).

As our analysis can serve as a transcriptional readout that can aid in defining the homology of organs, we also performed clustering analysis between all pairs of organ-specific genes in the ten species and observed root-, seed-, flower, leaf-, meristem- and male-specific clusters (Supplementary Fig. 4). Interestingly, the male samples in *Physcomitrium* and *Marchantia* formed a distinctive cluster (Supplementary Fig. 4), suggesting that flagellated sperm of bryophytes employ a unique male transcriptional program compared with non-motile sperm of angiosperms.

To reveal which biological processes are preferentially expressed in the different organs across the ten species, we performed a functional enrichment analysis of Mapman bins, transcription factors, and kinases (Fig. 2c, Supplementary Fig. 5, Supplementary Table 4, 5). The analysis revealed that many functions were depleted in male and root samples in at least 50% of the species, indicating that most male and roots' cellular processes were significantly repressed (p-value < 0.05, Fig. 2c, Supplementary Fig. 5). As expected, genes associated with photosynthesis were enriched in leaves but depleted in roots, root meristems, and male samples. Genes expressed in roots were enriched in solute transport functions, enzyme classification (enzymes not associated with other processes), RNA biosynthesis, secondary metabolism, phytohormone action, and cell wall organization (Fig. 2c). Interestingly, female and male reproductive cells were enriched for 'not assigned' bin, indicating that these organs are enriched for genes with unknown functions.

Since the organ-specific genes (Supplementary Table 3) are likely important for the formation and function of the organ, we investigated organ-specific transcription factors (Supplementary Table 6) and receptor kinases (Supplementary Table 7). An enrichment analysis of transcription factors (69 families) and kinases (142 families) showed that apical meristem and root samples were highly enriched in transcription factors, while male and apical meristem were enriched for kinases (Fig. 2c). In apical meristems, some of the enriched transcription factor families (C2C2-YABBY, GRF) were associated with the regulation, development, and differentiation of meristem[25,26]. In roots, the enriched transcription factors (MYB, bHLH, WRKY, NAC) are related to biotic and abiotic stress response and root development[27,28]. These organ-specific genes are thus prime candidates for further functional analysis (Supplementary Table 7).

Phylostratigraphic analysis of organ-specific orthogroups

Organs, such as seeds and flowers, appeared at a specific time in plant evolution. To investigate whether there is a link between orthogroups' appearance and their expression patterns, we used the proteomes of 23 phylogenetically representative species and a species tree derived from the 1000K Plant initiative

(2019). Orthogroups (orthologous gene groups) were obtained using Orthofinder v2.4.0[29] (see material and methods) and their age (node in the species tree) was estimated using phylostratigraphy[30]. Briefly, for each orthogroup we searched its last common ancestor to place it to one node (phylostrata) of the species tree, where node 1 indicated the oldest phylostratum, and node 23 indicated the youngest, species-specific phylostratum (Supplementary Table 8). A total of 131,623 orthogroups were identified in the 23 Archaeplastida, of which 113,315 (86%) were species-specific, and the remaining 18,308 (14%) were assigned to internal nodes. Of these internal node orthogroups, most were ancestral (24% - node 1, 10% - node 3), represented the common ancestor of streptophytes (7%, node 6), land plants (7%, node 8), seed plants (10%, node 13), monocots (0.3%, node 18), or eudicots (1%, node 19) (Fig. 3a). Analysis of phylostrata in each species revealed a similar distribution of the orthogroups, with most of them belonging to node 1 (~34%) or being species-specific (~31%, Supplementary Fig. 6).

To investigate whether the different phylostrata show different expression trends, we surveyed orthogroups that contain at least two species with RNA-seq data, which resulted in 43,883 (33% of the total number of orthogroups) meeting this criterion. Then, each orthogroup was assigned to different expression profiles: ubiquitous (not specific in any organ), not conserved (e.g., root-specific in one species, flower-specific in others), or organ-specific (for details see material and methods, Supplementary Table 8 indicates expression profile of each orthogroup). The majority of the orthogroups in internal nodes (not species-specific) of the phylogenetic tree were assigned as ubiquitous (9,416), which corresponded to orthogroups that showed broad and not organ-specific expression (Fig. 3b). Interestingly, we observed a clear pattern of orthogroups becoming increasingly organ-specific as phylostratigraphic age decreased (<5% specific genes in node 1, vs. ~25% in node 13), indicating that younger orthogroups are recruited to specific organs (Fig. 3b). Using GO annotations of Arabidopsis genes with experimental evidence, we observed that organ-specific orthogroups have relevant functions for the assigned organ (Supplementary Table 9).

Next, we identified organ-specific orthogroups and investigated when they appeared during plant evolution. The number of orthogroups in internal nodes per organ varied from 12 (spore) to 228 (root), and we observed trends of organs across the internal nodes. In general, many organ-specific orthogroups

were present in nodes corresponding to monocots (Node 18, 20, 22). Expectedly, the 9,416 ubiquitous orthogroups were mostly of ancient (node 1-7) origin, suggesting that these old orthogroups tend to show a broader expression. The nonconserved groups had both old and more recent orthogroups. From the organ-specific families, leaves and spores were the groups containing more ancient families, while meristems had younger families. Flower, root, seeds, stem had few older families. Interestingly, when we compared male and female groups, we observed that the male-specific orthogroups had older orthogroups than the female-specific orthogroups (Fig. 3c).

Several studies revealed that new genes in animals tend to be preferentially expressed in male reproductive tissues, such as testis[31]. Similar observations have been made in *Arabidopsis*, rice, and soybean[32], where new genes were predominantly expressed in male reproductive cells[33], suggesting that these cells may act as an "innovation incubator" for the birth of *de novo* genes. Our gene expression data also revealed that male samples possess the youngest transcriptome in *Arabidopsis* (Fig. 3d, yellow bar), and in the male samples of *M. polymorpha*, *A. trichopoda*, *Z. mays*, *O. sativa*, *S. lycopersicum*, but not in *P. patens* (Fig. 3e, dark-blue cells for male, Supplementary Fig. 7). Pollen also expresses a substantial portion of old genes (species nodes 1-7 in Fig. 3c), probably representing an old transcription program present in gametes in Archaeplastida. With the unclear exception in *Physcomitrium*, we conclude that the observation that male samples express young genes is robust in the plant kingdom. However, we can not rule out the possibility of an underestimation of the age in male samples, since male-specific orthogroups seem to evolve fast (see 'Evolution of ubiquitous and organ-specific orthogroups'), and it has been observed that higher rates of evolution can lead to error in phylostratigraphic analysis [34].

Phylostratigraphic and gene expression analysis reveals that co-option drives the evolution of organs

The evolution of land plants involved many major innovations mediated by gains and losses of orthogroups and co-option of existing gene functions[20]. Most of the changes are related to land

adaptations comprising requirements for structural support, uptake of water, prevention of desiccation and gas exchange[35]. To better understand this complex process, we first analyzed the enrichment/depletion of organ-specific and ubiquitous genes in each node of the species tree (Supplementary Table 10). In line with previous results (Fig. 3b), ubiquitous genes were enriched for genes that appeared before the divergence of land plants and depleted for genes that appeared when plants colonized land (node 8, Fig. 4a). In line with the basal function (photosynthesis) of leaves, leaf-specific genes were enriched in ancestral nodes and the species-specific nodes of *M. polymorpha* (thallus samples) and *S. moellendorffii* (microphyll), and depleted in species-specific nodes of the seed plants (Fig. 4a).

Leaf-specific orthogroups were acquired mainly in two ancestral nodes, before the divergence of land plants and before the divergence of seed plants (Fig. 4b). Most of the orthogroups were gained in node 1 (34 families, Supplementary Table 11). Leaves have multiple origins in land plants[36], however, the programs for oxygenic photosynthesis originated in ancient organisms[37]. In agreement, before the divergence of land plants, we observed enrichment for functions related to photosynthesis (<node 8, before land plants), and after the divergence of land plants, we detected enrichment for additional functions such as external stimuli response, cytoskeleton organization, phytohormone action, and protein modification (Supplementary Table 12).

Interestingly, stem-, root-, and flower-specific genes shared a similar pattern and appeared to be enriched in nodes 4-8, 10-13, 15, and 20, and depleted in the species-specific nodes of vascular plants, except for *P. abies* for stems and *S. moellendorffii* for flowers. Although the origin(s) of roots, stems, and flowers are associated with vascular plants[38–40], we observed gene family expansions before the divergence of land plants (Fig. 4b) and in nodes as old as node 3 (2 orthogroups) for stems, node 1 (1 orthogroup) for roots, and node 3 (1 orthogroup) for flowers (Supplementary Table 11). Previous studies suggested that the evolution of novel morphologies was mainly driven by the reassembly and reuse of pre-existing genetic mechanisms, as exemplified by conserved transcriptional programs between flowers and cones in gymnosperms [36,41]. It was indicated that primitive root programs may have been present before the divergence of lycophytes and euphyllophytes[42]. Also, before the divergence of

charophytes from land plants, an ancestral origin was proposed for the SVP subfamily, which plays a crucial role in the control of flower development[43]. A recent study has shown that a moss (*Polytrichum commune*) possesses a vascular system functionally comparable to that of vascular plants[44]. These results support the idea that primitive stem-, root-, and flower-specific orthogroups existed prior to vascular plants' divergence. After the divergence of land plants, we can observe that there is incremental gene family gain in monocots for all three organs (roots, stems, flowers, Fig. 4b, indicated by red nodes), and also to a lesser extent in the ancestral node of seed plants. Specifically, for stem, we observed more gains in gymnosperms and more losses in eudicots. Functional enrichment analysis supports only enrichment in nodes corresponding to land plants (>node 8, before land plants) and not in older nodes (Supplementary Table 12).

Male-specific genes were enriched in angiosperms (node 15), monocots (node 20), eudicots (nodes 19, 21), and species-specific nodes, while female-specific genes were enriched only in monocots (nodes 18, 22), eudicots (node 19), and species-specific nodes (Fig. 4a). Additional male-specific families were gained in older nodes than female-specific families (intensity of the red color in the ancestral node of land plants, Fig. 4b). For male orthogroups, we observed six waves of gains (>15 orthogroups) in nodes 3, 8 (land plants), 13 (seed plants), 15 (angiosperms), 19 (eudicots), 20 (monocots). From these nodes, parallel to gains, we also observed many losses (>=10 orthogroups in three nodes 13 (seed plants), 15 (angiosperms), and 19 (eudicots) (Supplementary Table 11). For female-specific families, we observed three main waves of gains (>10 orthogroups) in nodes 13 (seed plants), 14 (gymnosperms), 20 (monocots), and different waves of losses (Supplementary Table 11). Male orthogroups showed enrichment for protein modification, enzyme classification, RNA biosynthesis, cell cycle organization, phytohormone action, and female orthogroups showed enrichment only for RNA biosynthesis (Supplementary Table 12). Considering gains and losses of orthogroups, male-specific families were gained mainly in the node ancestral to land plants, and in monocots, and for female-specific families in seed plants and gymnosperms (Fig. 4b).

In summary, the genetic programs for organ-specific genes are present in older nodes, before the divergence of land plants. Monocots seem to be the group with more gene family gains, which is in agreement with previous studies[45].

Evolution of ubiquitous and organ-specific orthogroups

Understanding the evolution of a gene is key to understanding the evolution of its function. We have observed that most of the organ-specific orthogroups appear early in evolution, before the divergence of land plants and the establishment of most organs (Fig. 4). Since gene duplication is considered an important source of functional innovation, we decided to test if organ-specific orthogroups experienced more duplications during their evolution than ubiquitously-expressed orthogroups. To test this, we used the ubiquitous and organ-specific orthogroups with a size of at least two sequences (13,329 orthogroups) and analysed the number of duplications observed (see material and methods). Interestingly, the number of duplications is much higher in orthogroups with a ubiquitous expression profile than any other organ-specific group (Supplementary Fig. 8a). Conversely, the organ-specific orthogroups predominantly show one or two duplications.

In order to test whether the organ-specific orthogroups evolve faster than ubiquitously-expressed orthogroups, we calculated the evolutionary rates as the ratio of nonsynonymous to synonymous substitution rates (dN/dS) for each single-copy orthogroup (see materials and methods). A total of 1,621 orthogroups were analysed and average pairwise dN, dS, and dN/dS was calculated for each group. Spore-specific orthogroups showed very high dS values (~35.7) and were not included in this analysis. The median dN/dS for ubiquitous and organ-specific orthogroups were less than 1, suggesting purifying selection (Supplementary Fig. 8b), which has been observed in previous studies[46,47]. When we compared the dN/dS distribution of ubiquitous genes against each of the organ-specific groups, we observed that male and stem orthogroups have significant lower median dN/dS (Wilcoxon rank sum test, P=1.4e-2 and 1.5e-2, respectively), and female and leaf orthogroups significant higher values (Wilcoxon rank sum test, P=3.4e-2 and 2.9e-2) (Supplementary Fig. 8b). For female and leaf orthogroups, the higher

dN/dS values observed were mainly due to a significant difference in the nonsynonymous substitution rate (dN, Supplementary Fig. 8c), which suggests higher rates of adaptive evolution. Interestingly, a recent study also observed higher dN/dS values in genes expressed in style and ovules in *Solanum* species, supporting our findings [48]. However, the lower dN/dS values observed in male and leaf are mainly explained by significantly higher synonymous substitution rates (dS), which is a proxy for mutation rate and could indicate that these orthogroups are evolving faster. Other studies showed that genes expressed in pollen tend to have lower dN/dS values than genes not expressed in pollen, which is attributed to stronger purifying selection on genes expressed in haploid gametophyte[49]. Furthermore, high dS values were observed in genes predominantly expressed in the sperm and pollen tube of *Arabidopsis*[32]. We observed that male samples express younger transcriptomes (TAI values, Fig. 3e) and since proteins that evolve rapidly could underestimate the phylostratigraphic age[34], we can not exclude the possible effect of this higher evolutionary rates in male orthogroups on the transcriptome age index (TAI). However, we also observed higher dS for seeds, stems, and roots (Supplementary Fig. 8d) and which was not met with high TAI values (Fig. 3e).

To study the relationship between the age and evolution of an orthogroup, we compared rates of evolution across the different nodes (phylostrata) of the species tree, and observed higher dN/dS, higher dN, and lower dS in younger nodes (Supplementary Fig. 8e,f,g and Supplementary Table 13). Interestingly, node 14 (gymnosperms) shows the highest median dN/dS and node 1, the lowest median value which is significantly different from younger nodes (Supplementary Table 13). We can observe that older orthogroups have significantly higher dS values, which points to fast evolving genes. Previous studies showed that older orthogroups have lower dN/dS, but did not observe large differences in dS values[46]. Worth to mention that monocots (node 20) seem to evolve faster than gymnosperms (node 14), and gymnosperms show significantly higher dN/dS than angiosperms (node 17 and 20) explained mainly for a major accumulation of nonsynonymous mutations. The difference in evolutionary rates between gymnosperms and angiosperms has been observed and discussed in previous studies[47].

Comparisons of transcriptional programs of gametes

Sexual reproduction is a complex process, requiring a dramatic reprogramming of the transcriptome during the diploid-to-haploid transition[50]. In diploid flowering plants, sexual reproduction involves the production of haploid male and female gametes and fertilization of the female ovule by male gametes mediated by pollination (Fig. 5a). The pollen delivers the sperm cell(s) to the ovary by a pollen tube, and the fertilized ovules grow into seeds within a fruit (Fig. 5a). The two haploid bryophytes in our study differ in their sexual reproduction. *Physcomitrium* is monoecious and bears both sperm and eggs on one individual (Fig. 5b), and *Marchantia* is dioecious and bears only egg or sperm, but never both (Fig. 5c). However, both species produce motile sperm that require water droplets to fertilize the egg, generating diploid zygotes. The zygotes divide by mitosis and grow into a diploid sporophyte. The sporophyte eventually produces specialized cells that undergo meiosis and produce haploid spores, which are released and germinate to produce haploid gametophytes (Fig. 5b,c).

To further study whether the transcriptional programs of sexual reproduction are conserved in land plants, we applied k-means clustering on the male- and female-specific genes over the RNA-seq samples representing different samples of male and female organs (Supplementary Table 1). For male-specific genes, the analysis assigned each sample to one or more clusters (Fig. 5d exemplifies male samples in *Arabidopsis* (for other species, see Supplementary Fig. 9), with a variable number of genes assigned to each cluster (Supplementary Table 14). We then inferred enriched biological processes (Fig. 5e, Supplementary Table 15), plotted average expression profiles (Fig. 5f), and used Jaccard distance to identify similar clusters across species (Fig. 5g). Interestingly, three clusters showed strong similarity and were specific to pollen tricellular, mature pollen, and pollen tube for Angiosperms (Fig. 5g, indicated by red lines). Functional enrichment analysis revealed that the corresponding samples were mainly enriched for cell wall organization, cytoskeletal organization, multi-process regulation, and protein modification (supported by five species, Fig. 5e). Conversely, other clusters showed enrichment for genes without assigned functions, and depletion for many biological processes (Fig. 5e).

Female samples included were less diverse than male samples. In all species, each sample was assigned to a cluster with the exception of *O. sativa*, where ovule is divided into two clusters (Supplementary Fig. 10, Supplementary Table 16, 17). Interestingly, when we measured the Jaccard distance among all clusters (including the species with one female sample), we observed no grouping of similar clusters, indicating that the female gamete transcriptomes were poorly conserved (Supplementary Fig. 10). Functional enrichment analysis showed enrichment mainly for not assigned functions and RNA processing, and depletion for many biological processes (Supplementary Fig. 10). The *G. biloba* ovule cluster (GINBI-0, ovule) showed enrichment for many functions, but ovule samples of other species did not support this observation. Despite the small number of samples included, these results provide evidence that female gamete transcriptomes are poorly conserved across the different species analyzed.

Analysis of signaling networks underpinning male gametophyte development and function

Gene co-expression networks help to identify sets of genes involved in related biological processes and highlight regulatory relationships[51]. Since we identified different gene clusters for male sub-samples (see above), we decided to test whether the genes assigned to different clusters are co-expressed. For this purpose, we reconstructed the co-expression networks of the ten species and analyzed whether the number of observed connections was similar to the number of expected connections (see material and methods). Interestingly, the clusters with expression profiles related to sperm had the least number of connections with other clusters for *O. sativa*, *Z. mays*, *A. trichopoda,* and *A. thaliana* (Fig. 6a). However, this pattern was not clear in *S. lycopersicum*, where the sperm cluster had connections with the cluster of generative cells. Specifically, for *A. thaliana* the co-expression network revealed that cluster C5 (sperm) is not well connected with other clusters (Fig. 6b), suggesting that the sperm cell transcriptome is distinctive, confirming earlier observations[52]. The connections between clusters followed a pattern from cluster C0 to C4, which highlighted the interaction of genes among the different developmental stages of male gametogenesis. The number of transcription factors and kinases present in the co-expression network changed among the different clusters, where transcription factors seemed

to be more abundant in cluster C0 (microspore), while kinases were more abundant in cluster C3 (mature pollen) (Fig. 6b, indicated by the sizes of rectangles, Supplementary Table 18).

Transcription factors and kinases are regulatory proteins essential for plant growth and development. To uncover the regulatory mechanism underlying male gametogenesis, we analyzed all the predicted transcription factors and kinases in all the male clusters of *A. thaliana*. First, we searched the literature describing the experimentally-verified function for all the transcription factors and kinases present in the five clusters (Supplementary Table 19). Then we classified the function of each gene as follows: no effect related to male gametogenesis (none), no experimentally described function (unknown), and important for microspore, bicellular, mature pollen, pollen tube, and sperm function. Interestingly, most of the genes are described as unknown (Fig. 6c), indicating no experiments associated with those genes. It is important to note that the genes classified as 'none' have been found to have an effect in other organs, but since pollen phenotype can be easily missed, this does not rule out the possibility of these genes being associated with male development. Also, many of those genes show effects in roots, and it has been shown that some genes are active during tip growth of root hairs and pollen tubes[53]. We observed that the transcription factors were important at different stages of male development, with main phenotypes affecting pollen tube and sperm function. Conversely, kinases only showed an effect on pollen tubes, which is in line with their intercellular communication involvement. Interestingly, we observed that genes present in the pollen tube cluster (ARATH-4) only affected pollen tube function, but pollen tube function can also be affected by genes from earlier stages of pollen development (ARATH1-3). In the case of sperm function, transcription factors expressed in tricellular pollen have the greatest effect, but we also observed the involvement of genes expressed in microspore, mature pollen and sperm (Fig. 6c).

Comparative gene expression analyses with the EVOREPRO database

To provide easy access to the data and analyses generated by our consortium, we have constructed an online database available at www.evorepro.plant.tools. The database is preloaded with the expression

data used in this study and also includes *Vitis vinifera* (eudicot, grapevine), *Chlamydomonas reinhardtii* (chlorophyte), and *Cyanophora paradoxa* (glaucophyte), bringing the total number of species to 13. The database can be queried with gene identifiers and sequences but also allows sophisticated, comparative analyses.

To showcase a typical user scenario, we identified genes specifically expressed in male organs (defined as, e.g., >35% reads of a gene expressed in male organs for *Arabidopsis*, Supplemental Fig. 1). This can be accomplished for one (https://evorepro.sbs.ntu.edu.sg/search/specific/profiles) or two (https://evorepro.sbs.ntu.edu.sg/specificity_comparison/) species, where the latter option can reveal specific expression profiles that are conserved across species (Fig. 7a). For this example, we selected *Arabidopsis* and *Amborella* as species A and B from the drop-down menus, respectively, and used orthogroups comprising only land plants, which uses all species found under node 8 in the species tree (Fig. 3a). Alternatively, the user can also select orthogroups constructed with seed plants (11 species found under node 13, Fig. 3a) or Archaeplastida (23 species found under node 1, Fig. 3a) sequences. Next, to select male organs for comparisons, we specified 'Tissue specificity' and 'Male' as a method to group the RNA-seq samples according to the definitions in Table 1. The slider near 'SPM cutoff' allows the user to adjust the SPM value (the slider ranges from SPM 0.5 to 1), which interactively reveals many genes are deemed organ-specific at a given SPM value cutoff. We left the slider at the default value (0.85) and clicked on the 'Compare specificities' button. The analysis revealed that 319 orthogroups are expressed specifically in the male organs of both *Amborella* and *Arabidopsis* (Fig. 7b), while the table below showed the identity of the genes and orthogroups (Fig. 7c, Supplementary Table 21). Interestingly, among the conserved genes, we observed *GCS1/HAP2*, which is required for pollen tube guidance and fertilization[54]. The table also contains links that redirect the user to pages dedicated to the genes and orthogroups. For example, clicking on the *Arabidopsis GCS1/HAP2* gene identifier redirects the user to a gene page containing the DNA/protein sequences (https://evorepro.sbs.ntu.edu.sg/sequence/view/17946), expression profile (Fig. 7d), gene family, co-expression network, and Gene Ontology functional enrichment analysis of the gene[55]. As expected, the interactive, exportable expression profiles confirmed that the *Arabidopsis GCS1/HAP2* and the

493 *Amborella* ortholog (https://evorepro.sbs.ntu.edu.sg/sequence/view/45084, Fig. 7e) are male-specific,
494 with the highest expression in sperm and pollen. Clicking on the gene family identifier
495 (OG_05_0008081) redirects to the gene family page
496 (https://evorepro.sbs.ntu.edu.sg/family/view/139708), which among others, contains an interactive
497 phylogenetic tree (Fig. 7f, https://evorepro.sbs.ntu.edu.sg/tree/view/88288) and heatmap (Fig. 7g,
498 https://evorepro.sbs.ntu.edu.sg/heatmap/comparative/tree/88288/row) showcasing the male- enriched
499 expression profiles for most of the genes in this family. Therefore, this approach can be used to identify
500 conserved, organ-specific genes across two species and study family-wide expression patterns.

501 Alternatively, the database can be used to identify conserved co-expression clusters of functionally
502 enriched genes. To demonstrate this tool, we navigated to
503 https://evorepro.sbs.ntu.edu.sg/search/enriched/clusters and entered 'pollen' into GO text box, selected
504 'pollen tube' as query and clicked on 'Show clusters'. The analysis revealed 5 co-expressed clusters
505 significantly ($P<0.05$) enriched for 'pollen tube' gene ontology term in *Arabidopsis*. We clicked on one
506 of the clusters (cluster 13, https://evorepro.sbs.ntu.edu.sg/cluster/view/113), redirecting us to a page
507 dedicated to the cluster. As expected, the cluster is significantly ($P<0.05$) enriched for genes involved
508 in pollen tube growth, cell wall organization and kinase activity, which are processes required to expand
509 and direct the pollen tube to the ovule. The page contains the identity of the 152 genes found in this
510 cluster, their average expression profiles, co-expression network
511 (https://evorepro.sbs.ntu.edu.sg/cluster/graph/113), and orthogroups and protein domains found in the
512 cluster.

513 Furthermore, a table labeled 'Similar Clusters' reveals the identity of similar (defined by Jaccard index,
514 see methods) co-expression clusters in other species, which can be used to identify functionally
515 equivalent clusters across species rapidly. To exemplify this, we first clicked on 'Jaccard index' table
516 header to sort the similar clusters and clicked on the 'Compare' link next to Cluster 39 from *Amborella*
517 (https://evorepro.sbs.ntu.edu.sg/graph_comparison/cluster/113/769/1). This redirected us to a co-
518 expression network page showing the genes (nodes), co-expression relationships (gray edges), and
519 orthologous genes (colored shapes of nodes connected by dashed edges) conserved in the two clusters.

The analysis revealed many conserved genes essential for pollen function, such as *ANX2*[56], *BUPS2* (*At2g21480*)[57], *PI4K Gamma-1*[58], *PTEN1*[59], *RIC1*[60], and *ATM1*[61]. To conclude, this approach can be used to uncover functionally equivalent, conserved transcriptional programs.

Discussion

To study the evolution of plant organs and gametes, we have generated and analyzed gene expression for ten land plants, comprising representatives of bryophytes, lycophytes, gymnosperms, sister to all angiosperms, monocots and eudicots. Our analyses' main advantage is that the conclusions are drawn from comparative analyses of ten species, which cover the largest collection of representatives of land plants. The comparative analysis revealed that each organ type typically expressed >50% of genes, with the exception of the male gametes, which showed expression of ~38% of genes, on average (Fig. 1d). Conversely, male gametes and roots showed the highest number (5.3% and 5.0%, respectively) of specifically expressed genes (Fig. 1f), suggesting that these non-photosynthesizing cell types and tissues are highly unique and specialized.

Despite the substantial heterogeneity of the growth conditions of the plants, the different developmental stages of the sampled organs, and different representation of the various tissues found in the organs (e.g., buds, stamen filaments, carpels in *Arabidopsis* vs. whole flowers in tomato, Table 1) we observed a significant and robust conservation of the transcriptional programs of the analyzed organs. With the surprising exception of female gametes, the corresponding transcriptomes tend to be more similar across the analyzed samples (Fig. 2b, Supplementary Fig. 3, Supplementary Fig. 4). As also observed in previous studies, roots, male, and seeds express conserved expression programs[42,62]. Another exception is seen in the leaf-like organs of bryophytes (leaflets and thallus for *Physcomitrium* and *Marchantia*, respectively), indicating that these organs have evolved independently from the leaves of flowering plants or that they have significantly diverged since the last common ancestor of flowering plants and bryophytes.

Next, we examined expression patterns of expressed orthogroups as a function of their age. We report a clear trend of older orthogroups having more ubiquitous (i.e., less organ-specific) expression, while younger orthogroups show an increasingly higher proportion of organ-specific expression (Fig. 3b-c). This indicates that newly-acquired genes are typically recruited to perform some specialized function in a plant organ, tissue, or cell type, rather than being integrated into fundamental biological pathways. As expected, male gametes show the highest expression of the youngest genes (Fig. 3d-e, Supplementary Fig. 7), which is in line with previous studies[33,63]. Interestingly, *Physcomitrium* gametes did not show this pattern, which is a finding that warrants further studies.

To study how new functions were gained or lost as the organs and gametes evolved, we studied which phylostrata are enriched or depleted in the different organs (Fig. 4a). Interestingly, we observe a significant enrichment for orthogroups that appeared long before the corresponding organ (Fig. 4a), showing that the establishment of organs relies heavily on the co-option of existing genetic material, as suggested previously[20,36,41]. Flowers (appearance in angiosperms), stems (appearance in vascular plants) and roots (appearance in vascular/seed plants) show similar patterns of enrichment and depletion of genes (Fig. 4a). This is surprising, as these organs appeared at different stages of plant evolution, which suggests that the co-option underlying the establishment of novel organs follows a similar pattern of gene gains and losses. Based on the diverse patterns of gains and losses of organ-specific orthogroups (Fig. 4b) we conclude that monocot-specific families show substantial net gains in genes that are specifically expressed in male gametes, seeds, stems, roots or in apical and root meristems (Fig. 4b), suggesting that during monocots evolution organ-specific transcriptomes were enriched with novel functions. Surprisingly, eudicots show an opposite pattern, exhibiting more net losses of organ-specific families in flowers, female and male gametes, leaves, stems, roots, and apical meristems (Fig. 4b). Similar patterns of gene losses were also observed in two major groups of the animal kingdom (Ecdysozoa and Deuterostomia), suggesting that reductive evolution of protein coding genes plays a major role in shaping genome evolution[64]. This surprising pattern of loss of functions in eudicots merits investigation by further analysis, which is made possible by identifying the corresponding orthogroups (Supplementary Table 11) and genes (Supplementary Table 8).

Our comparative analysis of male gamete development reveals that transcriptional programs of mature pollen form well-defined clusters and are thus conserved across species (Fig. 5f-g). The mature pollen clusters are enriched for processes related to signaling (protein modification comprising protein kinases) and cell wall remodeling (Fig. 5e), which are likely representing processes mediating pollen germination, pollen tube growth, and sperm cell delivery. Conversely, the earlier stages of male gamete development showed less defined clusters and enrichment for genes with unknown function (bin 'not assigned', Fig. 5e), suggesting that the processes taking place in the early stages of pollen development are yet to be uncovered. Furthermore, the female gametes show poor clustering, indicating overall low conservation of the transcriptional programs and enrichment of genes with unknown function for most clusters (Supplementary Fig. 10c). These results indicate that genes expressed during early male gamete and female gamete formation warrant closer functional analysis, which is now made possible by our identification of these genes (Supplementary Table 14, 16). Of particular interest are the male-specific transcription factors and kinases that we identified (Fig. 6c), assumingly involved in various stages of pollen development and function (Supplementary Table 19). As a large fraction of these genes are not yet characterized, their involvement in male gametogenesis and function should be further investigated.

To provide easy access to the 13 expression atlases, organ-specific genes, functional enrichment analyses, co-expression networks, and various comparative tools, we provide the EVOREPRO database (www.evorepro.plant.tools) to the community (Fig. 7). This database represents a valuable resource for further study and validation of key genes involved in organogenesis and land plants reproduction.

An even deeper understanding of the origin and evolution of plant organs will require an analysis of more plant species (especially streptophyte algae, ferns and gymnosperms), together with inclusion of information about the presence of non-coding DNA (e.g., cis-regulatory elements) and non-coding RNA (e.g., sRNAs, miRNAs).

Methods

Plant growth, RNA isolation and sequencing

The protocols used to generate RNA-sequencing data for *Physcomitrium*, *Marchantia*, tomato, maize, *Arabidopsis* and *Amborella* are described in Supplementary Methods.

Compiling gene expression atlases

RNA data of different samples from nine species (*Physcomitrium patens*, *Marchantia polymorpha*, *Ginkgo biloba*, *Picea abies*, *Amborella trichopoda*, *Oryza sativa*, *Zea mays*, *Arabidopsis thaliana*, *Solanum lycopersicum*) were grouped in ten different classes (organs) (flower, female, male, seeds, spore, leaf, stem, apical meristem, root meristem, root) (Table 1, Supplementary Table 1). For male and female reproductive organs samples we also included different sub-samples (female: egg cell, ovary, ovule; Male: microspore, bicellular pollen, tricellular pollen, mature pollen, pollen tube, generative cell, sperm) for each species (Table 1, Supplementary Table 1). A total of 4,806 different RNA sequencing samples were used, from which 4,672 were downloaded from the SRA database and 134 obtained from our experiments (see above). Publicly available RNA-seq experiments data were downloaded from ENA (https://www.ebi.ac.uk/ena/browser/home). For more details, see Supplementary Methods.

Identifying organ-specific genes

Organ-specific genes based on expression data were detected by calculating the specificity measure (SPM), using a similar method as described in[65]. For each gene, we calculated the average TPM value in each sample (e.g., root, leaf, seeds). Then, the SPM value of a gene in a sample was computed by dividing the average TPM in the sample by the sum of the average TPM values of all samples. The SPM value ranges from 0 (a gene is not expressed in a sample) to 1 (a gene is fully sample-specific). To identify sample-specific genes, for each of the ten species, we first identified a SPM value threshold above which the top 5% SMP values were found (Supplementary Fig. 1, red line). Then, if a gene's SPM value in a sample was equal to or larger than the threshold, the gene was deemed to be specifically expressed in this sample.

Similarity of organ-specific transcriptomes between samples and species

To estimate whether organ-specific transcriptomes (see above) are similar, we calculated Jaccard distance d_J between orthogroup sets. These orthogroup sets were found by identifying the orthogroups of organ-specific genes per each species. Then pairwise d_J was calculated for all the samples and used as input for the clustermap. The d_j ranges between 0 (the two sets of orthogroups are identical) to 1 (the two sets have no orthogroups in common).

To estimate whether a species' organ-specific transcriptome was significantly similar to a corresponding sample in the other species (e.g. *Arabidopsis* root vs. rice root, tomato root), we tested whether the $d_J$values comparing the same sample were smaller (i.e. more similar) than d_J values comparing the sample to the other samples (e.g., *Arabidopsis* root vs. rice flower, rice leaf, tomato flower, tomato leaf). We used Wilcoxon rank-sum to obtain the p-values, which were adjusted using a false discovery rate (FDR) correction[66] using a cutoff of 0.05.

Phylogenomic and phylostratigraphic analysis

We used proteomes of 23 species representing key phylogenetic positions in the plant kingdom (see Supplementary Table 20) to construct orthologous gene groups (orthogroups) with Orthofinder v2.4.0[29]. A species tree, of the 23 individuals, based on a recent phylogeny including more than 1000 species[67] was used for the phylostratigraphic analysis. The phylostratum (node) of an orthogroup was assessed by identifying the oldest clade found in the orthogroup using ETE v3.0[68]. For more details, see Supplementary Methods.

Transcriptomic age index calculation

Transcriptome age index (TAI) is the weighted mean of phylogenetic ranks (phylostrata) and we calculated it for every sample[63]. We used the species tree from One Thousand Plant Transcriptomes

Initiative, 2019[67]. The nodes in the tree were assigned numbers ranging from 1 (oldest node) to 22 (youngest node, Fig. 3a) by traversing the tree using ETE v3.0 (Huerta-Cepas et al. 2016) with default parameters. The age (phylostratum) of an orthogroup and all genes belonging to the orthogroup, were derived by identifying the last common ancestor found in the orthogroup using ETE v3.0[68]. In the case of species-specific orthogroups the age of the orthogroup was assigned as 23. Finally, all genes with TPM values <2 were excluded and the TAI was calculated for the remaining genes by dividing the product of the gene's TPM value and the node number by the sum of TPM values.

Functional annotation of genes and identification of transcription factor and kinase families

The proteomes of the ten species included in the transcriptome dataset were annotated using the online tool Mercator4 v2.0 (https://www.plabipd.de/portal/web/guest/mercator4/-/wiki/Mercator4/recent_changes). Transcription factors and kinases were predicted using iTAK v1.7a[69]. Additional transcription factors were identified using the online tool PlantTFDB v5.0 (http://planttfdb.cbi.pku.edu.cn/prediction.php)[70]. For more details, see Supplemental Methods.

Functional enrichment analysis

Functional enrichment of the list of organ-specific and cluster-specific genes of each species, and genes gained in each node, was calculated using the bins predicted with Mercator 4 v2.0. Briefly, for a group of m genes (e.g., genes specifically expressed in Arabidopsis root), we first counted the number of Mapman bins present in the group, and then evaluated if these bins were significantly enriched or depleted by calculating an empirical p-value. Transcription factor and kinase enrichment was calculated following the same procedure. For more details, see Supplemental Methods.

Identification of orthogroup expression profiles

In order to analyse the expression profiles at phylostrata level, orthogroups were classified as ‘organ-specific’, 'ubiquitous', and 'not conserved'. ‘Organ-specific’ orthogroups are orthogroups containing organ-specific genes and can be sub-classified according to the organ (flower-, female-, male-, seeds-, spore-, leaf-, apical meristem-, stems-, root meristem-, root-specific). ‘Ubiquitous’ are orthogroups that are expressed in different organs for each species, i.e., they do not show a ‘organ-specific’ expression profile. ‘Not conserved’ are orthogroups that have different organ-specific expression profiles in different species (e.g., orthogroups containing root-specific genes for *Arabidopsis* and male-specific genes for *Solanum*). Only orthogroups with species with sufficient expression data were used. More specifically, we only analyzed orthogroups that were: i) species-specific with transcriptome data or, ii) contained at least two species with transcriptome data. To identify organ-specific orthogroups, we required , iii) >50% of genes of the orthogroup should support the expression profile, iv) >=50% of the species with transcriptome data present in the node should support the expression profile.

Gene enrichment analysis per phylostrata

In order to analyse gene enrichment of specific organs across the different phylostrata in the species tree (Fig. 3a), we used all the organ-specific genes of the ten species included. For each species and for each defined sample (ubiquitous, flower, female, male, seeds, spore, leaf, stem, apical meristem, root meristem, root) we counted the number of genes present in each node of the species tree, and then evaluated if the number of organ-specific genes were significantly enriched or depleted by calculating an empirical p-value as described for functional enrichment analysis. Then, we evaluated each organ and counted the number of species that show significant enrichment/depletion ($p<0.05$) in each node of the species tree. We obtained a normalized value per each node by calculating the difference of species showing enrichment and species showing depletion and dividing it by the total number of species that show enrichment/depletion.

Gene family comparisons

For each organ-specific (flower, female, male, seeds, spore, leaf, stem, apical meristem, root meristem, root) and ubiquitous expression profiles we mapped loss and gain of organ-specific orthogroups onto the species tree (Fig. 3a). All the orthogroups classified as organ-specific (see above) were analysed independently and gain and loss was computed using the approach described in[71] with ETE v3.0[68]. Briefly, a gene family gain was inferred at the last common ancestor of all the species included in the family and a loss when a species did not have orthologs in the particular gene family. Groups of monophyletic species that have lost the gene were counted as one loss. Then, we collapsed the values of the nodes of the species tree to fit the different clades included (Fig. 4b), and we calculated the difference between the total gains and the total losses to obtain an absolute value for each node. The values of each expression profile were normalized dividing the values by the maximum absolute value in a way that we got a range from -1 to 1 (negative values for losses and positive values for gains). Finally, per each expression profile (ubiquitous, flower, female, male, seeds, spore, leaf, stem, apical meristem, root meristem, root) a graphical representation of the different clades showing the nodes with a intensity of color proportional to the normalized values of gains and losses was plotted using ETE v3.0[68].

Gene duplications and evolutionary rates of ubiquitous and organ-specific orthogroups

To analyse gene duplication, ubiquitous and organ-specific orthogroups with at least two sequences (13,329) were selected. The orthogroups with two sequences (2,188) were analysed separately, and if the two sequences belonged to the same species, one duplication was assumed. For each orthogroup with at least three sequences (11,141) gene trees were reconstructed. The protein sequences of each orthogroup were aligned using the same approach as described in the PhylomeDB pipeline[72] and phylogenetic trees were built using IQ-TREE v2.1.2[73]. For more details, see Supplemental Methods.

Identification of gamete-specific transcriptional profiles by clustering analysis

We analyzed the male and female organ-specific genes and their different sub-samples (Supplementary Table 1), to identify transcriptional profiles by clustering analysis. For the clustering analysis we only included species with at least two sub-samples (*Amborella trichopoda*, *Oryza sativa*, *Zea mays*, *Arabidopsis thaliana*, *Solanum lycopersicum*). The male samples were divided into: microspore, bicellular pollen, tricellular pollen, mature pollen, pollen tube, generative cell, and sperm cell for Angiosperms; and sperm for bryophytes. The female samples were divided into egg cell, ovary, and ovule. For each gene, the average TPM in each sub-sample was calculated, and the average TPM values were scaled by dividing with the highest average TPM value for the gene. The k-means clustering method from the sklearn.cluster package was used to fit the scaled average TPM values to the number of clusters (*k*) ranging from 1 to 20. The sklearn.cluster package contains multiple methods to evaluate the influence of the clustering parameters, and we used the elbow method to find the optimal number *k*, where *k* that produced a sum of squared distances<80% of *k*=1 was selected (Supplementary Fig. 11).

Constructing the co-expression network and establishing the EVOREPRO database

Coexpression networks were calculated using the CoNekT framework[55], which was also used to establish the EVOREPRO database available at www.evorepro.plant.tools. For each species, all the genes that were co-expressed in each male cluster were analysed to test whether the number of connections observed is similar to the expected number. For this, we divided the number of observed connections between the genes of two clusters (eg. cluster 1 and cluster 2) by the expected value (product of the number of genes in cluster 1 x number of genes in cluster 2). These values were used to perform a pearson correlation analysis and the results were presented in heatmaps. The networks present in the male clusters were visualized using Cytoscape v3.8.0[74]. The network files are available from www.evorepro.plant.tools/species/.

Data availability

The fastq files are available for *Arabidopsis* (E-MTAB-9456), *Amborella* (E-MTAB-9190), *Marchantia* (E-MTAB-9457), *Physcomitrium* (E-MTAB-9466), maize (E-MTAB-9692) and tomato (E-MTAB-9725). The data can be obtained from https://www.ebi.ac.uk/ena.

References

1. Jill Harrison, C. Development and genetics in the evolution of land plant body plans. *Philos. Trans. R. Soc. Lond. B. Biol. Sci* **372**, (2017).
2. Fürst-Jansen, J. M. R., de Vries, S. & de Vries, J. Evo-physio: on stress responses and the earliest land plants. *J. Exp. Bot.* **71**, 3254–3269 (2020).
3. Brown, R. C. & Lemmon, B. E. Spores before sporophytes: hypothesizing the origin of sporogenesis at the algal-plant transition. *New Phytol.* **190**, 875–881 (2011).
4. Edwards, D., Morris, J. L., Richardson, J. B. & Kenrick, P. Cryptospores and cryptophytes reveal hidden diversity in early land floras. *New Phytol.* **202**, 50–78 (2014).
5. Kenrick, P. & Crane, P. R. The origin and early evolution of plants on land. *Nature* **389**, 33–39 (1997).
6. Berner, R. A. GEOCARBSULF: A combined model for Phanerozoic atmospheric O2 and CO2. *Geochim. Cosmochim. Acta* **70**, 5653–5664 (2006).
7. Beerling, D. J., Osborne, C. P. & Chaloner, W. G. Evolution of leaf-form in land plants linked to atmospheric CO2 decline in the Late Palaeozoic era. *Nature* **410**, 352–354 (2001).
8. Menand, B. *et al.* An ancient mechanism controls the development of cells with a rooting function in land plants. *Science* **316**, 1477–1480 (2007).
9. Hater, F., Nakel, T. & Groß-Hardt, R. Reproductive multitasking: the female gametophyte. *Annu. Rev. Plant Biol.* **71**, 517–546 (2020).
10. Hackenberg, D. & Twell, D. The evolution and patterning of male gametophyte development. *Curr. Top. Dev. Biol.* **131**, 257–298 (2019).
11. Amici, G. B. Observations microscopiques sur diverses espèces de plantes. *Ann Sei Nat Bot* **2**, 211–248 (1824).

12. Johnson, M. A., Harper, J. F. & Palanivelu, R. A Fruitful Journey: Pollen Tube Navigation from Germination to Fertilization. *Annu. Rev. Plant Biol.* **70**, 809–837 (2019).
13. Sprunck, S. Twice the fun, double the trouble: gamete interactions in flowering plants. *Curr. Opin. Plant Biol.* **53**, 106–116 (2020).
14. Borg, M. *et al.* The R2R3 MYB transcription factor DUO1 activates a male germline-specific regulon essential for sperm cell differentiation in Arabidopsis. *Plant Cell* **23**, 534–549 (2011).
15. Favery, B. *et al.* KOJAK encodes a cellulose synthase-like protein required for root hair cell morphogenesis in Arabidopsis. *Genes Dev.* **15**, 79–89 (2001).
16. Denninger, P. *et al.* Male-female communication triggers calcium signatures during fertilization in Arabidopsis. *Nat. Commun.* **5**, 4645 (2014).
17. Borges, F. *et al.* FACS-based purification of Arabidopsis microspores, sperm cells and vegetative nuclei. *Plant Methods* **8**, 44 (2012).
18. Borg, M. *et al.* An EAR-Dependent Regulatory Module Promotes Male Germ Cell Division and Sperm Fertility in Arabidopsis. *Plant Cell* **26**, 2098–2113 (2014).
19. Cyprys, P., Lindemeier, M. & Sprunck, S. Gamete fusion is facilitated by two sperm cell-expressed DUF679 membrane proteins. *Nat. Plants* **5**, 253–257 (2019).
20. Bowles, A. M. C., Bechtold, U. & Paps, J. The origin of land plants is rooted in two bursts of genomic novelty. *Curr. Biol.* **30**, 530-536.e2 (2020).
21. Rhee, S. Y. & Mutwil, M. Towards revealing the functions of all genes in plants. *Trends Plant Sci.* **19**, 212–221 (2014).
22. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
23. Pina, C., Pinto, F., Feijó, J. A. & Becker, J. D. Gene family analysis of the Arabidopsis pollen transcriptome reveals biological implications for cell growth, division control, and gene expression regulation. *Plant Physiol.* **138**, 744–756 (2005).
24. Steffen, J. G., Kang, I.-H., Macfarlane, J. & Drews, G. N. Identification of genes expressed in the Arabidopsis female gametophyte. *Plant J.* **51**, 281–292 (2007).
25. Bowman, J. L. The YABBY gene family and abaxial cell fate. *Curr. Opin. Plant Biol.* **3**, 17–22

(2000).

26. Kim, J. H. & Lee, B. H. GROWTH-REGULATING FACTOR4 ofArabidopsis thaliana is required for development of leaves, cotyledons, and shoot apical meristem. *J. Plant Biol.* **49**, 463–468 (2006).

27. Ding, Z. J. *et al.* Transcription factor WRKY46 modulates the development of Arabidopsis lateral roots in osmotic/salt stress conditions via regulation of ABA signaling and auxin homeostasis. *Plant J.* **84**, 56–69 (2015).

28. Long, T. A. *et al.* The bHLH transcription factor POPEYE regulates response to iron deficiency in Arabidopsis roots. *Plant Cell* **22**, 2219–2236 (2010).

29. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

30. Domazet-Loso, T., Brajković, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).

31. Begun, D. J., Lindfors, H. A., Kern, A. D. & Jones, C. D. Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. *Genetics* **176**, 1131–1137 (2007).

32. Gossmann, T. I., Saleh, D., Schmid, M. W., Spence, M. A. & Schmid, K. J. Transcriptomes of Plant Gametophytes Have a Higher Proportion of Rapidly Evolving and Young Genes than Sporophytes. *Mol. Biol. Evol.* **33**, 1669–1678 (2016).

33. Cui, X. *et al.* Young Genes out of the Male: An Insight from Evolutionary Age Analysis of the Pollen Transcriptome. *Mol. Plant* **8**, 935–945 (2015).

34. Moyers, B. A. & Zhang, J. Further simulations and analyses demonstrate open problems of phylostratigraphy. *Genome Biol. Evol.* **9**, 1519–1527 (2017).

35. Doyle, J. A. Phylogenetic analyses and morphological innovations in land plants. in *Annual Plant Reviews* (eds. Roberts, J. A., Evan, D., McManus, M. T. & Rose, J. K. C.) 1–50 (John Wiley & Sons, Ltd, 2018). doi:10.1002/9781119312994.apr0486.

36. Pires, N. D. & Dolan, L. Morphological evolution in land plants: new designs with old genes. *Philos. Trans. R. Soc. Lond. B. Biol. Sci* **367**, 508–518 (2012).

37. Cardona, T. Thinking twice about the evolution of photosynthesis. *Open Biol.* **9**, 180246 (2019).

38. Harrison, C. J. & Morris, J. L. The origin and early evolution of vascular plant shoots and leaves. *Philos. Trans. R. Soc. Lond. B. Biol. Sci* **373**, (2018).

39. Hetherington, A. J. & Dolan, L. Stepwise and independent origins of roots among land plants. *Nature* **561**, 235–238 (2018).

40. Specht, C. D. & Bartlett, M. E. Flower Evolution: The Origin and Subsequent Diversification of the Angiosperm Flower. *Annu. Rev. Ecol. Evol. Syst.* **40**, 217–243 (2009).

41. Pires, N. D. *et al.* Recruitment and remodeling of an ancient gene regulatory network during land plant evolution. *Proc Natl Acad Sci USA* **110**, 9571–9576 (2013).

42. Huang, L. & Schiefelbein, J. Conserved Gene Expression Programs in Developing Roots from Diverse Plants. *Plant Cell* **27**, 2119–2132 (2015).

43. Tanabe, Y. *et al.* Characterization of MADS-box genes in charophycean green algae and its implication for the evolution of MADS-box genes. *Proc Natl Acad Sci USA* **102**, 2436–2441 (2005).

44. Brodribb, T. J., Carriquí, M., Delzon, S., McAdam, S. A. M. & Holbrook, N. M. Advanced vascular function discovered in a widespread moss. *Nat. Plants* **6**, 273–279 (2020).

45. Ruprecht, C. *et al.* Phylogenomic analysis of gene co-expression networks reveals the evolution of functional modules. *Plant J.* **90**, 447–465 (2017).

46. Guo, Y.-L. Gene family evolution in green plants with emphasis on the origination and evolution of Arabidopsis thaliana genes. *Plant J.* **73**, 941–951 (2013).

47. Buschiazzo, E., Ritland, C., Bohlmann, J. & Ritland, K. Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC Evol. Biol.* **12**, 8 (2012).

48. Moyle, L. C., Wu, M. & Gibson, M. J. S. Reproductive proteins evolve faster than non-reproductive proteins among *Solanum* species. *BioRxiv* (2020) doi:10.1101/2020.11.30.405183.

49. Chibalina, M. V. & Filatov, D. A. Plant Y chromosome degeneration is retarded by haploid purifying selection. *Curr. Biol.* **21**, 1475–1479 (2011).

50. Borg, M. *et al.* Epigenetic reprogramming rewires transcription during the alternation of

858 generations in Arabidopsis. *elife* **10**, (2021).
859 51. Rao, X. & Dixon, R. A. Co-expression networks for plant biology: why and how. *Acta Biochim*
860 *Biophys Sin (Shanghai)* **51**, 981–988 (2019).
861 52. Borges, F. *et al.* Comparative transcriptomics of Arabidopsis sperm cells. *Plant Physiol.* **148**,
862 1168–1181 (2008).
863 53. Becker, J. D., Takeda, S., Borges, F., Dolan, L. & Feijó, J. A. Transcriptional profiling of
864 Arabidopsis root hairs and pollen defines an apical cell growth signature. *BMC Plant Biol.* **14**, 197
865 (2014).
866 54. von Besser, K., Frank, A. C., Johnson, M. A. & Preuss, D. Arabidopsis HAP2 (GCS1) is a sperm-
867 specific gene required for pollen tube guidance and fertilization. *Development* **133**, 4761–4769
868 (2006).
869 55. Proost, S. & Mutwil, M. CoNekT: an open-source framework for comparative genomic and
870 transcriptomic network analyses. *Nucleic Acids Res.* **46**, W133–W140 (2018).
871 56. Boisson-Dernier, A. *et al.* Disruption of the pollen-expressed FERONIA homologs ANXUR1 and
872 ANXUR2 triggers pollen tube discharge. *Development* **136**, 3279–3288 (2009).
873 57. Zhu, L. *et al.* The Arabidopsis CrRLK1L protein kinases BUPS1 and BUPS2 are required for
874 normal growth of pollen tubes in the pistil. *Plant J.* **95**, 474–486 (2018).
875 58. Alves-Ferreira, M. *et al.* Global expression profiling applied to the analysis of Arabidopsis stamen
876 development. *Plant Physiol.* **145**, 747–762 (2007).
877 59. Gupta, R., Ting, J. T. L., Sokolov, L. N., Johnson, S. A. & Luan, S. A tumor suppressor homolog,
878 AtPTEN1, is essential for pollen development in Arabidopsis. *Plant Cell* **14**, 2495–2507 (2002).
879 60. Zhou, Z. *et al.* Arabidopsis RIC1 severs actin filaments at the apex to regulate pollen tube growth.
880 *Plant Cell* **27**, 1140–1161 (2015).
881 61. Liang, Y. *et al.* MYB97, MYB101 and MYB120 function as male factors that control pollen tube-
882 synergid interaction in Arabidopsis thaliana fertilization. *PLoS Genet.* **9**, e1003933 (2013).
883 62. Szövényi, P., Waller, M. & Kirbis, A. Evolution of the plant body plan. *Curr. Top. Dev. Biol.* **131**,
884 1–34 (2019).
885 63. Domazet-Lošo, T. & Tautz, D. A phylogenetically based transcriptome age index mirrors

ontogenetic divergence patterns. *Nature* **468**, 815–818 (2010).

64. Guijarro-Clarke, C., Holland, P. W. H. & Paps, J. Widespread patterns of gene loss in the evolution of the animal kingdom. *Nat. Ecol. Evol.* **4**, 519–523 (2020).

65. Xiao, S.-J., Zhang, C., Zou, Q. & Ji, Z.-L. TiSGeD: a database for tissue-specific genes. *Bioinformatics* **26**, 1273–1275 (2010).

66. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

67. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).

68. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

69. Zheng, Y. *et al.* iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. *Mol. Plant* **9**, 1667–1670 (2016).

70. Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J. & Gao, G. PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res.* **48**, D1104–D1113 (2020).

71. Ballester, A.-R. *et al.* Genome, Transcriptome, and Functional Analyses of Penicillium expansum Provide New Insights Into Secondary Metabolism and Pathogenicity. *Mol. Plant Microbe Interact.* **28**, 232–248 (2015).

72. Huerta-Cepas, J. *et al.* PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res.* **39**, D556-60 (2011).

73. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

74. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

Acknowledgments

I.J is supported by Singaporean Ministry of Education grant MOE2018-T2-2-053, while M.M is supported by NTU Start-Up Grant. ERA-CAPS EVO-REPRO I2163 and FWF grant P30802 to F.B.; FCT ERA-CAPS-0001-2014 and PTDC-BIA-FBT-28484-2017 to J.D.B; ERA-CAPS EVO-REPRO DR 334/12-1 to S.S. and T.D. DH was supported by ERA-CAPS UK Biotechnology and Biological Research Council Grant BB/N005090 awarded to DT; M.B. was supported through the FWF Lise Meitner fellowship M1818. The Vienna BioCenter Core Facilities GmbH (VBCF) Plant Sciences Facility acknowledges funding from the Austrian Federal Ministry of Education, Science and Research and the City of Vienna. L.S was supported by CSF grant 17-23183S. C.M. and D.Ho. were supported by Czech Ministry of Education, Youth and Sport (LTC18034 and LTAIN19030) through the European Regional Development Fund-Project "Centre for Experimental Plant Biology": No. CZ.02.1.01/0.0/0.0/16_019/0000738.The Genomics Unit of Instituto Gulbenkian de Ciência was partially supported by ONEIDA Project (LISBOA-01-0145-FEDER-016417) co-funded by FEEI - "Fundos Europeus Estruturais e de Investimento" from "Programa Operacional Regional Lisboa 2020" and by national funds from FCT - "Fundação para a Ciência e a Tecnologia". C.S.M acknowledges a doctoral fellowship from FCT (PD/BD/114362/2016) under the Plants for Life PhD Program. J.D.B received salary support from FCT through an "Investigador FCT" position. MJ and JG were supported by a US National Science Foundation grant (IOS-1540019).

Help with sample generation: Lenka Záveská Drábková and David Reňák. Marchantia growth was performed by the Plant Sciences Facility at Vienna BioCenter Core Facilities GmbH (VBCF), member of the Vienna BioCenter (VBC), Austria. Maximilian Weigend, Cornelia Löhne and Bernhard Reinken (Botanical Garden of the University of Bonn, Germany) are acknowledged for providing *Amborella trichopoda* plant material. Devendra Shivhare is acknowledged for a preliminary analysis of *Physcomitrium* RNA-seq data.

We would like to thank Debbie Maizels (http://www.scientificart.com) for the illustrations on Fig.1 and Fig. 5.

Author Contributions

Conceived and designed the analysis: JDB, MM

Collected the data: ACL, MFT, SGP, CSM, IJ, LS, CM, DHo, DH

Contributed data or analysis tools: FB, MB, SS, TD, DT

Performed the analysis: IJ, CF, SP, ACL, MM

Wrote the paper: IJ, JDB, MM

Competing interests

The authors declare no competing interests.

Figure legends

Fig. 1: Expression atlases for seven land plant species. Depiction of the different organs, tissues, and cells collected for (**a**) *P. patens* (**b**) *Marchantia polymorpha*, and (**c**) angiosperms. **d**, The percentage of genes (x-axis) found to be expressed (defined as TPM>2) in organs (y-axis) of the different species (indicated by colored bars as in (f)). The numbers beneath the organs (y-axis) indicate the average percentage of genes for all species. **e**, Expression profiles of organ-specific genes from *Arabidopsis thaliana*. Genes are in rows, organs in columns and the genes are sorted according to the expression profiles (e.g., flower, female). The numbers at the top of each column indicate the total number of genes per organ. Gene expression is scaled to range from 0-1. Bars on the left of each heatmap show the organ-specific genes and correspond to the samples on the bottom: pink - Flower, purple - Female, yellow - Male, orange - Seeds/Spore, dark-green - Leaf, medium-green - Stem, light-green - Apical meristem, blue - Root meristem, brown - Root. **f**, Percentage of organ-specific *Arabidopsis* genes with PO annotations for the 10 organs. The 'Others' category indicates the genes with annotations that could correspond to more than one organ or samples not included in this study. **g**, The percentage of genes with specific expression in the ten species.

Fig. 2: Comparison of organ-specific transcriptomes. a, Bar plot showing the Jaccard distances (y-axis) when comparing the same samples (x-axis, e.g., male-male) and one sample versus the others (e.g., male-others) for *Arabidopsis thaliana*. Lower values indicate a higher similarity of the transcriptomes. The sample size (n) is indicated below each boxplot. The two-sided Wilcoxon rank-sum statistic was used to obtain the p-values indicated above the boxplots. All the boxplots show the distribution of all samples with dots, the median (center line), first and third quartile (upper and lower hinges), and the whiskers that extend to a maximum of 1.5 x interquartile range. **b**, Significantly similar transcriptomes are indicated by blue cells (light blue $p<0.05$ and dark blue $p<0.01$). Species are indicated by the mnemonic: PHYPA - *Physcomitrium patens*, MARPO - *Marchantia polymorpha*, SELML - *Selaginella moellendorffii*, GINBI - *Ginkgo biloba*, PICAB - *Picea abies*, AMBTC - *Amborella trichopoda*, ORYSA - *Oryza sativa*, MAIZE - *Zea mays*, ARATH - *Arabidopsis thaliana*, SOLLC - *Solanum lycopersicum*. The two-sided Wilcoxon rank-sum statistic was used to obtain the p-values. **c**, Heatmap showing the significant (p-value < 0.05) functional enrichment (orange cell) or depletion (blue cell) in the ten organ classes (y-axis) in at least 50% species. The heatmap indicates Mapman bins (photosynthesis-not assigned), transcription factors, and kinases. In all cases a one-sided empirical p-value was calculated using the 'Functional enrichment analysis' method (Supplementary Materials). The individual p-values are presented in Supplementary Table 4 and 5.

Fig. 3: Genomic analysis of organ-specificity of orthogroups. a, Species tree of the 23 species for which we have inferred orthogroups. The phylogenetic relationship was based on One Thousand Plant

Transcriptomes Initiative, 2019. Species in red are the ones with transcriptomic data available. Blue numbers in the nodes indicate the node number (e.g., 1: node 1). The tree's red circles show the percentage of orthogroups found at each node (largest: node 1 - 24% of all orthogroups, smallest: node 21 - 0.1%). **b**, Percentage of expression profile types of orthogroups per node. The expression profile types are: ubiquitous (light gray, orthogroup is not organ-specific), not conserved (light blue, organ-specificity not conserved in different species), or organ-specific (e.g., brown: root-specific). **c**, Percentage of phylostrata (nodes) within the different expression profile types. **d**, Transcriptome age index (TAI) of the different organ-specific genes in *Arabidopsis thaliana*. The boxplots show the TAI values (y-axis) in the different organs (x-axis), where a high TAI value indicates that the organ expresses a high number of younger genes. The sample size (n) is indicated above each boxplot. All the boxplots show the distribution of all samples with dots, the median (center line), first and third quartile (upper and lower hinges), and the whiskers that extend to a maximum of 1.5 x interquartile range. **e**, Summary of the average TAI value in the ten species. The organs are shown in rows, while the species are shown in columns. The TAI values were scaled to 1 for each species by dividing values in a column with the highest column value.

Fig. 4: Evolutionary analysis of organs. a, Enrichment and depletion of organ-specific genes per node in the species tree (nodes in the x-axis are the same as in Fig. 3a). The colors correspond with the number of species showing enrichment in each case (dark red: all species show enrichment, dark blue: all species show depletion). Horizontal bars below the node numbers show the main clades in different colors (Bry: Bryophytes, Gymn: Gymnosperms). **b**, Cladograms of the main lineages showing gain (in red) and loss (blue) of orthogroups with ubiquitous and organ-specific expression profiles.

Fig. 5: Comparison of male development across species. Overview of sexual reproduction in (**a**) Angiosperms, (**b**) *Physcomitrium*, and (**c**) *Marchantia*. **d**, Heatmaps showing the expression of male samples genes for *Arabidopsis thaliana*. Genes are in columns, sample names in rows. Gene expression is scaled to range between 0-1. Darker color corresponds to stronger gene expression. Bars to the bottom indicate the k-means clusters. **e**, Heatmap showing enrichment (orange) and depletion (blue) of functions in the found clusters. Light colors: $p<0.05$, dark colors: $p< 0.01$. In all cases a one-sided empirical p-value was calculated using the 'Functional enrichment analysis' method (Supplementary Materials). The individual p-values are presented in Supplementary Table 15. **f**, Heatmap showing the average normalized TPM value per cluster for all the species. **g**, Clustermap is showing the Jaccard distance between pairs of clusters of all the species.

Fig. 6: A network analysis of male clusters. a, Heatmaps show the number of observed connections divided by the number of expected connections. Darker colors indicate more connections between clusters. **b**, *A. thaliana* co-expression network clusters showing the edges between the different clusters (indicated as ARATH-0-5). The size of the panels indicate the number of genes in each cluster.

Transcription factors, kinases, and other genes are shown in red, yellow, and blue, respectively. **c**, Percentage of genes of each *A. thaliana* male cluster. The colors indicate the different stages of male development that a given gene is known to be involved in. For example, the majority of transcription factors in cluster ARATH-4 (highest expression in the pollen tube, Fig. 5f) are important for pollen tube growth (green bars).

Fig. 7: Features of the EVOREPRO database. a, Compare specificities tool. The dropdown menus allow selection of the species, orthogroups, organs, tissues, cell types, and SPM value cutoffs. The analysis is started by clicking on the 'Compare specificity' button. **b.** The Venn diagram shows the number of unique and common orthogroups of male-specific genes in *Arabidopsis* and *Amborella*. The default SPM value cutoff of 0.85 was used for both species. **c.** The table shows the identity of genes and orthogroups (first column) that are specifically expressed in male organs of *Arabidopsis* (second column) and *Amborella* (third column). Each row contains a gene family, and each cell can contain multiple comma-separated genes. Red triangles containing exclamation marks indicate genes with low expression (<10TPM). **d.** Expression profile of *GCS1* from *Arabidopsis*. The colored columns indicate the average expression values in the different samples, while gray points indicate the minimum and maximum expression values. The y-axis indicates the TPM value. **e.** Expression profile of *GCS1*-like gene from *Amborella* (*AMTR_s00069p00106210*). For clarity, the gray point indicating the maximum value in the sperm sample is omitted. **f.** Phylogenetic tree of the gene family OG_05_0008081 representing *GCS1*. The branches represent genes that are color-coded by species. The heatmap to the right of the gene identifiers indicates the scaled expression values in the major organ and cell types and ranges from low (yellow) to high (dark blue). Genes with TPM < 10 are indicated by filled red points, while the maximum gene expression is indicated by a blue bar to the right. **g.** Heatmap indicating the low (green) and high (red) expression of the *GCS1* gene family. **h.** Comparative analysis of co-expression clusters significantly (P<0.05) enriched for 'pollen tube' gene ontology term in *Arabidopsis* (cluster 13, left) and *Amborella* (cluster 39, right). Nodes indicate genes, while solid gray and dashed blue edges connect co-expressed and orthologous genes, respectively. We used 'label co-occurrences' as node options. For clarity, only part of each cluster is shown**.**

Tables

Table 1. Organs, tissues, and cell types used in the expression atlases analyzed. The different species are shown in columns, while the rows organize the organs, tissues and cell types into rows. N/A indicates that a given species does not have a corresponding organ/tissue, while '-' indicates that our dataset does not include data for the corresponding organ. Total number of experiments per sample are indicated in parentheses.

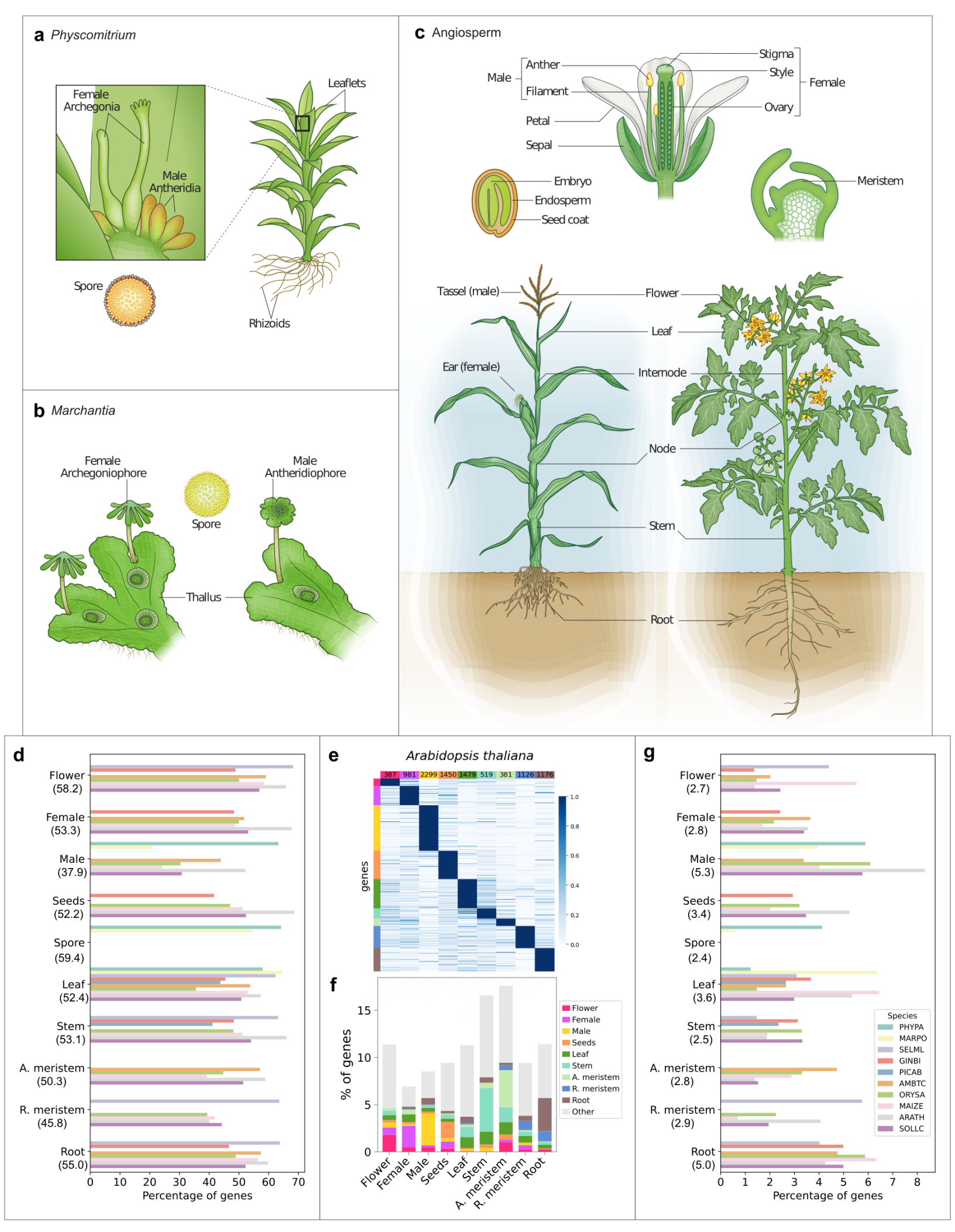
Organ/tissue /cell type	Marchantia	Physcomitrium	Selaginella	Gingko	Spruce	Amborella	Arabidopsis	Tomato	Rice	Maize

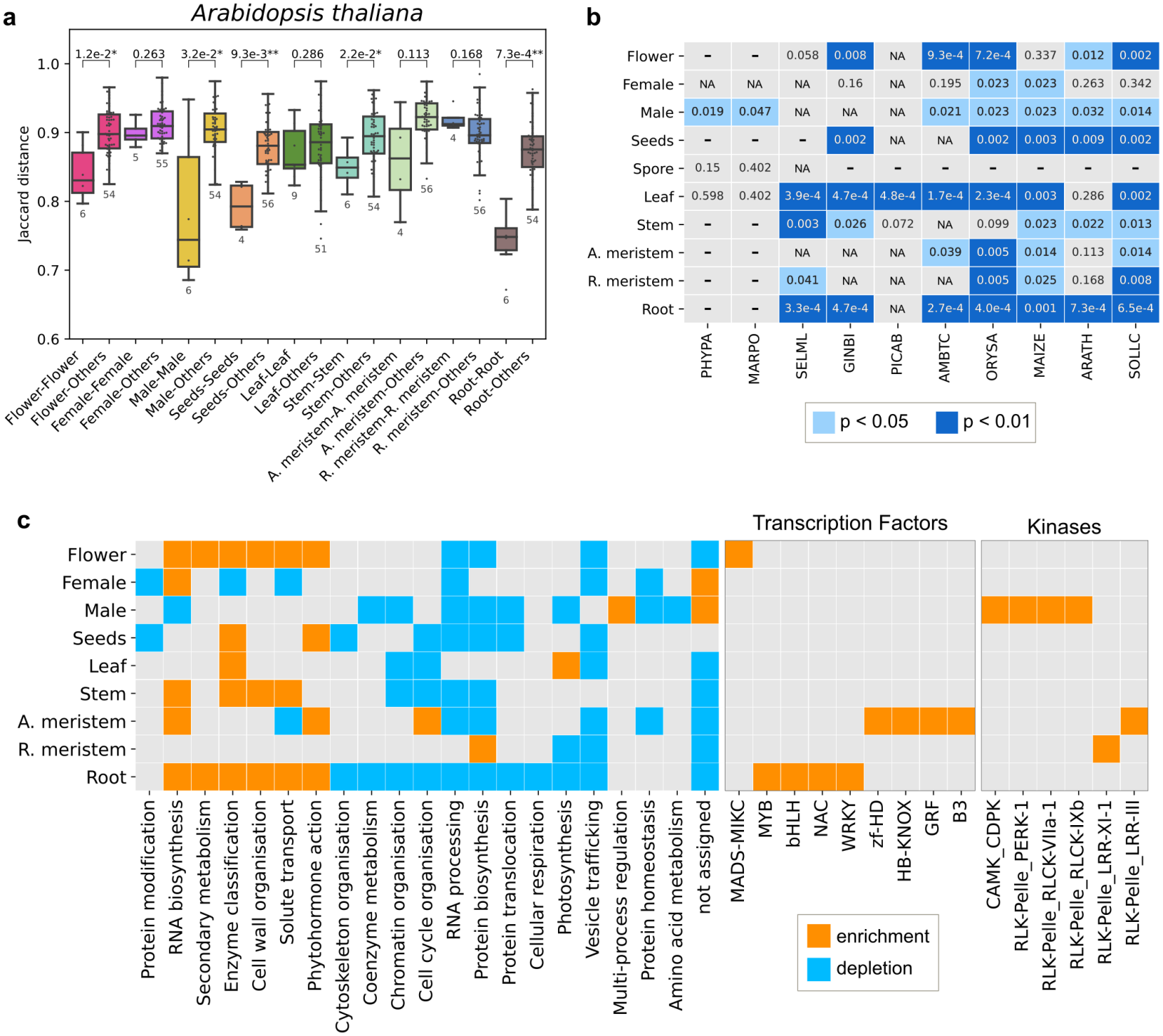
Flower	N/A	N/A	strobili (2)	microstrobilus (2), strobili (5)	N/A	flowers (6), buds (3), tepals (3)	carpels (14), stamen filaments (2), stigmatic tissue (2), petals (2), receptacles (8), sepals (4)	flowers (12), buds (7)	panicles (10), buds (2)	tassels (23), ear (22)
Female	-	-	-	ovules (9)	-	ovary (3), egg apparatus cell (3)	ovule (26), egg cell (10)	ovary (6), ovule (8), ovary wall (4)	ovary (14), ovule (40), egg cell (18)	nucellus (2), ovary (3), ovule (3), embryo sac (2)
Male	sperm (3)	Sperm (2)	-	-	-	pollen (mature, tube) (9), generative cell (2), microspores (3), sperm (3)	sperm (6), pollen (mature, tube, bicellular, tricellular) (26), microspore (6)	pollen (mature, tube) (44), microspore (3), generative cell (3), sperm cell (3)	pollen (tri-cellular, mature) (14), sperm (5)	pollen (mature, tube) (45), sperm cell (7), microspore (5)
Seeds	N/A	N/A	N/A	kernel (5)	-	-	endosperm (9), seed (young) (10), seed (4), seed (germinating) (6)	seeds (5-30 DPA) (94)	seeds (65), seed (1)	seed (11), kernel (11), endosperm (13), seeds (20), pericarp and aleurone (1)
Spore	sporeling (14)	germinating spores (3), spore capsule (12)	-	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Leaf	thallus (38)	leaflets (43)	microphyll (2)	leafs (81)	needles (63)	leaf (3)	leaf (14)	leaves (49)	leaves (644)	leaves (133)
Stem	N/A	N/A	top stem (2), bottom stem (2)	cambium (9), stem (3)	phloem (40), xylem (33), cambium (2)	-	stems (72)	stems (10)	stems (27)	stems (18)
Apical meristem	-	-	-	-	-	apical meristem (2)	apical meristem (30)	apical meristem (10)	apical meristem (16)	apical meristem (3)

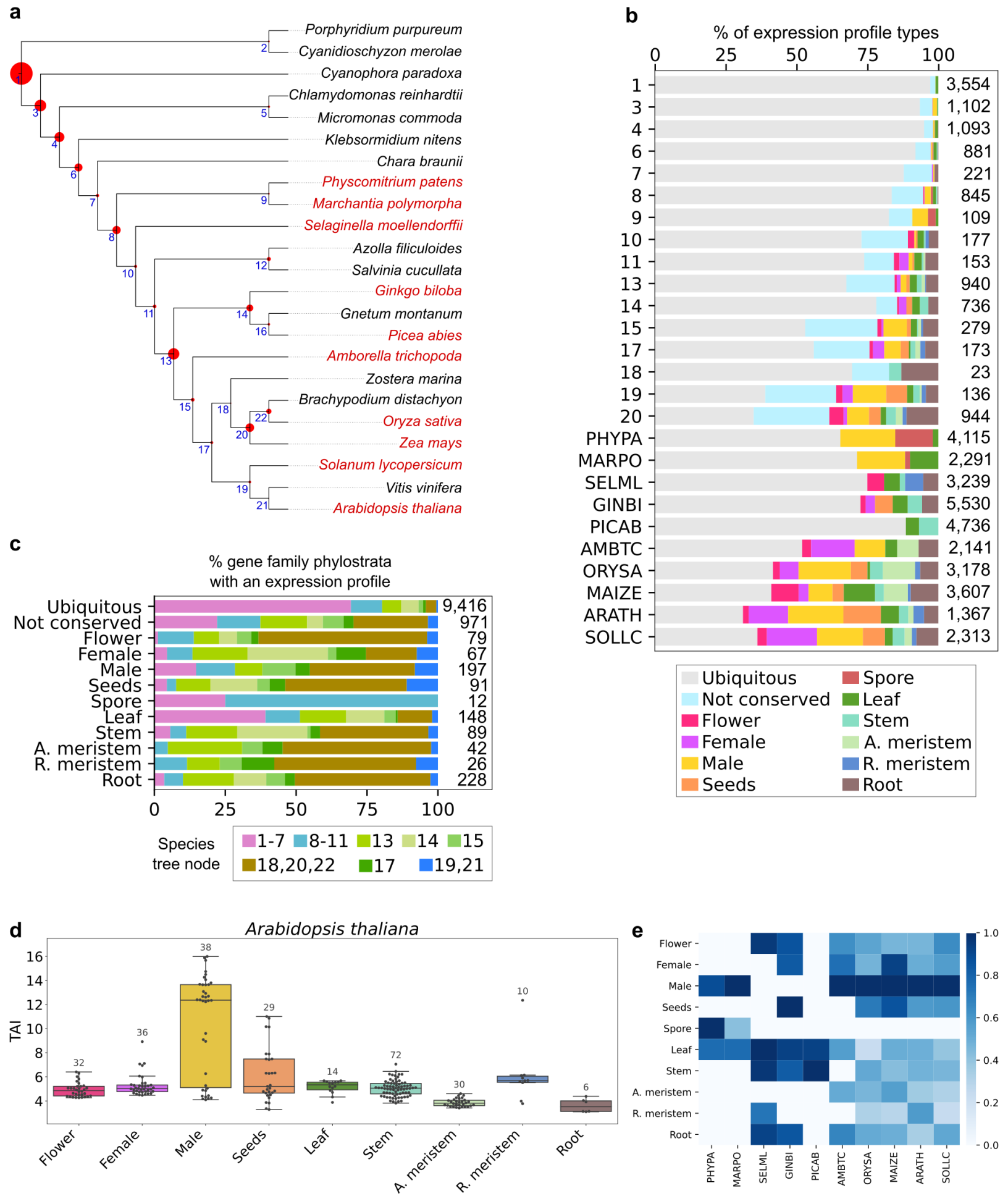
Root meristem	N/A	N/A	meristematic zone (3)	-	-	-	meristematic and QC zone (10)	meristematic zone (3)	meristematic zone (2)	meristematic zone (2)
Root	N/A	N/A	roots (5), rhizophores (2)	root (3)	-	roots (3)	apex (2), elongation zone (1), tip (3)	elongation zone (3), differentiation zone (3), root (4), root hair cells (2)	differentiation zone (2), roots (28), elongation zone (3)	roots (97), stele (4), elongation zone (4), maturation zone (1)

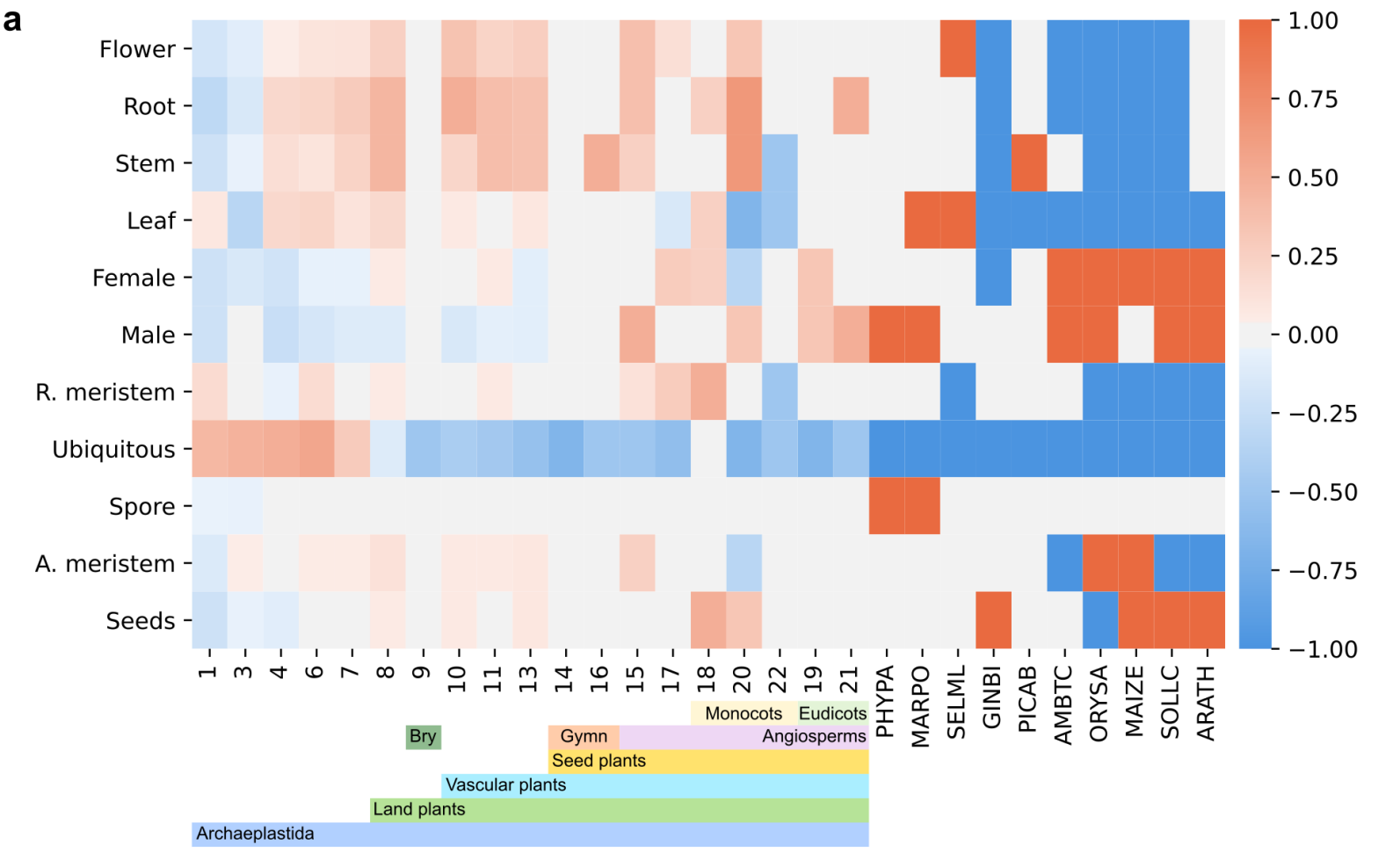
1056

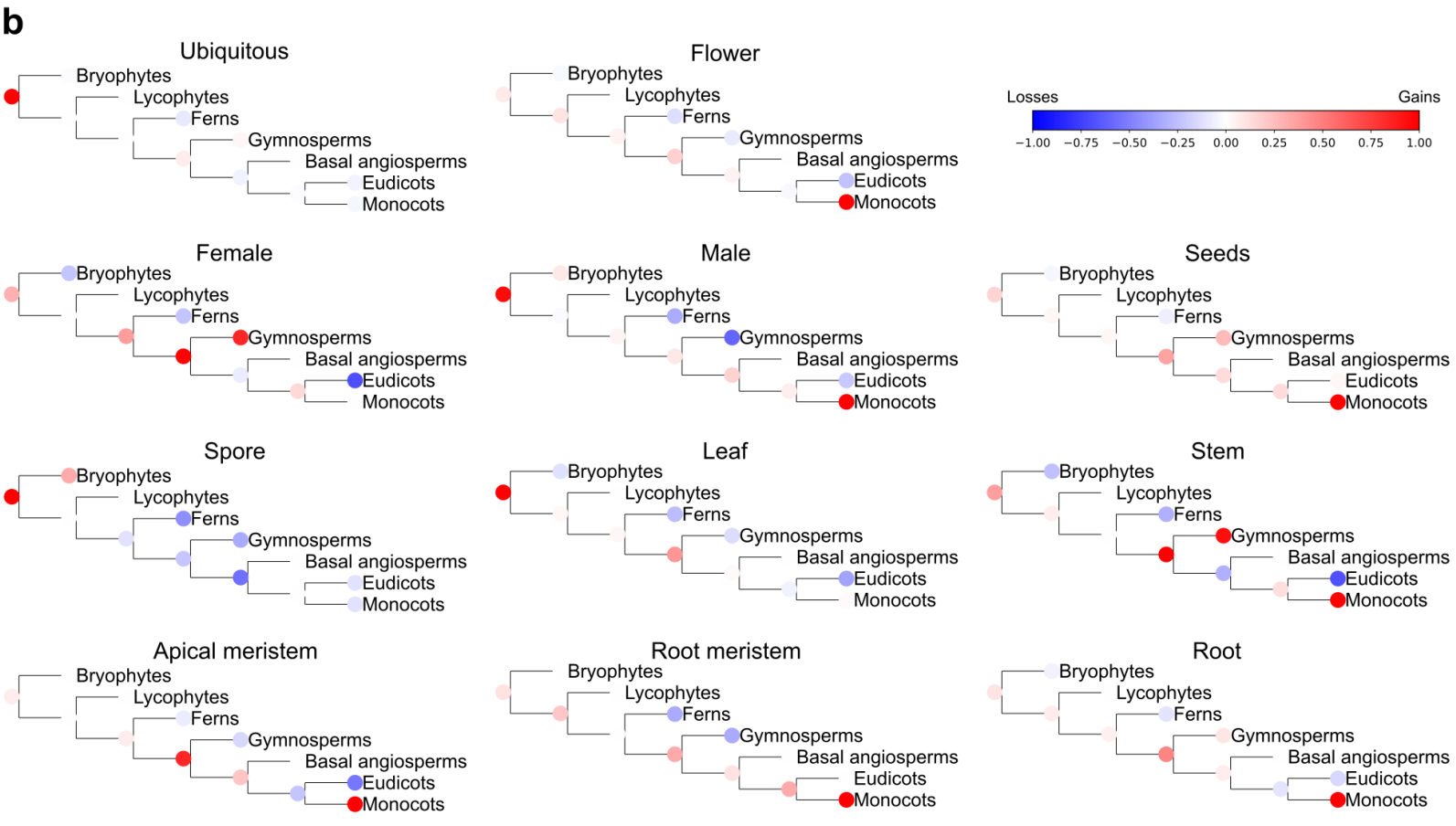
1057

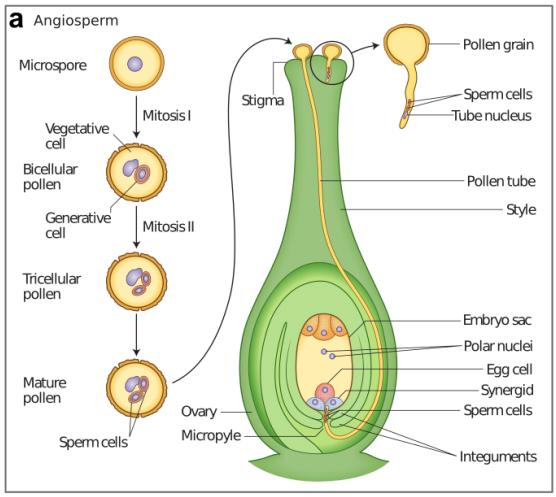

a Physcomitrium
Female Archegonia
Male Antheridia
Leaflets
Spore
Rhizoids
b Marchantia
Female Archegoniophore
Spore
Male Antheridiophore
Thallus
c Angiosperm
Male
Anther
Filament
Petal
Sepal
Stigma
Style
Ovary
Female
Embryo
Endosperm
Seed coat
Meristem
Tassel (male)
Ear (female)
Flower
Leaf
Internode
Node
Stem
Root
d
Flower (58.2)
Female (53.3)
Male (37.9)
Seeds (52.2)
Spore (59.4)
Leaf (52.4)
Stem (53.1)
A. meristem (50.3)
R. meristem (45.8)
Root (55.0)
Percentage of genes
e
Arabidopsis thaliana
387 981 2299 1450 1479 519 381 1126 1176
genes
f
% of genes
Flower
Female
Male
Seeds
Leaf
Stem
A. meristem
R. meristem
Root
Other
g
Flower (2.7)
Female (2.8)
Male (5.3)
Seeds (3.4)
Spore (2.4)
Leaf (3.6)
Stem (2.5)
A. meristem (2.8)
R. meristem (2.9)
Root (5.0)
Species
PHYPA
MARPO
SELML
GINBI
PICAB
AMBTC
ORYSA
MAIZE
ARATH
SOLLC
Percentage of genes

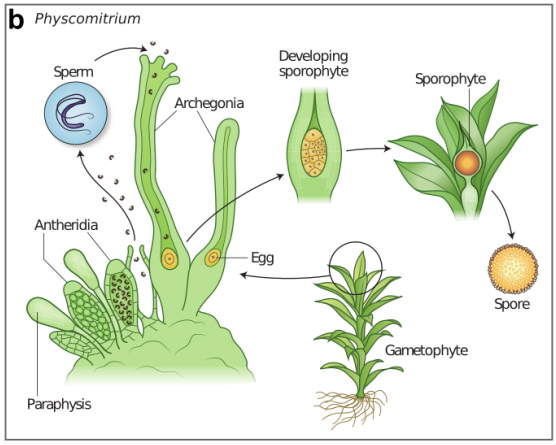

a
Arabidopsis thaliana
Jaccard distance
1.2e-2*
0.263
3.2e-2*
9.3e-3**
0.286
2.2e-2*
0.113
0.168
7.3e-4**
Flower-Flower
Flower-Others
Female-Female
Female-Others
Male-Male
Male-Others
Seeds-Seeds
Seeds-Others
Leaf-Leaf
Leaf-Others
Stem-Stem
Stem-Others
A. meristem-A. meristem
A. meristem-Others
R. meristem-R. meristem
R. meristem-Others
Root-Root
Root-Others
b
Flower - - - 0.058 0.008 NA 9.3e-4 7.2e-4 0.337 0.012 0.002
Female - NA NA NA 0.16 NA 0.195 0.023 0.023 0.263 0.342
Male - 0.019 0.047 NA NA NA 0.021 0.023 0.023 0.032 0.014
Seeds - - - - 0.002 NA NA 0.002 0.003 0.009 0.002
Spore - 0.15 0.402 NA - - - - - - -
Leaf - 0.598 0.402 3.9e-4 4.7e-4 4.8e-4 1.7e-4 2.3e-4 0.003 0.286 0.002
Stem - - - 0.003 0.026 0.072 NA 0.099 0.023 0.022 0.013
A. meristem - - - NA NA NA 0.039 0.005 0.014 0.113 0.014
R. meristem - - 0.041 NA NA NA 0.005 0.025 0.168 0.008
Root - - 3.3e-4 4.7e-4 NA 2.7e-4 4.0e-4 0.001 7.3e-4 6.5e-4
PHYPA
MARPO
SELML
GINBI
PICAB
AMBTC
ORYSA
MAIZE
ARATH
SOLLC
p < 0.05
p < 0.01
c
Transcription Factors
Kinases
Flower
Female
Male
Seeds
Leaf
Stem
A. meristem
R. meristem
Root
Protein modification
RNA biosynthesis
Secondary metabolism
Enzyme classification
Cell wall organisation
Solute transport
Phytohormone action
Cytoskeleton organisation
Coenzyme metabolism
Chromatin organisation
Cell cycle organisation
RNA processing
Protein biosynthesis
Protein translocation
Cellular respiration
Photosynthesis
Vesicle trafficking
Multi-process regulation
Protein homeostasis
Amino acid metabolism
not assigned
MADS-MIKC
MYB
bHLH
NAC
WRKY
zf-HD
HB-KNOX
GRF
B3
CAMK_CDPK
RLK-Pelle_PERK-1
RLK-Pelle_RLCK-VIIa-1
RLK-Pelle_RLCK-IXb
RLK-Pelle_LRR-XI-1
RLK-Pelle_LRR-III
enrichment
depletion

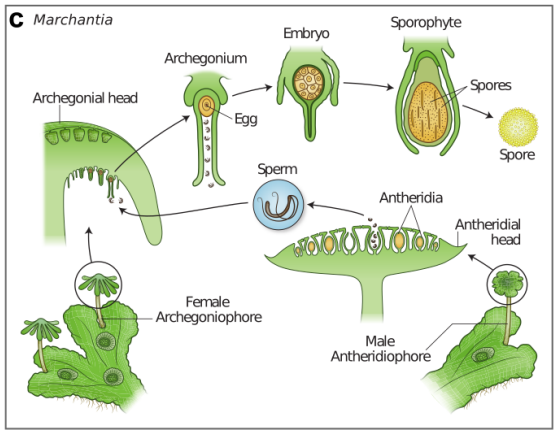

a
Porphyridium purpureum
Cyanidioschyzon merolae
Cyanophora paradoxa
Chlamydomonas reinhardtii
Micromonas commoda
Klebsormidium nitens
Chara braunii
Physcomitrium patens
Marchantia polymorpha
Selaginella moellendorffii
Azolla filiculoides
Salvinia cucullata
Ginkgo biloba
Gnetum montanum
Picea abies
Amborella trichopoda
Zostera marina
Brachypodium distachyon
Oryza sativa
Zea mays
Solanum lycopersicum
Vitis vinifera
Arabidopsis thaliana
b
% of expression profile types
0
25
50
75
100
1 3,554
3 1,102
4 1,093
6 881
7 221
8 845
9 109
10 177
11 153
13 940
14 736
15 279
17 173
18 23
19 136
20 944
PHYPA 4,115
MARPO 2,291
SELML 3,239
GINBI 5,530
PICAB 4,736
AMBTC 2,141
ORYSA 3,178
MAIZE 3,607
ARATH 1,367
SOLLC 2,313
Ubiquitous
Not conserved
Flower
Female
Male
Seeds
Spore
Leaf
Stem
A. meristem
R. meristem
Root
c
% gene family phylostrata
with an expression profile
Ubiquitous 9,416
Not conserved 971
Flower 79
Female 67
Male 197
Seeds 91
Spore 12
Leaf 148
Stem 89
A. meristem 42
R. meristem 26
Root 228
0
25
50
75
100
Species tree node
1-7
8-11
13
14
15
18,20,22
17
19,21
d
Arabidopsis thaliana
TAI
16
14
12
10
8
6
4
32
36
38
29
14
72
30
10
6
Flower
Female
Male
Seeds
Leaf
Stem
A. meristem
R. meristem
Root
e
Flower
Female
Male
Seeds
Spore
Leaf
Stem
A. meristem
R. meristem
Root
PHYPA
MARPO
SELML
GINBI
PICAB
AMBTC
ORYSA
MAIZE
ARATH
SOLLC
1.0
0.8
0.6
0.4
0.2
0.0

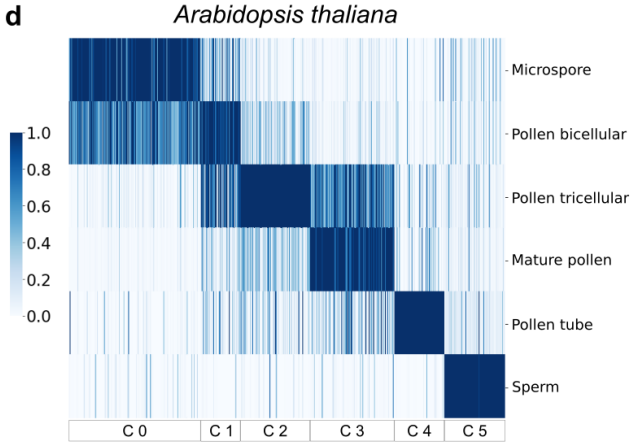

a
Flower
Root
Stem
Leaf
Female
Male
R. meristem
Ubiquitous
Spore
A. meristem
Seeds
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
1 3 4 6 7 8 9 10 11 13 14 16 15 17 18 20 22 19 21 PHYPA MARPO SELML GINBI PICAB AMBTC ORYSA MAIZE SOLLC ARATH
Monocots
Eudicots
Bry
Gymn
Angiosperms
Seed plants
Vascular plants
Land plants
Archaeplastida


b
Ubiquitous
Flower
Losses
Gains
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00
Female
Male
Seeds
Spore
Leaf
Stem
Apical meristem
Root meristem
Root
Bryophytes
Lycophytes
Ferns
Gymnosperms
Basal angiosperms
Eudicots
Monocots

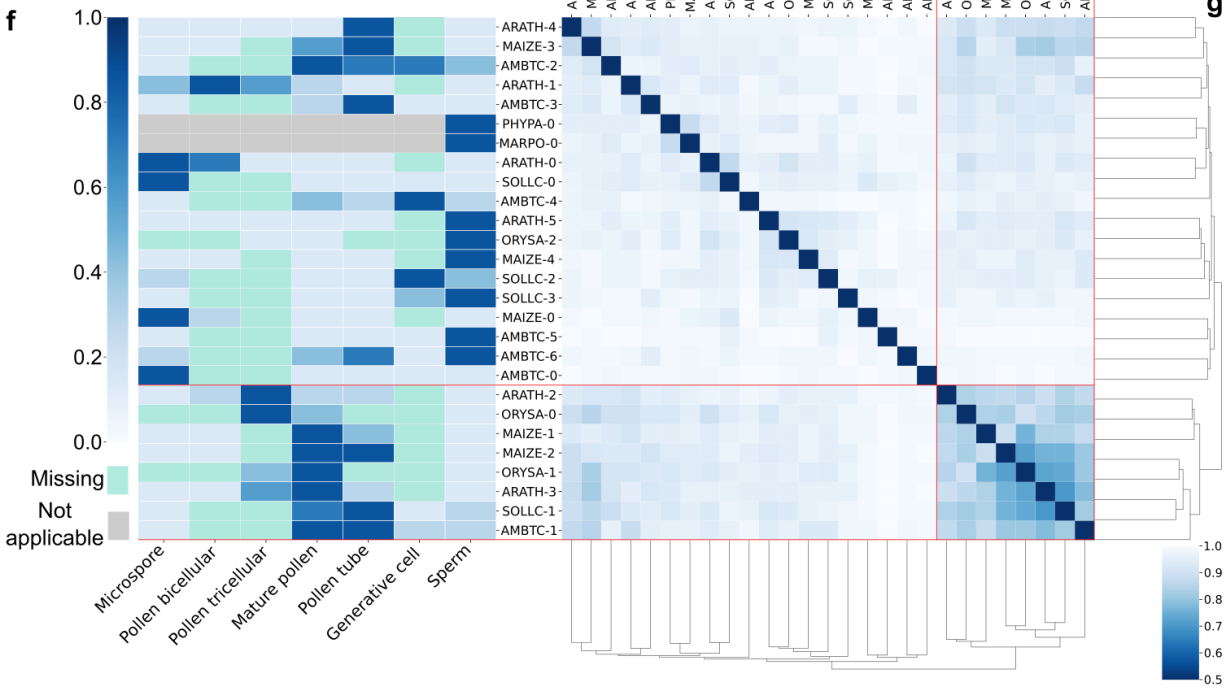

a Angiosperm
Microspore
Mitosis I
Vegetative cell
Bicellular pollen
Generative cell
Mitosis II
Tricellular pollen
Mature pollen
Sperm cells
Stigma
Pollen grain
Sperm cells
Tube nucleus
Pollen tube
Style
Embryo sac
Polar nuclei
Egg cell
Synergid
Sperm cells
Ovary
Micropyle
Integuments

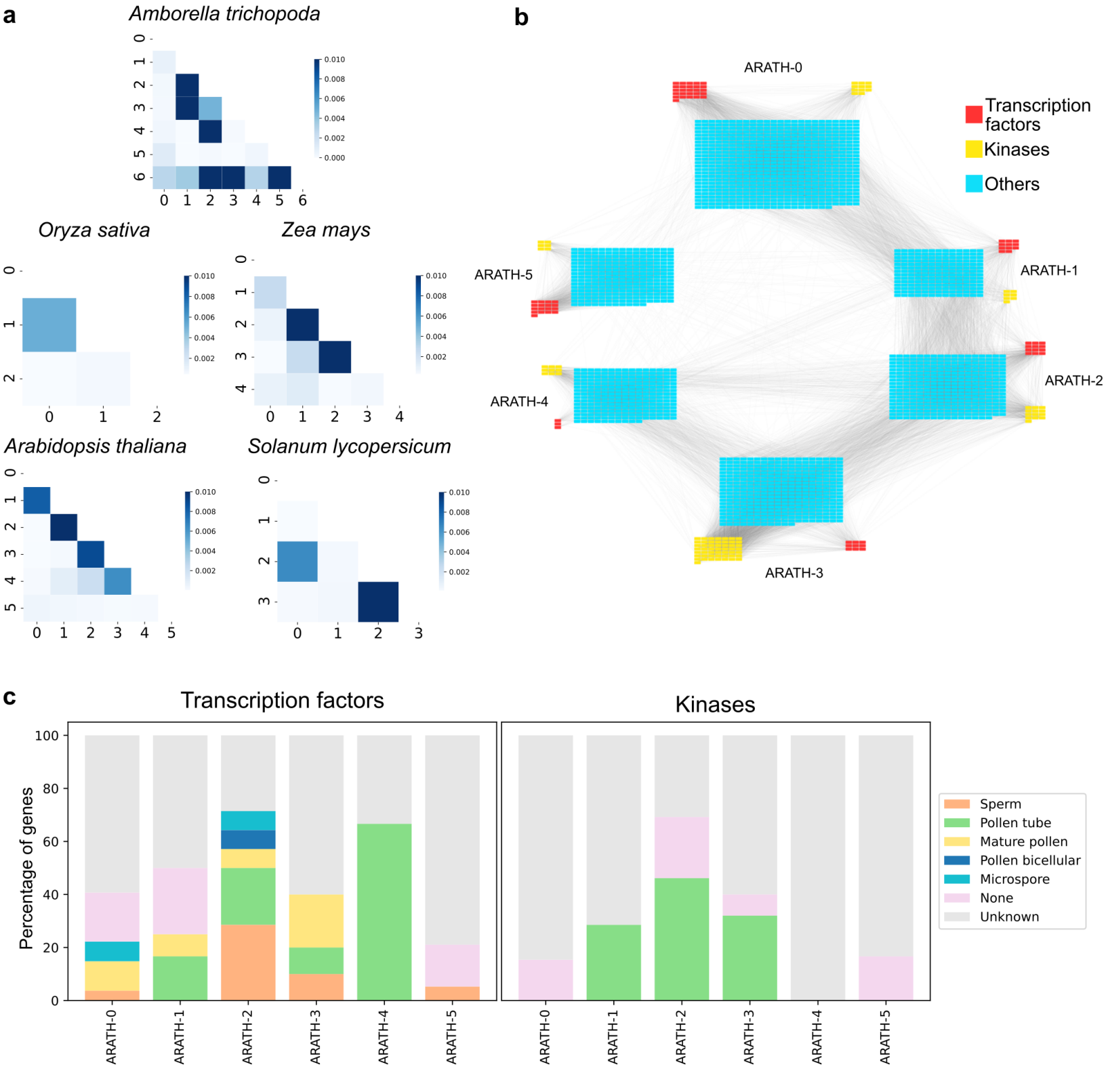

b Physcomitrium
Sperm
Archegonia
Developing sporophyte
Sporophyte
Antheridia
Egg
Spore
Gametophyte
Paraphysis


c Marchantia
Archegonium
Embryo
Sporophyte
Archegonial head
Egg
Spores
Spore
Sperm
Antheridia
Antheridial head
Female Archegoniophore
Male Antheridiophore

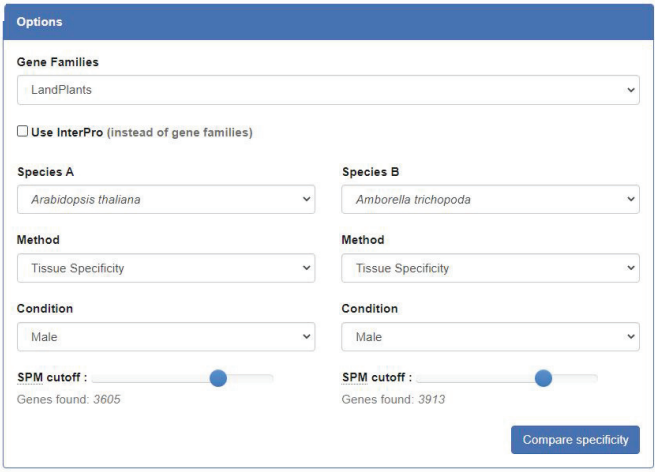

d Arabidopsis thaliana
1.0
0.8
0.6
0.4
0.2
0.0
Microspore
Pollen bicellular
Pollen tricellular
Mature pollen
Pollen tube
Sperm
C 0
C 1
C 2
C 3
C 4
C 5


e
not assigned
Vesicle trafficking
Solute transport
Secondary metabolism
RNA processing
RNA biosynthesis
Protein translocation
Protein modification
Protein homeostasis
Protein biosynthesis
Phytohormone action
Photosynthesis
Nutrient uptake
Multi-process regulation
Lipid metabolism
External stimuli response
Enzyme classification
DNA damage response
Cytoskeleton organisation
Coenzyme metabolism
Chromatin organisation
Cellular respiration
Cell wall organisation
Cell cycle organisation
Carbohydrate metabolism
Amino acid metabolism
ARATH-4
MAIZE-3
AMBTC-2
ARATH-1
AMBTC-3
PHYPA-0
MARPO-0
ARATH-0
SOLLC-0
AMBTC-4
ARATH-5
ORYSA-2
MAIZE-4
SOLLC-2
SOLLC-3
MAIZE-0
AMBTC-5
AMBTC-6
AMBTC-0
ARATH-2
ORYSA-0
MAIZE-1
MAIZE-2
ORYSA-1
ARATH-3
SOLLC-1
AMBTC-1


f
1.0
0.8
0.6
0.4
0.2
0.0
Missing
Not applicable
Microspore
Pollen bicellular
Pollen tricellular
Mature pollen
Pollen tube
Generative cell
Sperm
g
1.0
0.9
0.8
0.7
0.6
0.5


a
Amborella trichopoda
Oryza sativa
Zea mays
Arabidopsis thaliana
Solanum lycopersicum
b
ARATH-0
ARATH-1
ARATH-2
ARATH-3
ARATH-4
ARATH-5
Transcription factors
Kinases
Others
c
Transcription factors
Kinases
Percentage of genes
Sperm
Pollen tube
Mature pollen
Pollen bicellular
Microspore
None
Unknown

A

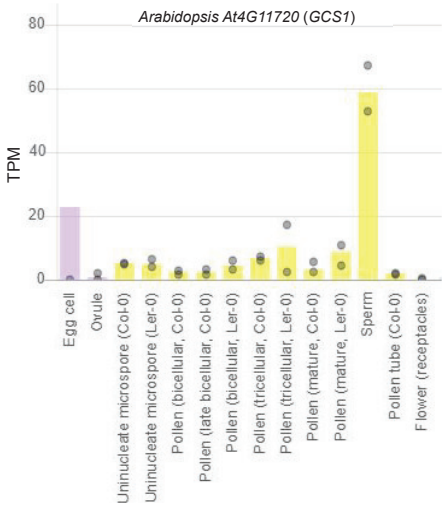


B

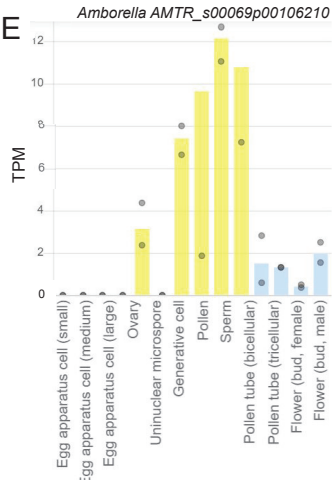
C

OG_05_0008081	AT4G11720 (GCS1)	AMTR_s00069p00106210 (evm_27.TU.AmTr_v1.0_scaffold00069.68)	intersection
OG_05_0002277	AT5G27980	AMTR_s00004p00042590 ⚠ (evm_27.TU.AmTr_v1.0_scaffold00004.22)	intersection
OG_05_0000548	AT5G11412 ⚠ AT5G53720 ⚠	AMTR_s00038p00212100 (evm_27.TU.AmTr_v1.0_scaffold00038.171)	intersection
OG_05_0007637	AT5G49150 (GEX2)	AMTR_s00019p00129570 (evm_27.TU.AmTr_v1.0_scaffold00019.109)	intersection
OG_05_0012083	AT3G17060	AMTR_s00123p00050980 (evm_27.TU.AmTr_v1.0_scaffold00123.3)	intersection
OG_05_0008080	AT1G04470, AT2G33420	AMTR_s00067p00188350 (evm_27.TU.AmTr_v1.0_scaffold00067.201)	intersection
OG_05_0003207	AT1G01640	AMTR_s00077p00025220 (evm_27.TU.AmTr_v1.0_scaffold00077.11)	intersection
OG_05_0001012	AT1G06030 ⚠ AT4G10260	AMTR_s00001p00265040 (evm_27.TU.AmTr_v1.0_scaffold00001.402)	intersection

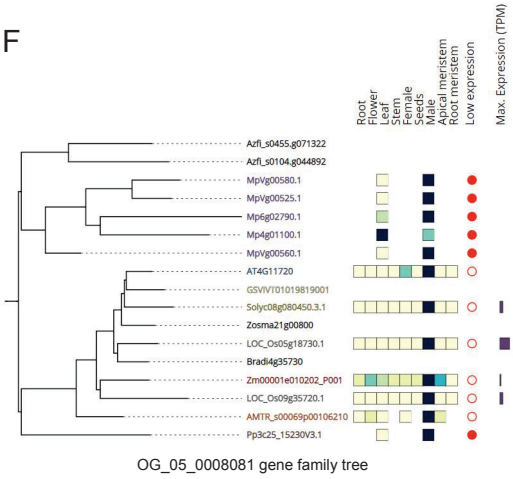
D



E

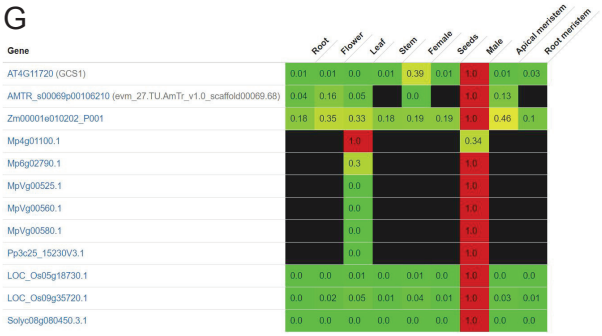


F

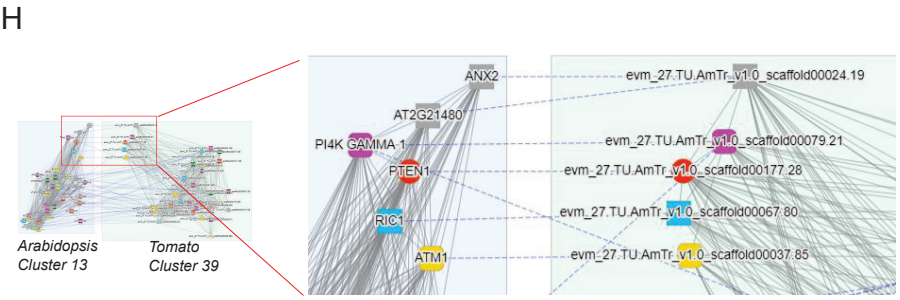


OG_05_0008081 gene family tree

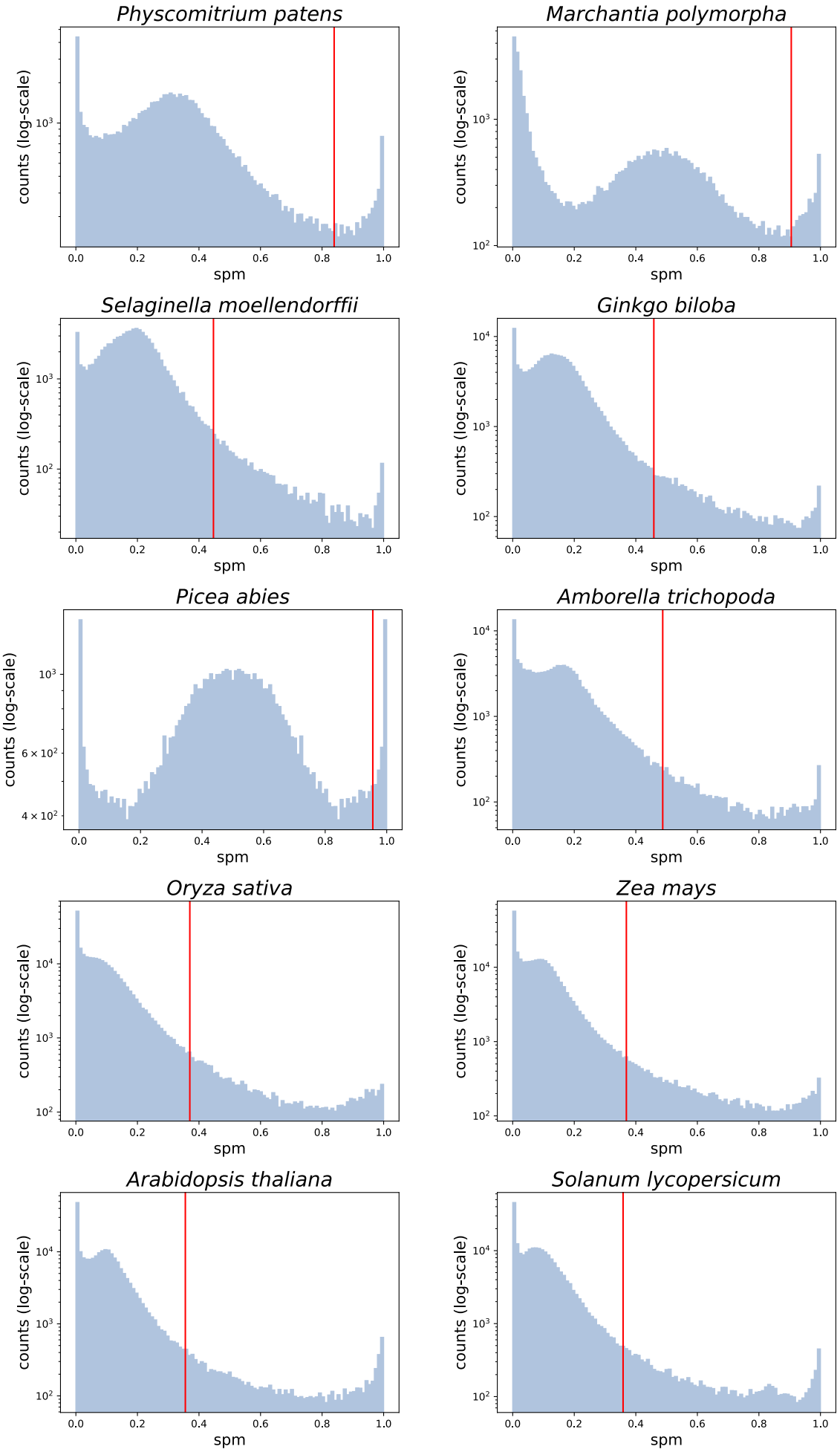
G



H

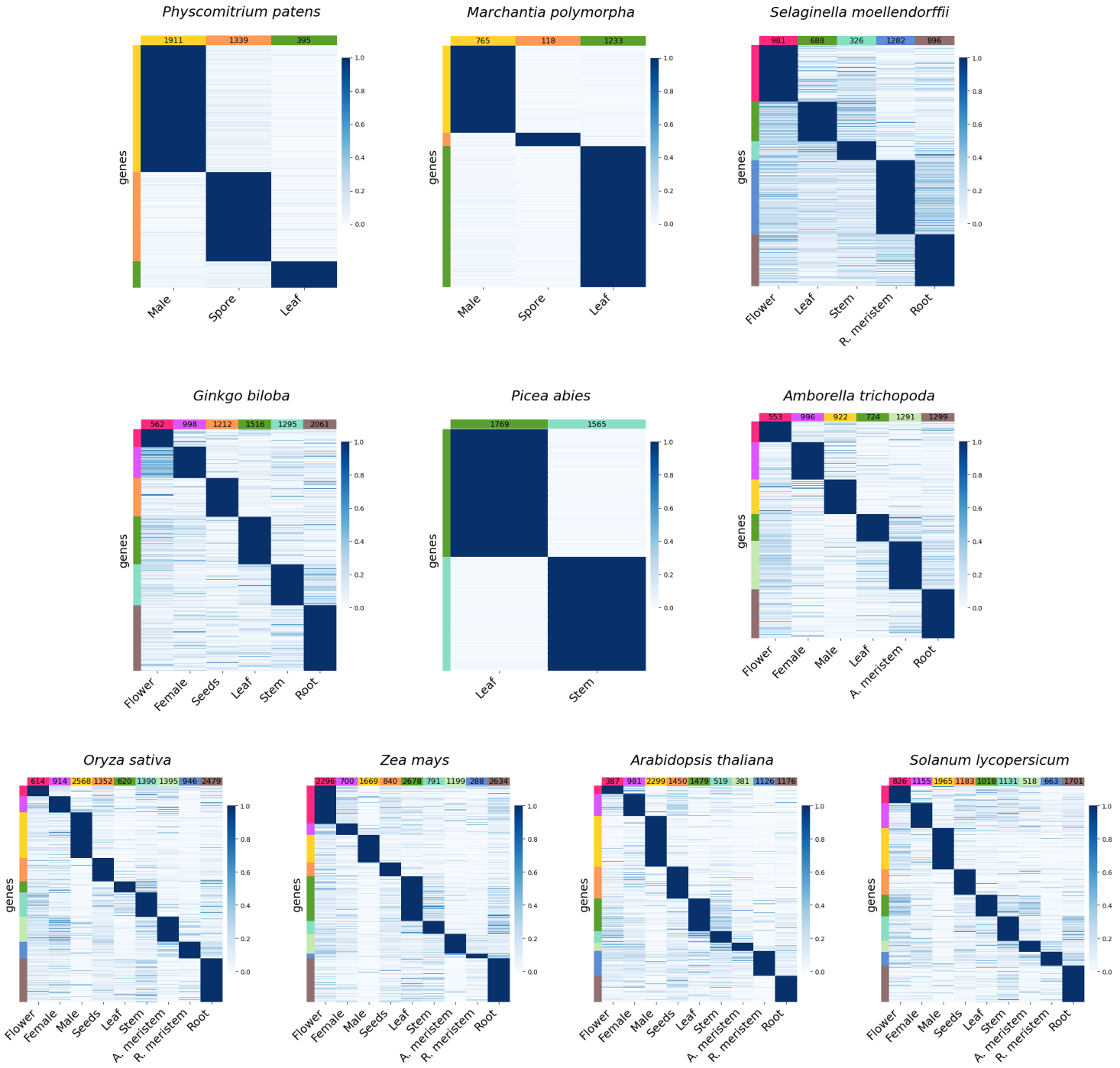


Supplementary Fig. 1



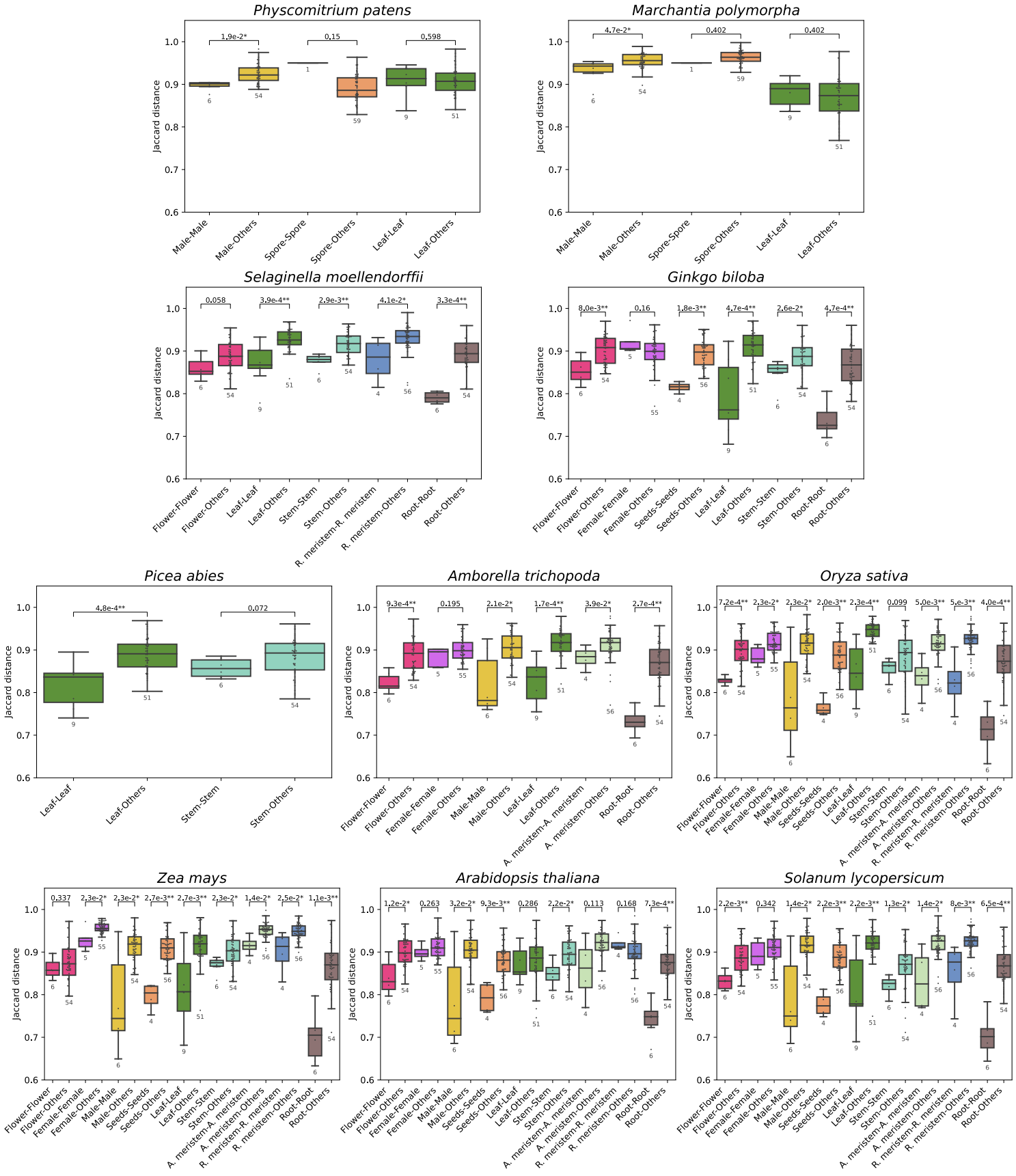
Supplementary Fig. 1: Distribution of SPM values in the ten species. The x-axis indicates the specificity measure (SPM), while the y-axis indicates the log10-transformed frequency of the SPM values observed for all genes across the samples. The vertical red line indicates the SPM value cutoff, below which 95% of values are found.

Supplementary Fig. 2



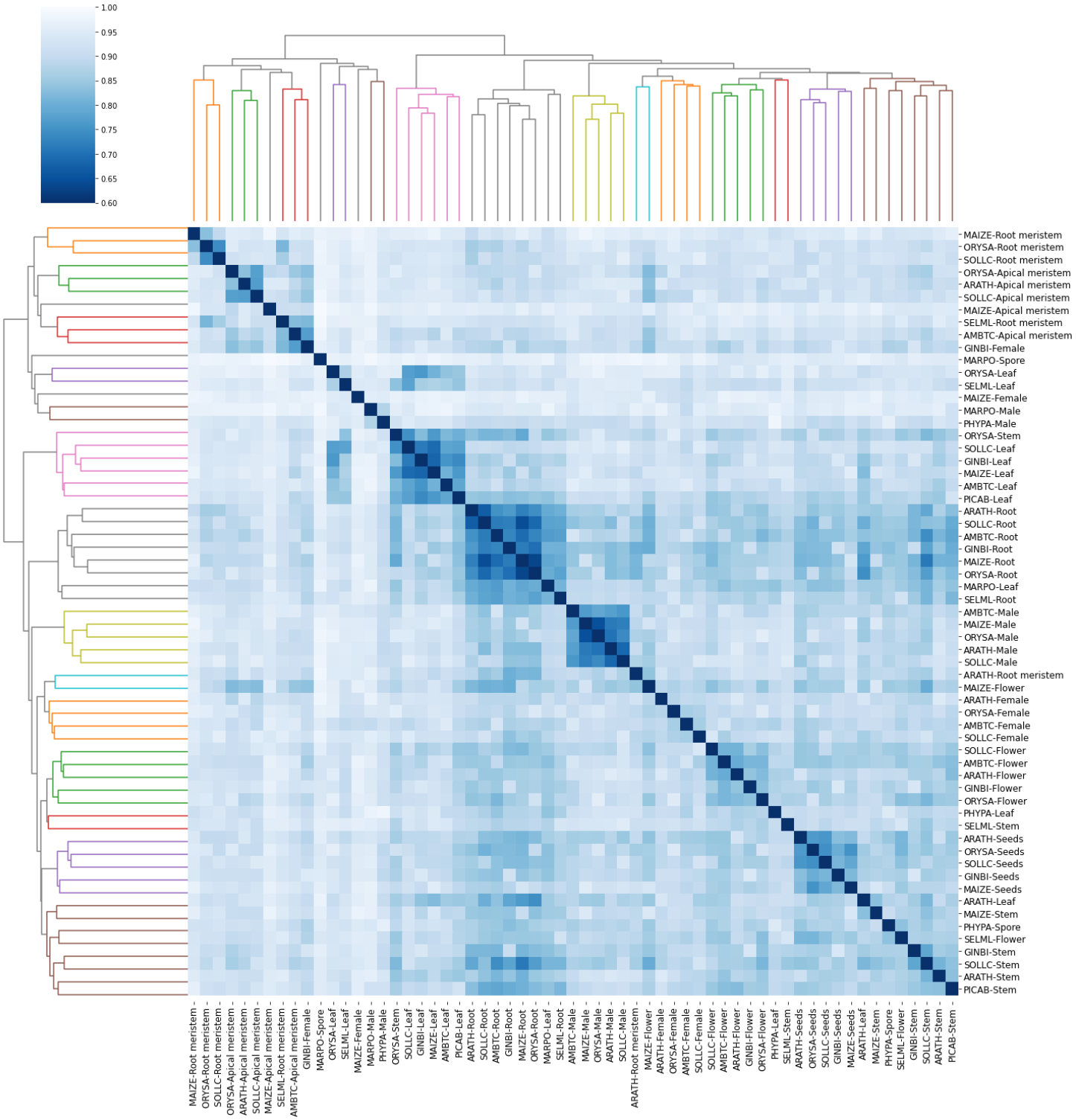
Supplementary Fig. 2: Expression profiles of the genes that were deemed to be specifically expressed in one of the organs/tissues/cells (sample) of the ten species used in this study. Genes are in rows, samples in columns, and the genes are sorted according to the expression profiles (e.g., flower, female). The numbers at the top of each column indicate the total number of specific genes in each sample. Gene expression is scaled to range from 0-1. Bars on the left of each heatmap show the organ-specific genes and correspond to the samples on the bottom: pink - Flower, purple - Female, yellow - Male, orange - Seeds/Spore, dark-green - Leaf, medium-green - Stem, light-green - Apical meristem, blue - Root meristem, brown - Root.

Supplementary Fig. 3



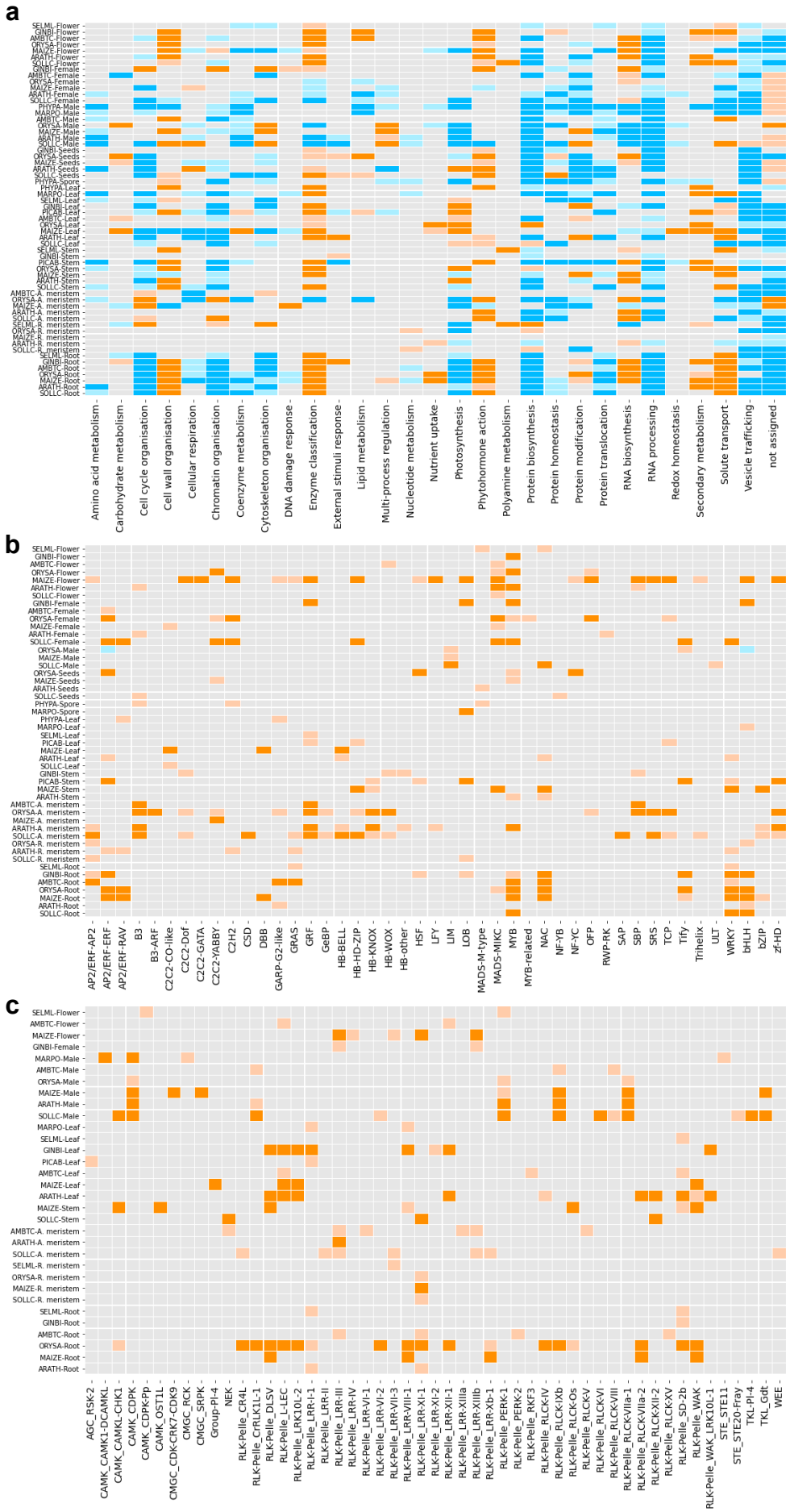
Supplementary Fig. 3: Bar plot showing the Jaccard distances when comparing the same organs (i.e., male-male) and one organ versus the others (i.e., male-others) for the ten species included in this study. The sample size (n) is indicated below each boxplot. The two-sided Wilcoxon rank-sum statistic was used to obtain the p-values indicated above the boxplots. All the boxplots show the distribution of all samples with dots, the median (center line), first and third quartile (upper and lower hinges), and the whiskers that extend to a maximum of 1.5 interquartile range.

Supplementary Fig. 4



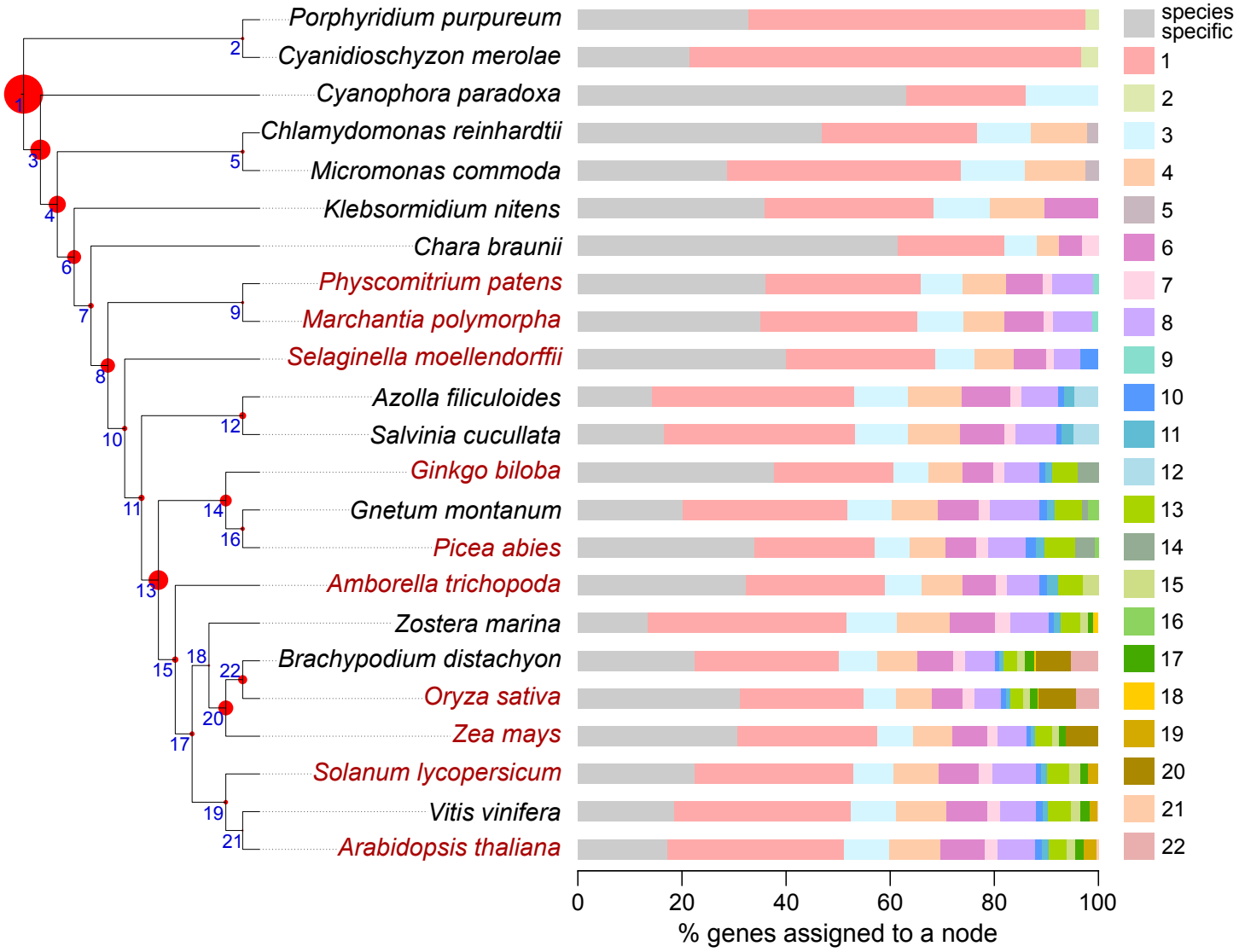
Supplementary Fig. 4: Comparing transcriptome similarities of the samples of the ten species. We used the Jaccard Index to calculate the similarity of transcriptomes of all samples in the dataset. The heatmap shows which transcriptomes of organs across species are similar by hierarchical clustering (dark blue). A lower value indicates a stronger similarity between two organs (white). For example, when comparing Arabidopsis root to roots from other species, we observe more similar transcriptomes than Arabidopsis root to non-root samples. The dendrograms on top and the left show the different clusters formed when the distance is <1.3.

Supplementary Fig. 5



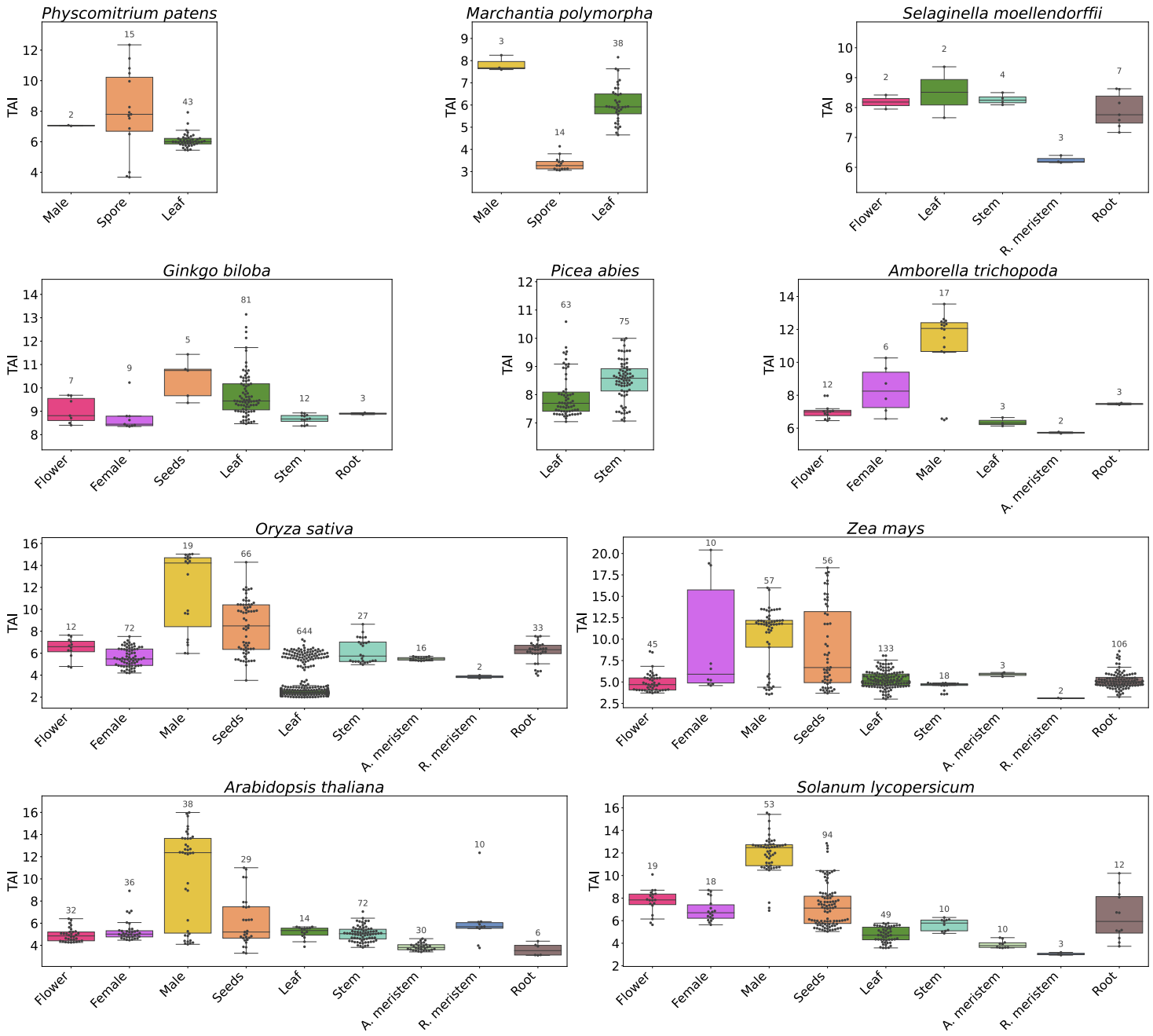
Supplementary Fig. 5. Functional enrichment analysis of the organ-specific transcriptomes. Organs are shown on the y-axis and functions in the x-axis for MapMan bins (a), transcription factors (b), and kinases (c). Orange and blue colors indicate enrichment and depletion, respectively. The intensity of the color is in correlation with the p-value (dark orange/blue: p <0.01, light orange/blue: p<0.05). In all cases a one-sided empirical p-value was calculated using the 'Functional enrichment analysis' method (Supplementary Materials). The individual p-values are presented in Supplementary Table 5.

Supplementary Fig. 6



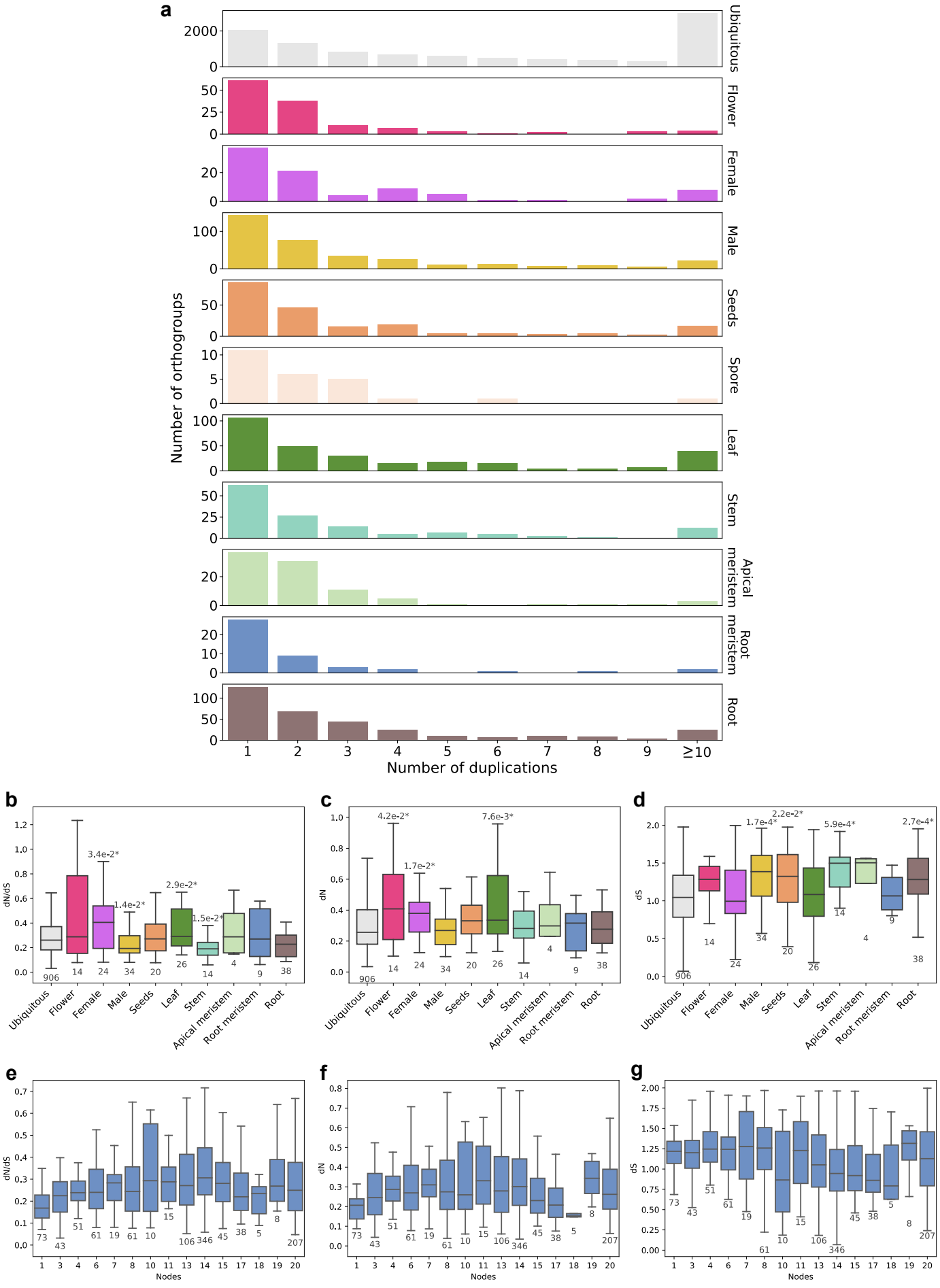
Supplementary Fig. 6: Cladogram of the 23 species included in the analysis. The phylogenetic relationship was based on One Thousand Plant Transcriptomes Initiative, 2019. Species in red are associated with transcriptomic data in this study. Blue numbers in the nodes indicate the node number (e.g., 1: NODE_1). The tree's red circles show the percentage of orthogroups found in each node (largest and smallest amounts: Node_1 - 24% and NODE_21 - 0.1%). Bars on the right show the percentage of genes per species that are present in each node. The nodes are shown in different colors, as indicated in the right bar.

Supplementary Fig. 7



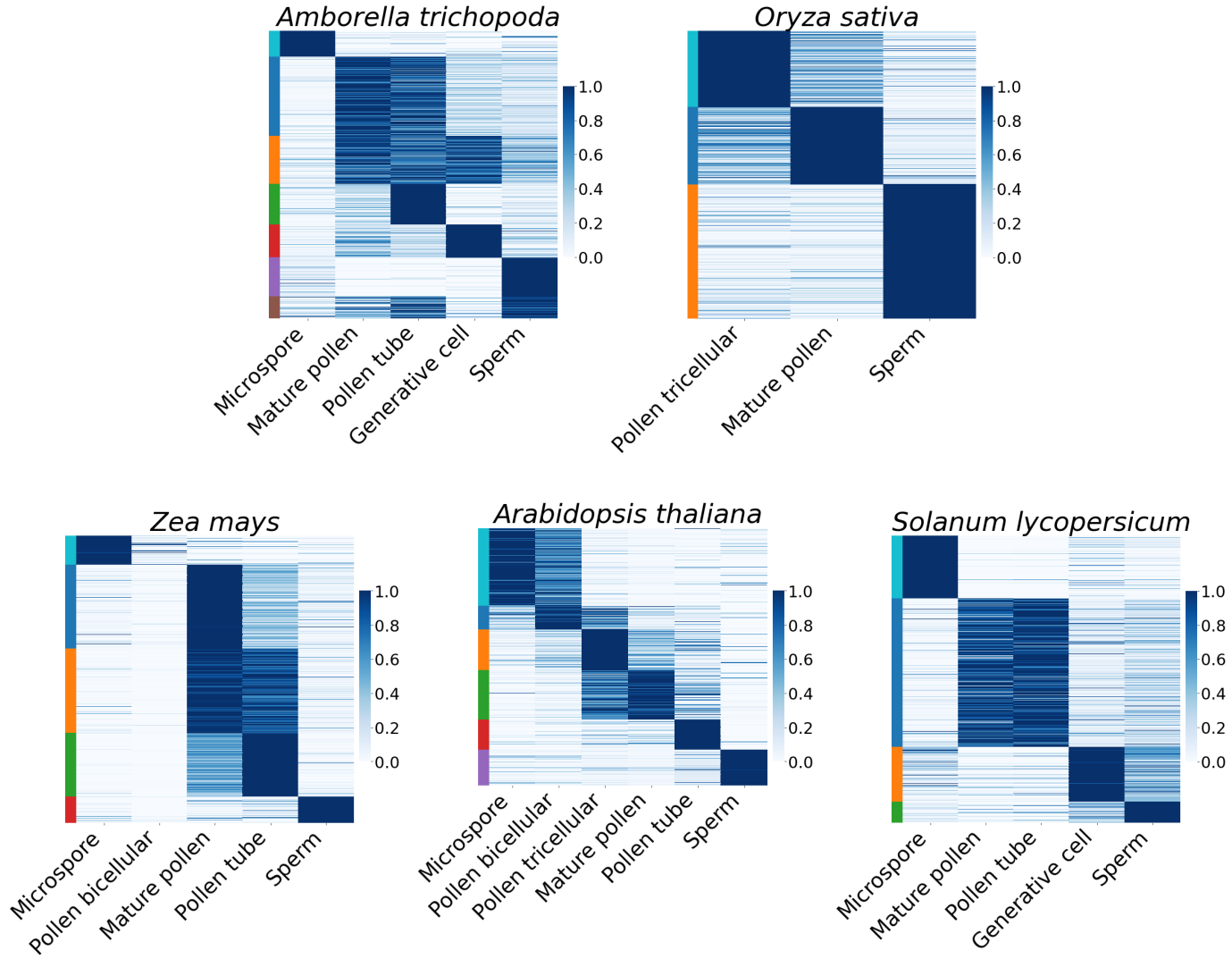
Supplementary Fig. 7: Transcriptomic age index in the ten species. The sample size (n) is indicated above each boxplot. All the boxplots show the distribution of all samples with dots, the median (center line), first and third quartile (upper and lower hinges), and the whiskers that extend to a maximum of 1.5 interquartile range.

Supplementary Fig. 8



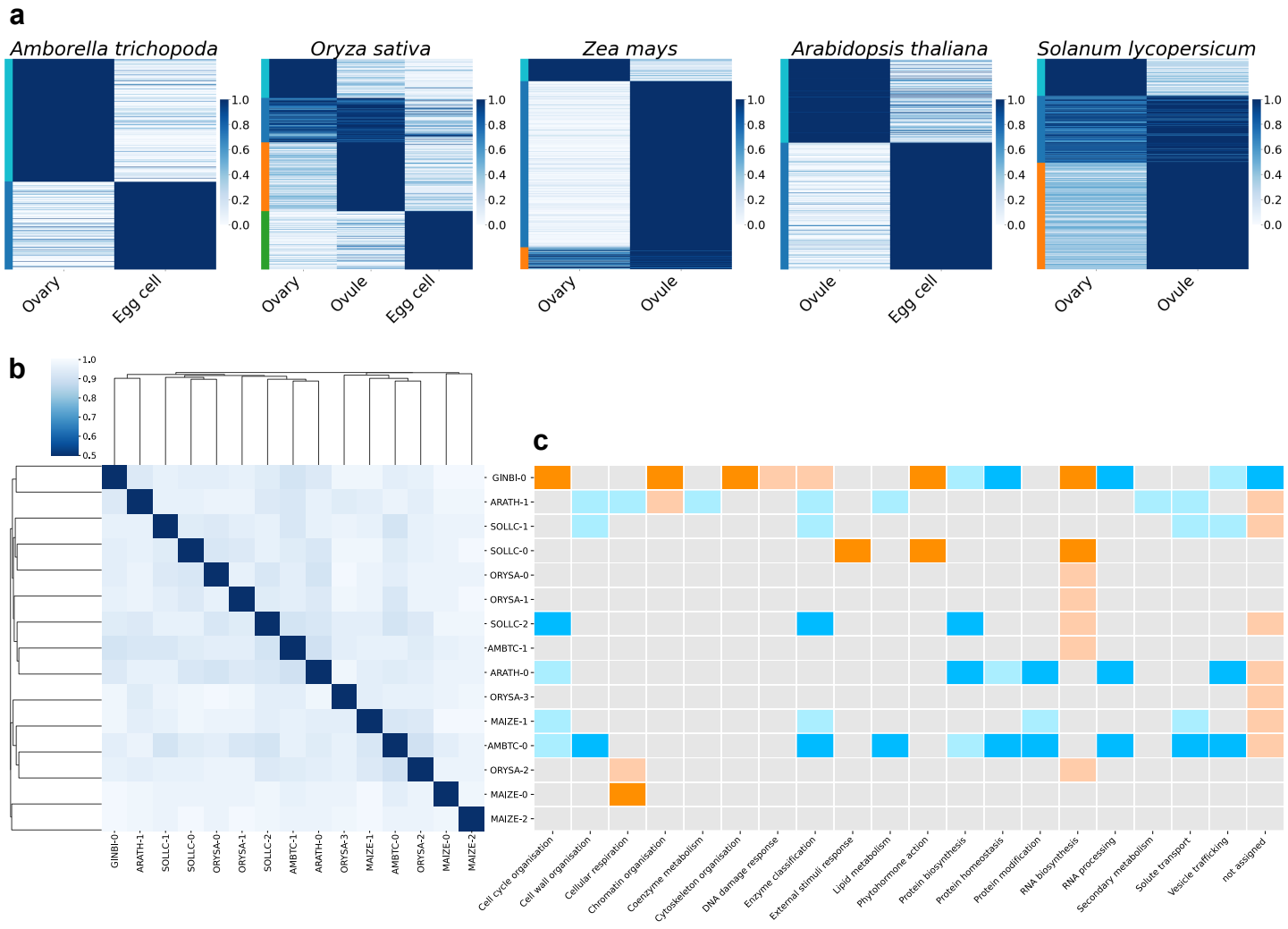
Supplementary Fig. 8: Number of duplications and evolutionary rates of orthogroups with organ-specific and ubiquitous expression profile. a, Number of orthogroups that have 1 to >=10 duplications. Ratio of nonsynonymous to synonymous substitution rates (dN/dS) (b), dN (c), dS (d) of the ubiquitous and organ-specific orthogroups. The numbers above the bars show the p-values of the comparison between the organ-specific and ubiquitously expressed genes using the two-sided Wilcoxon rank-sum test (p<0.05). Different colors indicate the different expression profiles. dN/dS (e), dN (f), dS (g) of the orthogroups in the different nodes of the species tree (Supplementary Fig. 6). All boxplots (b-g) show the sample size (n) below each boxplot, the median (center line), first and third quartile (upper and lower hinges), and the whiskers that extend to a maximum of 1.5 interquartile range.

Supplementary Fig. 9



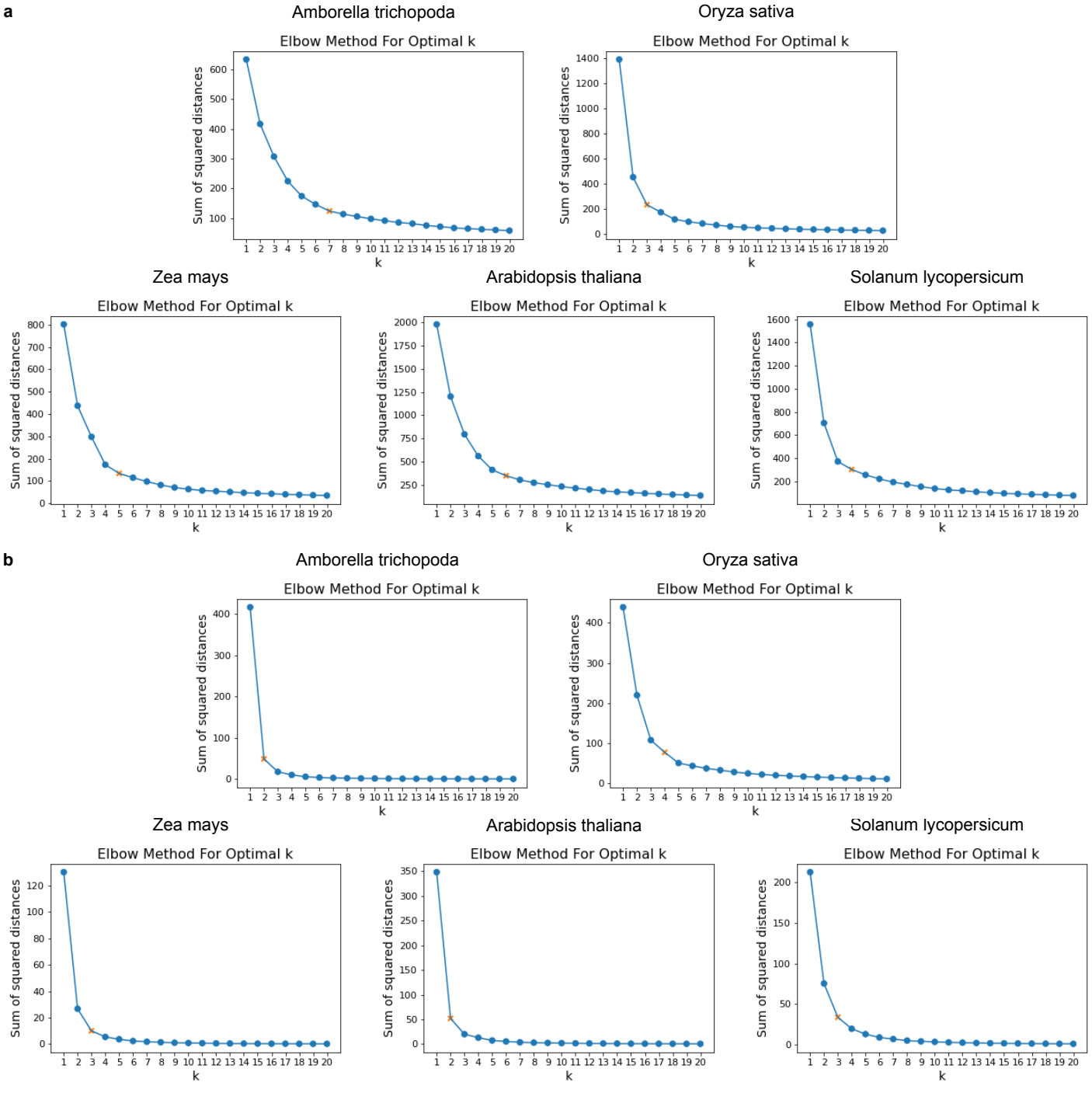
Supplementary Fig. 9: Expression of male developmental stages genes for five species. Genes are in rows, developmental stages in columns. Gene expression is scaled to range from 0-1. Darker color corresponds to a stronger positive correlation. Bars in the left mark the different clusters.

Supplementary Fig. 10



Supplementary Fig. 10: Analysis of the expression profile in different development stages of female organs. Heat map showing the normalized TMP of genes per each development stage for five species. Bars on the left indicate the clusters. b, Jaccard distance between the clusters. c, Heatmap showing enrichment and depletion of functions. Orange and blue indicate enrichment and depletion, respectively (light colors: $p < 0.05$, dark colors: $p < 0.01$). In all cases a one-sided empirical p-value was calculated using the 'Functional enrichment analysis' method (Supplementary Materials). The individual p-values are presented in Supplementary Table 17.

Supplementary Fig. 11



Supplementary Fig. 11: Identifying k value with the elbow method. The orange mark indicates the k value where the sum of squared distances was less than 80% of the highest value found at k =1. a, For the male samples, and b, for the female samples.

Supplementary Methods

Physcomitrium growth conditions, RNA isolation and sequencing

Plant growth

The Gransden wild-type strain from *P. patens* Bruch & Schimp[1] was used for this study. To initiate plant growth and culture, 3 mature sporophytes were sterilized using a 5% commercial bleach solution for 5 minutes and rinsed twice in molecular grade water. Sterilized sporophytes were then broken using a pipette tip and diluted into 5mL molecular grade water. Spore containing solution was then distributed into 4 sterile peat pellets (Jiffy-7, Jiffy Products International) and two 9 cm Petri dishes containing KNOPS medium (Reski and Abel, 1985) supplemented with 0.5 g/l ammonium tartrate dibasic (Sigma-Aldrich Co). Petri dishes were kept at 25°C, 50% humidity, and 16 h light (light intensity 80 µlum/m/s). Protonema samples were collected 10 days after spore germination.

Plants in Phytatray™ II (Sigma-Aldrich Co) containing 4 sterile peat pellets (Jiffy-7, Jiffy Products International) were grown for 6-8 weeks at 25°C, 50% humidity, and 16 h light (light intensity 80 µlum/m/s). Water was supplied to the bottom of each box. Leave samples were collected after 6 weeks, prior to induction of gametangia development. For gametangia and sporophyte development, water was again supplied to the bottom of each box containing four pellets and were transferred to 17°C, 8 h light, and 50% humidity (light intensity 50 µlum/m/s) to induce the development of reproductive structures[2]. Gametangia samples (archegonia, paraphysis and sperm cell packages) were collected 15 days after reproductive induction. Antheridia samples were collected at several time points during their development. Further development of the sporophyte was conducted under these conditions and sporophyte samples were collected at different time points during sporophyte development. S1 sporophytes were collected 7 days after sperm cell (SC) release, S2 sporophytes 15 days after SC release, S3 sporophytes 20 days after SC release (green spore capsules) and SM samples 28 days after SC release (brown spore capsules).

Sample preparation and sequencing

Leaves, protonema and sporophytes were collected under a stereoscope using tweezers, placed in 2.5 uL of RLT+ buffer (Qiagen), and shock frozen in liquid nitrogen. Before RNA-seq library preparation, these samples were mechanically disrupted using sterile pellet pestles (Z359947, Sigma-Aldrich Co). Antheridia, archegonia, paraphysis and sperm cell packages were collected using a Yokogawa CSU-W Spinning Disk confocal with 10x 0.25NA objective, using the brightfield channel and an Andor Zyla 4.2 sCMOS camera. For each of these samples the plants were prepared under a stereoscope, isolating the whole gametangia for ca. 10 shoots. They were placed in 20 uL of molecular grade water on a glass slide. Using a cover slip the gametangia were disrupted into individual antheridia by applying slight pressure. Slides were placed under a microscope and specific organs were identified and collected, using an Eppendorf CellTram® Air/Oil/vario micromanipulator with glass capillaries (borosilicate glass with fire polished ends, without filament GB100-9P) pulled with a Narishige PC-10 puller. Then they were transferred to another clean slide, and subsequently excessive liquid containing possible contaminations, such as cell debris, was removed. For paraphysis samples 8-15 individual paraphysis were collected directly into 2 uL of RLT+ buffer and flash frozen in liquid nitrogen. For antheridia samples 5 to 20 individual antheridia of each specific stage (9 to 15 days after induction, distinguished by size) were collected and then burst under a microscope by applying pressure on a cover slip applied to the samples on the slide. The slide was washed with 4 uL of RLT+ buffer and the buffer transferred into a PCR tube, subsequently flash frozen. Archegonia samples were prepared from 3-5 archegonia following the same procedure. Released sperm cell packages (2-5 per sample) were collected from gametangia preparations (as described above; antheridia 15 days after induction) without clean up, transferred into a tube with 2 uL of RLT+ buffer, flash frozen in liquid nitrogen and subsequently used for RNA-seq library preparation.

RNA-seq library preparation for all samples was performed as described in[3], with the addition of mixing the PCR tubes on a Thermomixer C (Eppendorf) every 15 minutes at 200 rpm for 1 min during the RT step. Libraries were sequenced on a NextSeq500 instrument with single-end 75 bp read length (SE75).

Marchantia growth conditions, RNA isolation and sequencing

Male accession of *Marchantia polymorpha* L., Takaragaike (Tak)-1 was grown on vermiculite under a long-day condition (16/8 h day/night) at 22 °C. To induce sexual reproduction, thalli developed from gemmae were transferred to a far-red light (700 – 780 nm, 44.3 µmol photons m^{-2} s^{-1}) supplemented light condition using LabLEDs (RHENAC GreenTec Ag). Sperms were released from antheridiophores by applying ddH_2O supplemented with RNasin® Ribonuclease Inhibitor (1 u/µl, Promega), collected in a 1.5 mL tube, and pelleted by centrifugation at 3,000 *g* for 5 min at 4 °C. RNA-seq libraries were generated from total RNA of isolated *M. polymorpha* sperm using Smart-seq2[4] using independent biological replicates. The libraries were sequenced on an Illumina Hiseq 2500 using 125 bp paired-end.

Amborella growth conditions, RNA isolation and sequencing

Plant material and isolation procedures

Amborella trichopoda male flowers were harvested from a male plant growing in the Botanical Garden in Bonn (Germany), in a shaded place inside a greenhouse under controlled conditions of 16-18ºC, constant humidity of 66% and 12-hour photoperiods. Buds and fully opened male flowers were gathered in 50 ml Falcon™ conical tubes (Thermo Fisher), placed without lid in a hermetically sealed plastic box containing a bed of silica gel.

Uninucleated microspores (UNM) were isolated at room temperature from flower buds of 4.5 mm length, as these were found to contain 98% uninucleated microspores. In brief, three samples with each 5 g buds were homogenized in 0.1 M mannitol and filtered with a 70-micron pore size PET strainer (PluriSelect). The filtered solution was processed by subsequent steps of percoll gradient separation, washing and centrifugation, as described previously[5].

Amborella generative cells (GC) were obtained from mature pollen grains that were purified like described previously[6]. Per replicate, 50 mg pollen was resuspended in 1 ml pollen germination medium and transferred into a 1.5 ml vial containing glass beads (0.4 – 0.6 mm). The vial was vortexed

continuously at 2,200 rpm for 4 minutes to crack the pollen grains and release its contents. The solution was filtered using a 15-micron PET strainer (PluriSelect). To stain the nuclei, a final concentration of 10X SYBR Green I was added and GCs were identified using an inverted microscope (Nikon) equipped with high-resolution 20X and 40X objectives suitable for fluorescent applications and suitable filters for SYBR Green I (497 nm excitation; 520 nm emission). For RNA-seq, three replicates of each 140 GC were harvested manually using an Eppendorf CellTram.

Amborella sperm cells (SC) were isolated at room temperature by adapting a method described for tomato sperm cell isolation[7]. In brief, three replicates with each 50 mg purified pollen were germinated as described[6]. 16 hours after germination, the medium was removed by filtration using a 15-micron PET strainer (PluriSelect) and the pollen tubes were incubated for 10 min in a 15% mannitol solution with 0.4% cellulase "Onozuka" R-10 and 0.2% macerozyme R-10 to release the sperm cells. The mixture was re-filtered using a 15-micron PET strainer and loaded on 5 ml 23% Percoll in 0.55 M mannitol and centrifuged at 1,000 x g for 30 min. Approximately 1 ml with SC, floating on the surface of the Percoll gradient, were harvested, washed with 1 ml RNAprotect® Cell Reagent (Qiagen) and centrifuged for 10 min at 2,500 x g. 50 µl of SC-enriched pellet (approximately 250 sperm cells each replicate) was used for RNA-seq library preparation.

Isolation and sampling of *Amborella* ovaries, egg apparatus cells, pollen tubes, pollen grains as well as male and female flowers, tepals, roots and leaves was done as described in previous studies[6,8].

RNA isolation and sequencing

RNA isolation from uninucleated microspores was performed by using the Spectrum™ Plant Total RNA Kit (Sigma-Aldrich) according to manufacturer's instructions. Total RNA from *Amborella* generative cells and sperm cells was extracted according to the "Purification of total RNA from animal and human cells" protocol of the RNeasy Plus Micro Kit (QIAGEN, Hilden, Germany). In brief, cells were stored and shipped on dry ice. After adding RLT Plus containing ß-mercaptoethanol the samples were homogenized by vortexing for 30 sec. Genomic DNA contamination was removed using gDNA Eliminator spin columns. Next ethanol was added and the samples were applied to RNeasy MinElute

spin columns followed by several wash steps. Finally total RNA was eluted in 12 µl of nuclease free water. Purity and integrity of the RNA was assessed on the Agilent 2100 Bioanalyzer with the RNA 6000 Pico LabChip reagent set (Agilent, Palo Alto, CA, USA).

The SMARTer Ultra Low Input RNA Kit for Sequencing v4 (Takara) was used to generate first strand cDNA from 2.5 ng UNM, 0.8 ng GC and 0.5 ng SC total RNA. Double stranded cDNA was amplified by LD PCR (10 for UNM, 13 cycles for GC and 15 cycles for SC) and purified via magnetic bead clean-up. Library preparation was carried out as described in the Illumina Nextera XT Sample Preparation Guide (Illumina, Inc., San Diego, CA, USA). 150 pg of input cDNA were tagmented by the Nextera XT transposome. The products were purified and amplified via a limited-cycle PCR program to generate multiplexed sequencing libraries. For the PCR step 1:5 dilutions of index 1 (i7) and index 2 (i5) primers were used. The libraries were quantified using the KAPA SYBR FAST ABI Prism Library Quantification Kit. Equimolar amounts of each library were used for cluster generation on the cBot (TruSeq SR Cluster Kit v3). The sequencing run was performed on a HiSeq 1000 instrument using the indexed, 2x100 cycles paired end (PE) protocol and the TruSeq SBS v3 Kit. Image analysis and base calling resulted in .bcl files, which were converted into .fastq files by the CASAVA1.8.2 software. Library preparation and RNA-seq were performed at the service facility "Center of Excellence for Fluorescent Bioanalytics (KFB)" (Regensburg, Germany; www.kfb-regensburg.de).

Arabidopsis growth conditions, RNA isolation and sequencing

Arabidopsis thaliana accession Columbia-0 (Col-0) plants were grown in controlled-environment cabinets at 22°C under illumination of 150 µmol/m2/sec with a 16-h photoperiod. Mature pollen grains (MPG) were harvested from open flowers of 5 to 6-week old plants by shaking into liquid medium (0.1 M D-mannitol) as described previously[79]. Microspores and developing pollen grains were released from anthers of closed flower buds and purified by Percoll density gradient centrifugation as described[5,9]. Populations of spores at five stages of development were isolated: uninucleate microspores (UNM), bicellular pollen (BCP), late bicellular pollen (LBC), tricellular pollen (TCP) and mature pollen (MPG).

For semi in vivo pollen tube growth, a transgenic marker line harboring MGH3p::MGH3-eGFP and ACT11p::H2B-mRFP[10] was used to pollinate WT emasculated pistils. After 2 hours, the pollinated pistil was excised and placed on double sided tape. The excised pistil was then cut at the junction of style and ovary and placed gently on solidified agarose pollen germination medium[11]. The pistil was incubated for an additional 4 hours for the pollen tubes to emerge from the cut end of the style. The pollen tubes were harvested using a 25G needle and immediately frozen in liquid nitrogen and subsequently used for the RNA-seq library preparation as described in[3].

Total RNA was isolated from each sample using the RNeasy Plant Kit (Qiagen) according to the manufacturer's instructions. RNA was DNAse-treated (DNA-freeTM Kit Ambion, Life Technologies) according to the manufacturer's protocol. RNA yield and purity were determined spectrophotometrically and using an Agilent 2100 Bioanalyzer. cDNA was prepared using a slightly modified SmartSeq2 protocol in which cDNA is synthesized from poly(A)+ RNA with an oligo(dT)-tailed primer[4,12]. The final libraries were prepared using a low-input Nextera protocol[13]. Libraries were sequenced on a NextSeq500 instrument with single-end 75 bp read length (SE75).

A transgenic line expressing EC1.1p:NLS-3xGFP was cultured and used for *Arabidopsis* egg cell isolation as previously described[14]. Three replicates of 25 to 30 pooled egg cells were used for RNA extraction, RNA-seq library preparation and Illumina Next Generation Sequencing[15].

Tomato growth conditions, RNA isolation and sequencing

Solanum lycopersicum (tomato accession Nagcarlang, LA2661) seeds were obtained from the Tomato Genetics Resource Center (TGRC, https://tgrc.ucdavis.edu/) and grown in the Brown University Greenhouse (Providence, RI, USA). Dry pollen grains were collected from stage 15 flowers[16] into 500µl eppendorf tubes. Pollen tubes were grown in 300µl of pollen growth medium in a 750µl eppendorf tube that was incubated in a 28°C water bath. Pollen tubes were grown at a density of ~1000 pollen grains/µl. The pollen germination medium[17] comprised 24% (w/v) polyethylene glycol (PEG) 4000, 0.01% (w/v) boric acid, 2% (w/v) Suc, 20 mm MES buffer, pH 6.0, 3 mm $Ca(NO_3)2{\cdot}4H_2O$, 0.02% (w/v)

$MgSO_4 \cdot 7H_2O$, and 1 mm KNO_3. Pollen tubes were grown for 1.5 hours, 3 hours, or 9 hours before they were collected by centrifugation (1000 x g) for 1 minute. Pollen germination medium was carefully removed by pipetting to avoid disrupting the loose pollen tube pellet. Independent pollen collections were made for each of three biological replicates at each time point. Eppendorf tubes containing pollen tubes were immediately flash frozen in liquid N_2, then stored at -80°C, or put directly on a dry-ice cooled metal block for cell disruption by grinding with a frozen plastic pestle (Kontes). Total RNA was extracted using the RNeasy Plant Kit (Qiagen). RNA samples were evaluated by Agilent 2100 Bioanalyzer (Brown University Genomics Core Facility) before RNA-seq library preparation (polyA selection) and Illumina HiSeq, (150bp, paired end) sequencing were performed by Genewiz (South Plainfield, New Jersey. USA).

Maize growth conditions, RNA isolation and sequencing

Maize plants (inbred line B73) were grown in an air-conditioned greenhouse at 26°C under illumination of about 400 μmol/m2/sec with a 16-h photoperiod (21°C night temperature) and air humidity between 60-65%. Fresh mature pollen grains were harvested as described[18]. Pollen tubes were germinated and grown for 2 hours *in vitro* using liquid pollen germination medium[19]. Total RNA was extracted from each three biological replicates of 100 mg pollen grains/pollen tubes by using a Spectrum™ Plant Total RNA Kit (Sigma-Aldrich) according to manufacturer's instructions. 250 ng of total RNA was each used for library construction. RNA-seq was carried out as described in the Illumina TruSeq Stranded mRNA Sample Preparation Guide for the Illumina HiSeq 1000 System (Illumina) and the KAPA Library Quantification Kit (Kapa Biosystems). Data from sperm cells, egg cells and various zygote stages were taken from published data[18].

Compiling gene expression atlases

RNA data of different samples from nine species (*Physcomitrium patens*, *Marchantia polymorpha*, *Ginkgo biloba*, *Picea abies*, *Amborella trichopoda*, *Oryza sativa*, *Zea mays*, *Arabidopsis thaliana*, *Solanum lycopersicum*) were grouped in ten different classes (organs) (flower, female, male, seeds,

spore, leaf, stem, apical meristem, root meristem, root) (Table 1, Supplementary Table 1). For male and female reproductive organs samples we also included different sub-samples (female: egg cell, ovary, ovule; Male: microspore, bicellular pollen, tricellular pollen, mature pollen, pollen tube, generative cell, sperm) for each species (Table 1, Supplementary Table 1). A total of 4,806 different RNA sequencing samples were used, from which 4,672 were downloaded from the SRA database and 134 obtained from our experiments (see above). Publicly available RNA-seq experiments data were downloaded from ENA[20], as described in CoNekt-Plants[21]. Proteomes and CDSs of each species were downloaded from different sources (Supplementary Table 20). The raw reads of each sample were mapped to the coding sequences (CDS) with Kallisto v.0.46.1[22] to obtain transcripts per million (TPM) gene expression values. If the reads came from single cell samples (egg cell, ovule, sperm, generative cell), we removed the samples that have <1M reads mapped, and for the other samples we removed those with <5M reads mapped (Supplementary Table 1). All those samples were used to calculate Highest Reciprocal Rank (HRR) networks, where two genes with HRR<100 were connected[23]. For comparative expression analysis, an additional filter was applied by keeping only samples with a Pearson correlation coefficient (PCC) >=0.8 to at least one other sample of the same type (e.g. flower to flower) (Supplementary Table 1). Additionally, we included the expression matrix of *Selaginella moellendorffii* which has 18 samples (Supplementary Table 1), and exclusively for the database (see section Constructing the co-expression network and establishing the EVOREPRO database) the expression matrices of two unicellular algae (*Chlamydomonas reinhardtii* and *Cyanophora paradoxa*) and *Vitis vinifera*[24]. Finally, genes with median expression levels >2 TPM were considered as expressed[25]. All expression matrices are available for download from http://www.gene2function.de/download.html.

Phylogenomic and phylostratigraphic analysis

We used proteomes of 23 species representing key phylogenetic positions in the plant kingdom (see Supplementary Table 20), to construct orthologous gene groups (orthogroups) with Orthofinder v2.4.0[26], where Diamond v0.9.24.125[27] was used as sequence aligner. A species tree, of the 23

individuals, based on a recent phylogeny including more than 1000 species[28] was used for the phylostratigraphic analysis. The phylostratum (node) of an orthogroup was assessed by identifying the oldest clade found in the orthogroup[29] using ETE v3.0[30]. Briefly, for each orthogroup all the corresponding species of the genes were identified, and then the node in the species tree was assigned by identifying the node of the last common ancestor of all these species. To test whether a specific phylostratum is enriched in an organ, we randomly selected (without replacement) the number of observed organ-specific genes 1000 times. The empirical p-values were obtained by calculating whether the observed number of orthogroups for each phylostratum was larger (when testing for enrichment) or smaller than (testing for depletion) than the number obtained from the 1000 sampling procedure. The p values were FDR corrected[31] using a cutoff of 0.05.

Functional annotation of genes and identification of transcription factor and kinase families

The proteomes of the ten species included in the transcriptome dataset were annotated using the online tool Mercator4 v2.0 (https://www.plabipd.de/portal/web/guest/mercator4/-/wiki/Mercator4/recent_changes). This tool assigns Mapman4 bins to genes[32]. Transcription factors and kinases were predicted using iTAK v1.7a[33]. Additional transcription factors were identified using the online tool PlantTFDB v5.0 (http://planttfdb.cbi.pku.edu.cn/prediction.php)[34].

GO and PO annotations of Arabidopsis were obtained from the database TAIR (https://www.arabidopsis.org/). We used only genes with GO experimental evidence codes: Inferred from Experiment (EXP), Inferred from Direct Assay (IDA), Inferred from Physical Interaction (IPI), Inferred from Mutant Phenotype (IMP), Inferred from Genetic Interaction (IGI) (for more information check: http://geneontology.org/docs/guide-go-evidence-codes/). In order to see if *Arabidopsis* organ-specific genes have a know annotation that corresponds to the assigned organ, we used the PO annotation and classified them into 10 groups: flower, female, male, seeds, leaf, stem, apical meristem, root meristem, root, and other. The group corresponding to 'others' include annotations for organs or tissues not included in this study (i.e. hypocotyl epidermis, fruit septum) and annotations that could

correspond to more than one organ (i.e. stomatal complex, guard cell). Then we calculated the percentage of organ-specific genes that have annotations and the percentage of genes that have annotations in agreement with the assigned organ. *Arabidopsis* GO annotations were used to annotate functionally organ-specific orthogroups.

Functional enrichment analysis

Functional enrichment of the list of organ-specific and cluster-specific genes of each species, and genes gained in each node, was calculated using the bins predicted with Mercator 4 v2.0. Briefly, for a group of *m* genes (e.g., genes specifically expressed in *Arabidopsis* root), we first counted the number of Mapman bins present in the group, and then evaluated if these bins were significantly enriched or depleted by calculating an empirical *p*-value using the resampling method. The empirical *p*-value that tests whether a Mapman bin (term) is enriched in a collection of *m* genes is defined as:

$$P - \text{value}_{\text{term}} = \frac{\sum_{n=1}^{N} I(\text{pred}_{\text{observed}} \leq \text{pred}_{\text{sampled}})}{N}$$

Where$pred_{Observed}$*is* the number of times a term is observed, $pred_{Sampled}$is the number of times the term is observed when the terms of *m* genes are randomly sampled (without replacement) from all genes in the genome. *N* is the number of permutations, which was set to 1000. *I* is an indicator function, which takes a value of 1 when the event (in this case $pred_{Observed} \leq pred_{Sampl}$) is true, and 0 when it is not. For functional depletion analysis, a similar approach was followed, with *I* taking a value of 1 when $pred_{observed} \geq pred_{sampled}$. To account for multiple hypothesis testing, we applied a false discovery rate (FDR) correction to the p-values[31] using a cutoff of 0.05. Transcription factor and kinase enrichment were calculated following the same procedure.

Gene duplications and evolutionary rates of ubiquitous and organ-specific orthogroups

To analyse gene duplication, ubiquitous and organ-specific orthogroups with at least two sequences (13,329) were selected. The orthogroups with two sequences (2,188) were analysed separately, and if the two sequences belonged to the same species, one duplication was assumed. For each orthogroup with at least three sequences (11,141) gene trees were reconstructed. The protein sequences of each orthogroup were aligned using the same approach as described in the PhylomeDB pipeline[35]. Briefly, alignments in forward and reverse direction were obtained using three programs (MUSCLE v3.8.1551[36], MAFFT v7.475[37], and Kalign v2.04[38]). Then, the six alignments were combined using M-COFFEE v13.45.0.4846264[39], and trimmed with trimAl v1.4.rev15[40] using a consistency cut-off of 0.16667 and a gap threshold of 0.5. Phylogenetic trees were built using maximum likelihood approach as implemented in IQ-TREE v2.1.2[41] using the best-fit model identified by ModelFinder[42]. All gene trees are available in Supplementary Table 22. Duplication events were inferred using ETE v3.0[30] using the species overlap method[43].

In order to evaluate the evolutionary rates across the different expression profiles and phylostrata, single-copy ubiquitous and organ-specific orthogroups with at least two sequences were selected (1,621 orthogroups). The protein alignments were back-translated using trimAl and the CDSs of each species. The number of synonymous substitutions per synonymous site (dS), the number of nonsynonymous substitutions per nonsynonymous site (dN) and the dN/dS ratio were estimated using codeML from PAML v4.9[44] with settings seqtype = 1, CodonFreq = 2, runmode = -2. Because low dS values and saturation of substitutions may result in inaccurate dN/dS, we excluded the genes showing $dS < 0.01$ and dS or $dN > 2$. High dN/dS values (>10) were also discarded[45]. For each orthogroup, the average dN, dS, dN/dS was estimated for all pairwise comparisons. We compared the values of ubiquitous orthogroups and all organ-specific orthogroups and obtained the p-values using the Wilcoxon rank-sum test. The values of 15 phylostrata with at least 5 orthogroups were compared using the Wilcoxon rank-sum test to obtain the p-values, which were adjusted using a false discovery rate (FDR) correction[31] using a cutoff of 0.05.

References

1. Ashton, N. W. & Cove, D. J. The isolation and preliminary characterisation of auxotrophic and analogue resistant mutants of the moss, Physcomitrella patens. *Molec. Gen. Genet.* **154**, 87–95 (1977).
2. Hohe, A., Rensing, S. A., Mildner, M., Lang, D. & Reski, R. Day Length and Temperature Strongly Influence Sexual Reproduction and Expression of a Novel MADS-Box Gene in the Moss *Physcomitrella patens*. *Plant Biol (Stuttg)* **4**, 595–602 (2002).
3. Misra, C. S. *et al.* Transcriptomics of Arabidopsis sperm cells at single-cell resolution. *Plant Reprod.* **32**, 29–38 (2019).
4. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
5. Dupláková, N., Dobrev, P. I., Reňák, D. & Honys, D. Rapid separation of Arabidopsis male gametophyte developmental stages using a Percoll gradient. *Nat. Protoc.* **11**, 1817–1832 (2016).
6. Flores-Tornero, M. *et al.* Transcriptomic and Proteomic Insights into Amborella trichopoda Male Gametophyte Functions. *Plant Physiol.* (2020) doi:10.1104/pp.20.00837.
7. Lu, Y., Wei, L. & Wang, T. Methods to isolate a large amount of generative cells, sperm cells and vegetative nuclei from tomato pollen for "omics" analysis. *Front. Plant Sci.* **6**, 391 (2015).
8. Flores-Tornero, M. *et al.* Transcriptomics of manually isolated Amborella trichopoda egg apparatus cells. *Plant Reprod.* **32**, 15–27 (2019).
9. Honys, D. & Twell, D. Transcriptome analysis of haploid male gametophyte development in Arabidopsis. *Genome Biol.* **5**, R85 (2004).
10. Borges, F. *et al.* FACS-based purification of Arabidopsis microspores, sperm cells and vegetative nuclei. *Plant Methods* **8**, 44 (2012).
11. Boavida, L. C. & McCormick, S. Temperature as a determinant factor for increased and reproducible in vitro pollen germination in Arabidopsis thaliana. *Plant J.* **52**, 570–582 (2007).
12. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
13. Baym, M. *et al.* Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS*

ONE **10**, e0128036 (2015).

14. Englhart, M., Šoljić, L. & Sprunck, S. Manual Isolation of Living Cells from the Arabidopsis thaliana Female Gametophyte by Micromanipulation. *Methods Mol. Biol.* **1669**, 221–234 (2017).
15. Sprunck, S. *et al.* Elucidating small RNA pathways in *Arabidopsis thaliana* egg cells. *BioRxiv* (2019) doi:10.1101/525956.
16. Brukhin, V., Hernould, M., Gonzalez, N., Chevalier, C. & Mouras, A. Flower development schedule in tomato Lycopersicon esculentum cv. sweet cherry. *Sex. Plant Reprod.* **15**, 311–320 (2003).
17. Covey, P. A. *et al.* A pollen-specific RALF from tomato that regulates pollen tube elongation. *Plant Physiol.* **153**, 703–715 (2010).
18. Chen, J. *et al.* Zygotic Genome Activation Occurs Shortly after Fertilization in Maize. *Plant Cell* **29**, 2106–2125 (2017).
19. Schreiber, D. N., Bantin, J. & Dresselhaus, T. The MADS box transcription factor ZmMADS2 is required for anther and pollen maturation in maize and accumulates in apoptotic bodies during anther dehiscence. *Plant Physiol.* **134**, 1069–1079 (2004).
20. Harrison, P. W. *et al.* The european nucleotide archive in 2018. *Nucleic Acids Res.* **47**, D84–D88 (2019).
21. Proost, S. & Mutwil, M. CoNekT: an open-source framework for comparative genomic and transcriptomic network analyses. *Nucleic Acids Res.* **46**, W133–W140 (2018).
22. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
23. Mutwil, M. *et al.* Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. *Plant Physiol.* **152**, 29–43 (2010).
24. Ferrari, C. *et al.* Expression Atlas of Selaginella moellendorffii Provides Insights into the Evolution of Vasculature, Secondary Metabolism, and Roots. *Plant Cell* **32**, 853–870 (2020).
25. Wagner, G. P., Kin, K. & Lynch, V. J. A model based criterion for gene expression calls using RNA-seq data. *Theory Biosci.* **132**, 159–164 (2013).
26. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative

genomics. *Genome Biol.* **20**, 238 (2019).

27. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
28. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
29. Domazet-Loso, T., Brajković, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).
30. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
31. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).
32. Schwacke, R. *et al.* MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. *Mol. Plant* **12**, 879–892 (2019).
33. Zheng, Y. *et al.* iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. *Mol. Plant* **9**, 1667–1670 (2016).
34. Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J. & Gao, G. PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res.* **48**, D1104–D1113 (2020).
35. Huerta-Cepas, J. *et al.* PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res.* **39**, D556-60 (2011).
36. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
37. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).
38. Lassmann, T. & Sonnhammer, E. L. L. Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* **6**, 298 (2005).

39. Wallace, I. M., O'Sullivan, O., Higgins, D. G. & Notredame, C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* **34**, 1692–1699 (2006).

40. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

41. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

42. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

43. Huerta-Cepas, J., Dopazo, H., Dopazo, J. & Gabaldón, T. The human phylome. *Genome Biol.* **8**, R109 (2007).

44. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

45. Villanueva-Cañas, J. L., Laurie, S. & Albà, M. M. Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biol. Evol.* **5**, 457–467 (2013).