nature plants

RESOURCE

https://doi.org/10.1038/s41477-021-00958-2

Check for updates

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

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[6,10], Tomokazu Kawashima[11,12], Michael Borg[11], Frédéric Berger[11], Jacob Goldberg[13], Mark Johnson[13], David Honys[9], David Twell[7], Stefanie Sprunck[3], Thomas Dresselhaus[3], Jörg D. Becker[6,10] ✉ and Marek Mutwil[1] ✉

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 programmes. However, no large-scale analyses of genomic and transcriptomic data have been performed for land plants. To remedy this, we generated gene expression atlases for various organs and gametes of ten plant species comprising bryophytes, vascular plants, gymnosperms and flowering plants. A 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, which indicates that male reproduction is highly specialized. The expression atlas capturing pollen development revealed numerous transcription factors and kinases essential for pollen biogenesis and function.

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 ultraviolet stress encountered on land, which allowed them to colonize most terrestrial surfaces. The earliest land plants were probably not equipped with these adaptations, and many of these adaptations were probably 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 CO_2 drawdown[6], which significantly shaped the geosphere and biosphere of Earth[7]. To enable soil attachment and nutrient uptake, the first land plants only had rhizoids, which are filamentous structures homologous to root hairs[8]. Roots later evolved to provide increased anchorage (and therefore increased height), nutrient uptake and enable survival in more arid environments. In parallel with innovations in 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, which 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 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 constitute 90% of all land plants and serve as the basis for the terrestrial food chain, either directly or indirectly. The use of model plants such as *Arabidopsis thaliana* and

[1]School of Biological Sciences, Nanyang Technological University, Singapore, Singapore. [2]Max Planck Institute for Molecular Plant Physiology, Potsdam-Golm, Germany. [3]Cell Biology and Plant Biochemistry, University of Regensburg, Regensburg, Germany. [4]Laboratory of Molecular Bacteriology, Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium. [5]VIB, Center for Microbiology, Leuven, Belgium. [6]Instituto Gulbenkian de Ciência, Oeiras, Portugal. [7]Department of Genetics and Genome Biology, University of Leicester, Leicester, UK. [8]School of Life Sciences, Gibbet Hill Campus, The University of Warwick, Coventry, UK. [9]Laboratory of Pollen Biology, Institute of Experimental Botany of the Czech Academy of Sciences, Prague, Czech Republic. [10]Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Oeiras, Portugal. [11]Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna, BioCenter (VBC), Vienna, Austria. [12]Department of Plant and Soil Sciences, University of Kentucky, Lexington, KY, USA. [13]Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University, Providence, RI, USA. ✉e-mail: jbecker@igc.gulbenkian.pt; mutwil@ntu.edu.sg

maize and technical advances allowing live-cell imaging of double fertilization have been instrumental to several major discoveries[16]. When assessing the 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, for example, by fluorescence-activated cell sorting[17]. Analyses 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* (also known as *GCS1*) and *GEX2* (ref. [18]). However, as novel genes that control the development of male and female gametes[9,10] or their functions[19] are still being discovered, 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. A comparison of 208 genomes revealed two bursts of genomic novelties in the ancestors of streptophytes and land plants, which were most probably 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 the 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 that 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 that are 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 (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 the following 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 were used to construct co-expression networks and to create an online EVOREPRO database to enable 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 (Methods). For all ten species, approximately 71% of their genes were expressed in at least one structure (Supplementary Table 2). Interestingly, the male sample had a lower percentage (38%), followed by root meristems (46%), while the other organs had 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 the top 5% of the SPM values was used to identify the organ-specific genes for all species (Supplementary Fig. 1 and 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 showed that the TPM values of the organ-specific genes in all cases are highest in the organs that the genes are specific to (Fig. 1e and 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 correspond to our predictions. We divided the PO annotations into 11 groups: 10 corresponding to the organs we studied and 1 named 'others', which included the annotations that could either correspond to more than one organ (that is, guard mother cell could correspond to leaf or stem) or represent organs and tissues not analysed in this study (for example, hypocotyl and coleoptile). From the total number of genes classified as organ-specific in *Arabidopsis* (9,798 genes), only an average of 11.4% had a 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 these genes showed correspondence with the organ to which they were assigned (that is, the higher percentage of flower-specific genes had PO annotations related to flowers; Fig. 1f), except for leaf-specific genes, for which most genes belonged 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*, for which 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 analysed 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 and Supplementary Table 2). In *A. thaliana*, the higher percentage of male-specific genes was in agreement with previous studies that showed high specialization of the male transcriptome[23]. Conversely, stem, spore, apical meristem, root meristem, flower and female showed values lower than 3% (Fig. 1g and Supplementary Table 2). This is in line with 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 an important part of the transcriptome, with male and root samples possessing the most specialized transcriptomes.

Conservation of organ transcriptomes across species. Our above analysis suggests that organ-specific gene expression is widespread;

Table 1 | Organs, tissues and cell types used in the expression atlases analysed

Organ, tissue, cell type	*Marchantia*	*Physcomitrium*	*Selaginella*	Gingko	Spruce	*Amborella*	*Arabidopsis*	Tomato	Rice	Maize
Flower	NA	NA	Strobili (2)	Microstrobilus (2), strobili (5)	NA	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 (tricellular, mature) (14), sperm (5)	Pollen (mature, tube) (45), sperm (7), microspore (5)
Seeds	NA	NA	NA	Kernel (5)	–	–	Endosperm (9), seed (young) (10), seed (4), seed (germinating) (6)	Seeds (5–30 days post anthesis) (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)	–	NA	NA	NA	NA	NA	NA	NA
Leaf	Thallus (38)	Leaflets (43)	Microphyll (2)	Leaves (81)	Needles (63)	Leaf (3)	Leaf (14)	Leaves (49)	Leaves (644)	Leaves (133)
Stem	NA	NA	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	NA	NA	Meristematic zone (3)	–	–	–	Meristematic and quiescent centre zone (10)	Meristematic zone (3)	Meristematic zone (2)	Meristematic zone (2)
Root	NA	NA	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)

NA (not applicable) indicates that a given species does not have a corresponding organ/tissue. The en dash (–) indicates that our dataset does not include data for the corresponding organ. The total number of experiments per sample are indicated in parentheses.

therefore, we set out to investigate whether these patterns are conserved across species. To this end, we employed a Jaccard distance method to investigate which organs specifically expressed similar sets of orthogroups. Values range 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, for example, the root-specific transcriptome is conserved across angiosperms, then the Jaccard distance of root versus root transcriptomes (for example, *Arabidopsis* root versus rice root) should be lower than when comparing root versus non-root transcriptomes (for example, *Arabidopsis* root versus rice leaf).

The analysis revealed that *Arabidopsis* flower-, male-, seed-, stem- and root-specific transcriptomes were significantly more similar to the corresponding organ in the other species (Wilcoxon rank-sum test $P < 0.05$; 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). Meanwhile, for other organs, some species showed significance for 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) and root meristem (4 out of 5) (Fig. 2b and Supplementary Fig. 3). Conversely, spore (0 out of 2) samples did not show similar transcriptomes across *Marchantia* and *Physcomitrium* (Fig. 2b and 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), which suggests that flagellated sperm of bryophytes employ a unique male transcription programme 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 and Supplementary Tables 4 and 5). The analysis revealed that many functions are depleted in male and root samples in at least 50% of the species, which indicates that most cellular processes in male and roots are significantly repressed ($P < 0.05$; Fig. 2c and 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 the 'not assigned' bin, which indicates that these organs are enriched for genes with unknown functions.

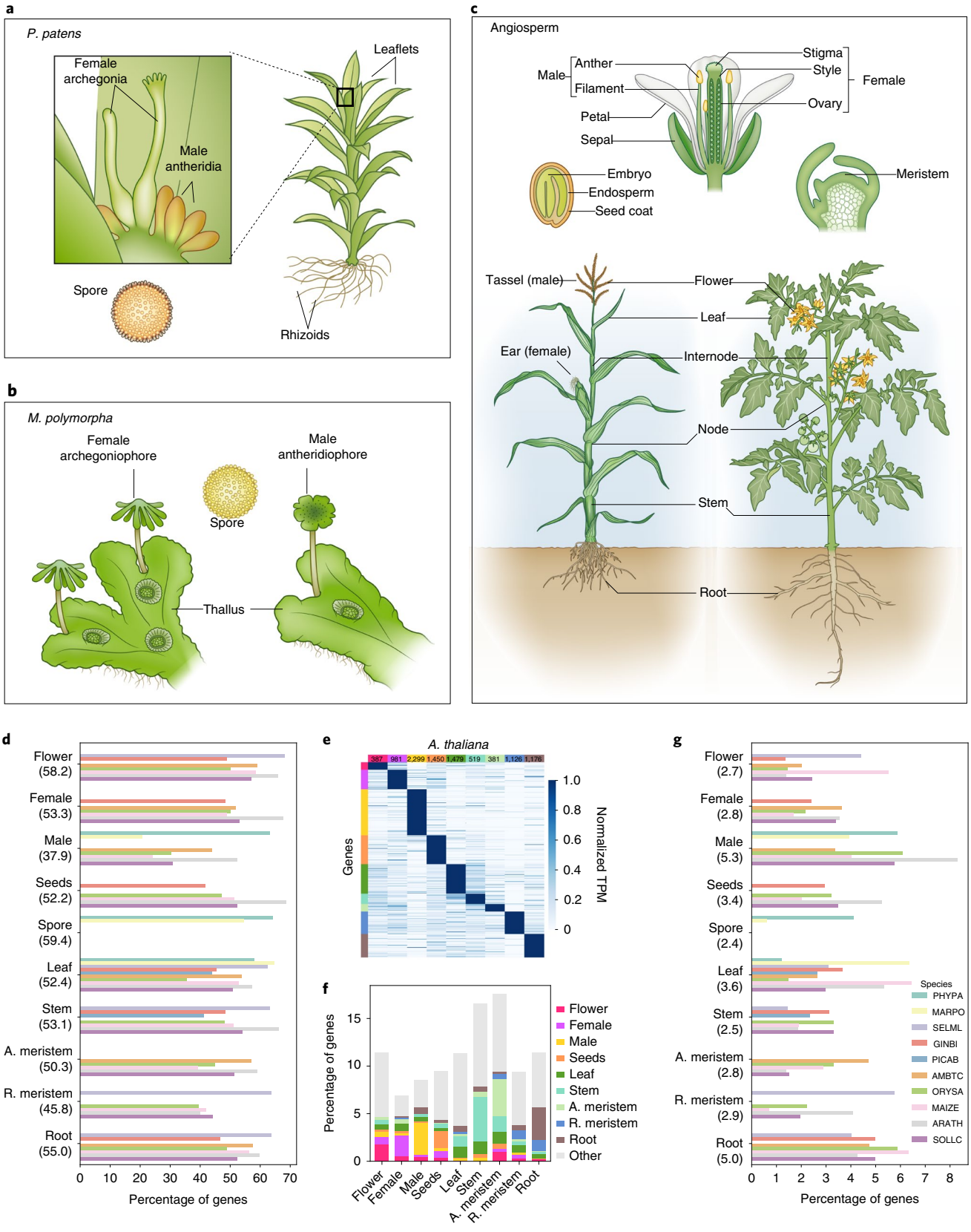
Since the organ-specific genes (Supplementary Table 3) are probably important for the formation and function of the organ, we investigated organ-specific transcription factors (Supplementary Table 6) and receptor kinases (Supplementary Table 7). An enrichment analysis of transcription factors (69 families) and kinases (142 families) showed that apical meristem and root samples are highly enriched in transcription factors, while male and apical meristem are enriched for kinases (Fig. 2c). In apical meristems, some of the enriched transcription factor families (C2C2-YABBY and GRF) were associated with the regulation, development and differentiation of the meristem[25,26]. In roots, the enriched transcription factors (MYB, bHLH, WRKY and NAC) were related to biotic and abiotic stress response and root development[27,28]. These organ-specific genes are therefore prime candidates for further functional analyses (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 the appearance of orthogroups and their expression patterns, we used the proteomes of 23 phylogenetically representative species and a species tree derived from the One Thousand Plant Transcriptomes Initiative, 2019. Orthogroups (orthologous gene groups) were obtained using Orthofinder (v.2.4.0)[29] (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 indicates the oldest phylostratum and node 23 indicates 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% for node 1, 10% for node 3), representing 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). An analysis of the phylostrata in each species revealed a similar distribution of the orthogroups, with most of them belonging to node 1 (~34%) or were species-specific (~31%; Supplementary Fig. 6).

To investigate whether the different phylostrata show different expression trends, we used RNA-seq data to survey orthogroups that contain at least two species. This resulted in 43,883 (33% of the total number of orthogroups) meeting this criterion. Then, each orthogroup was assigned to the following different expression profiles: ubiquitous (not specific in any organ), not conserved (for example, root-specific in one species, flower-specific in others) or organ-specific (see Methods for details and Supplementary Table 8 for the 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 the phylostratigraphic age decreased (<5% specific genes in node 1 versus ~25% in node 13), which indicates that younger orthogroups are recruited to specific organs (Fig. 3b). Using Gene Ontology (GO) annotations of *Arabidopsis* genes with experimental evidence, we observed that organ-specific orthogroups have relevant functions for the assigned organ (Supplementary Table 9).

Fig. 1 | Expression atlases for seven land plant species. a–c, Depiction of the different organs, tissues and cells collected for *P. patens* (**a**) *M. polymorpha* (**b**) and angiosperms (**c**). **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 coloured bars as in **f**). The numbers beneath the organs (*y* axis) indicate the average percentage of genes for all species. A. meristem, apical meristem; R. meristem, root meristem. **e**, Expression profiles of organ-specific genes from *A. thaliana*. Genes are in rows, organs in columns and the genes are sorted according to the expression profiles (for example, 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 to 1. Bars on the left of each heatmap show the organ-specific genes and correspond to the samples on the bottom. **f**, Percentage of organ-specific *Arabidopsis* genes with PO annotations for the ten organs. The 'Other' category indicates 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. Species are indicated by the following mnemonics: PHYPA, *P. patens*; MARPO, *M. polymorpha*; SELML, *S. moellendorffii*; GINBI, *G. biloba*; PICAB, *P. abies*; AMBTC, *A. trichopoda*; ORYSA, *O. sativa*; MAIZE. *Z. mays*; ARATH, *A. thaliana*; SOLLC, *S. lycopersicum*.

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 (nodes 18, 20 and 22). As expected, the 9,416 ubiquitous orthogroups were mostly of ancient (nodes 1–7) origin, which suggests that these old orthogroups tend to show



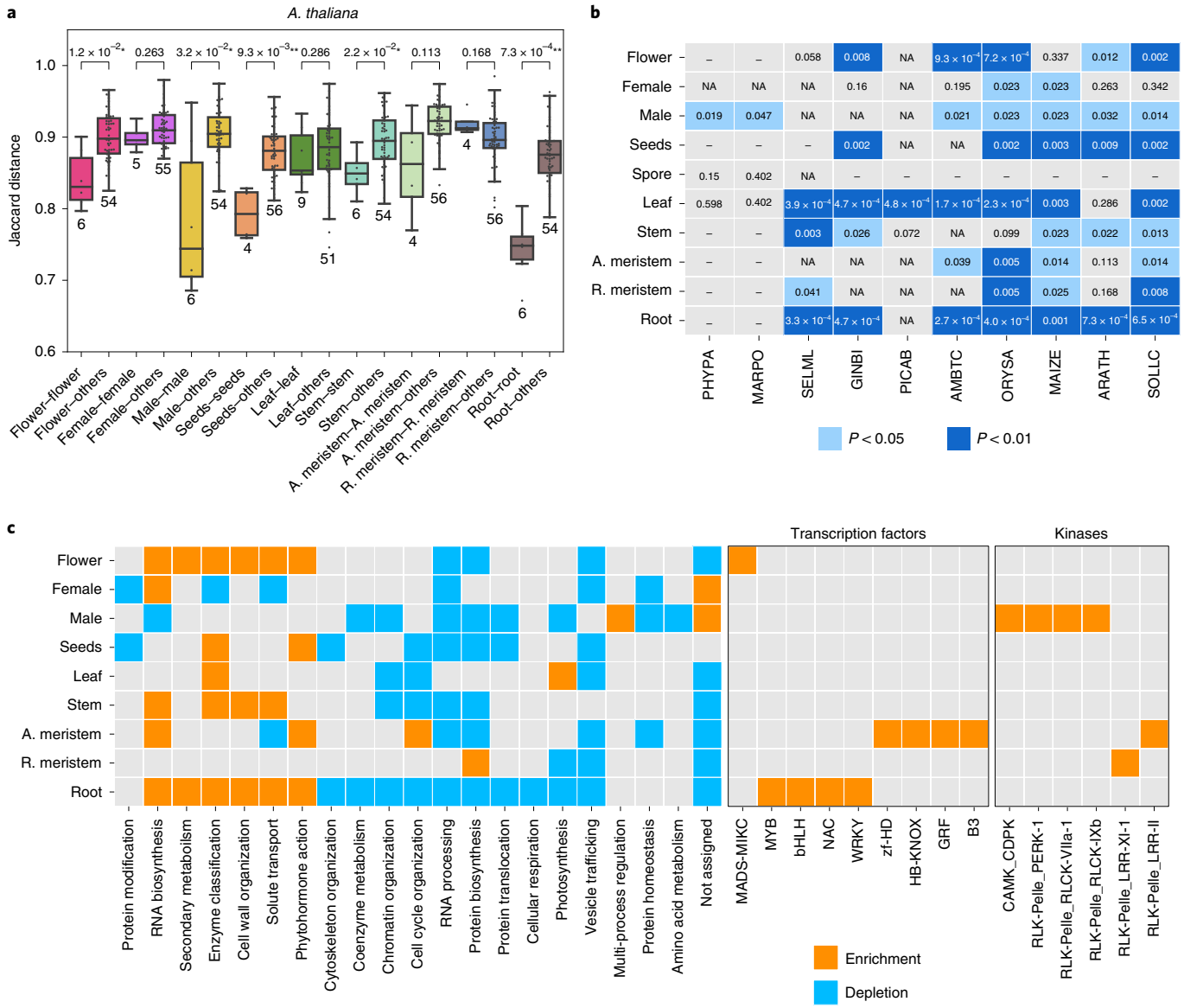


Fig. 2 | Comparison of organ-specific transcriptomes. a, Bar plot showing the Jaccard distances (*y* axis) when comparing the same samples (*x* axis, for example, male-male) and one sample versus the others (for example, male–others) for *A. thaliana*. Lower values indicate a higher similarity of the transcriptomes. The sample size (*n*) is indicated below each boxplot. 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 (centre line) and the first and third quartile (upper and lower hinges), and the whiskers extend to a maximum of 1.5× the interquartile range. $^{*}P < 0.05$, $^{**}P < 0.01$. **b**, Significantly similar transcriptomes are indicated by blue cells. Two-sided Wilcoxon rank-sum statistic was used to obtain the *P* values. **c**, Heatmap showing the significant ($P < 0.05$) functional enrichment or depletion 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, one-sided empirical *P* values were calculated using the functional enrichment analysis method (Supplementary Methods). The individual *P* values are presented in Supplementary Tables 4 and 5.

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 and 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 have 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], for which new genes were predominantly expressed in male reproductive cells[33]. This suggests 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, and Supplementary Fig. 7). Pollen also expressed a substantial proportion of old genes (species nodes 1–7 in Fig. 3c), which probably represents an old transcription programme 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 cannot rule out the possibility of an underestimation of the age in male samples,

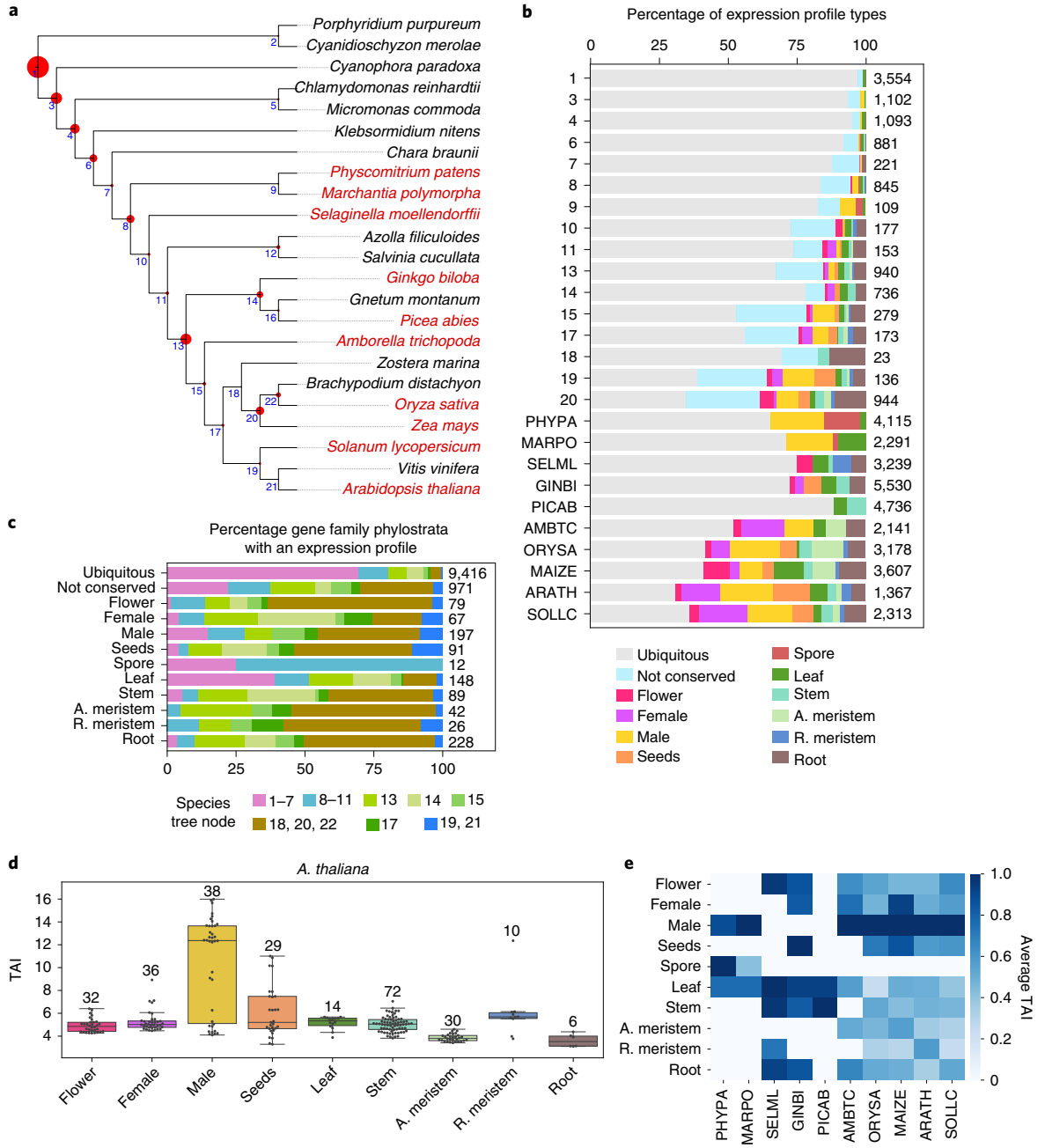


Fig. 3 | Genomic analysis of the organ specificity of orthogroups. a, Species tree of the 23 species for which we have inferred orthogroups. The phylogenetic relationship was based on the One Thousand Plant Transcriptomes Initiative, 2019. Species in red are the ones with transcriptomic data available. Blue numbers in the nodes indicate the node number (for example, 1: node 1). The red circles in the tree 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 as follows: ubiquitous (orthogroup is not organ-specific); not conserved (organ-specificity is not conserved in different species); or organ-specific (for example, root-specific). The numbers to the right of the chart indicate the number of orthogroups assigned to a node. **c**, Percentage of phylostrata (nodes) within the different expression profile types. The numbers to the right of the chart indicate the number of orthogroups assigned to the different expression profiles. **d**, Transcriptomic age index (TAI) values of the different organ-specific genes in *A. 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 (centre line) and the first and third quartile (upper and lower hinges), and the whiskers that extend to a maximum of 1.5× the 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.

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 analyses[34].

Phylostratigraphic and gene expression analyses reveal 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].

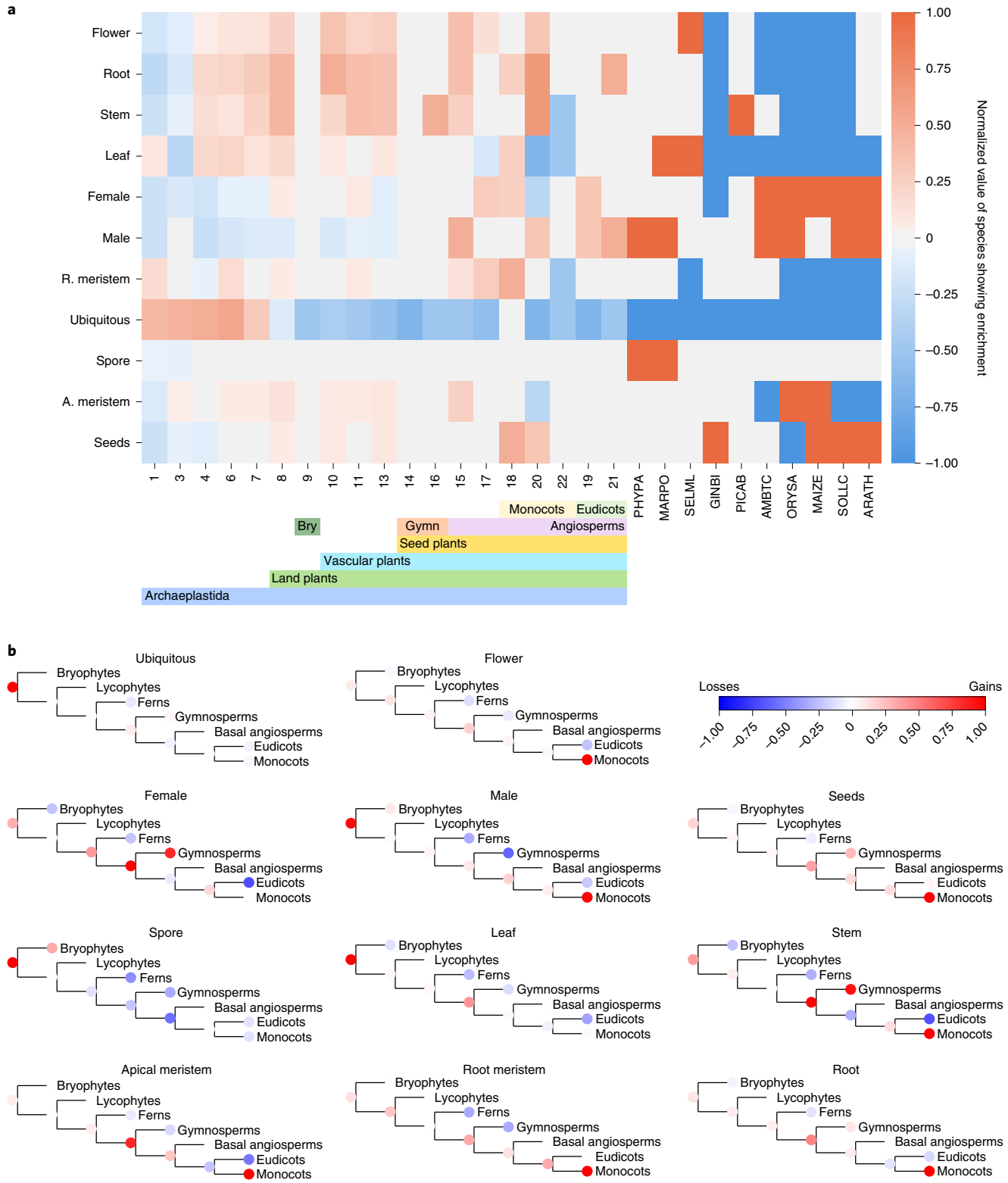


Fig. 4 | Evolutionary analysis of organs. a, Enrichment and depletion of organ-specific genes per node in the species tree (nodes in the *x* axis are the same as in Fig. 3a). The colours correspond to the number of species showing enrichment in each case (dark red: all species show enrichment, dark blue: all species show depletion). Horizontal bars below the node numbers show the main clades in different colours. 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.

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 analysed 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 programmes 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), while 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 have suggested that the evolution of novel morphologies was mainly driven by the reassembly and reuse of pre-existing genetic mechanisms, as exemplified by the conserved transcription programmes between flowers and cones in gymnosperms[36,41]. It was indicated that primitive root programmes 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 before the divergence of vascular plants. After the divergence of land plants, we can observe that there is incremental gene family gain in monocots for all three organs (roots, stems and flowers; Fig. 4b, indicated by red nodes) and 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. The 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 and 21) and species-specific nodes, while female-specific genes were enriched only in monocots (nodes 18 and 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 colour in the ancestral node of land plants, Fig. 4b). For male orthogroups, we observed 6 waves of gains (>15 orthogroups) in nodes 3, 8 (land plants), 13 (seed plants), 15 (angiosperms), 19 (eudicots) and 20 (monocots). From these nodes, parallel to gains, we also observed many losses (≥10 orthogroups) in 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) and 20 (monocots), and different waves of losses (Supplementary Table 11). Male orthogroups showed enrichment for protein modification, enzyme classification, RNA biosynthesis, cell cycle organization and phytohormone action, whereas female orthogroups showed enrichment only for RNA biosynthesis (Supplementary Table 12). Considering the gains and losses of orthogroups, for male-specific families, gains were mainly in the node ancestral to land plants and in monocots, whereas for female-specific families, gains were in seed plants and gymnosperms (Fig. 4b).

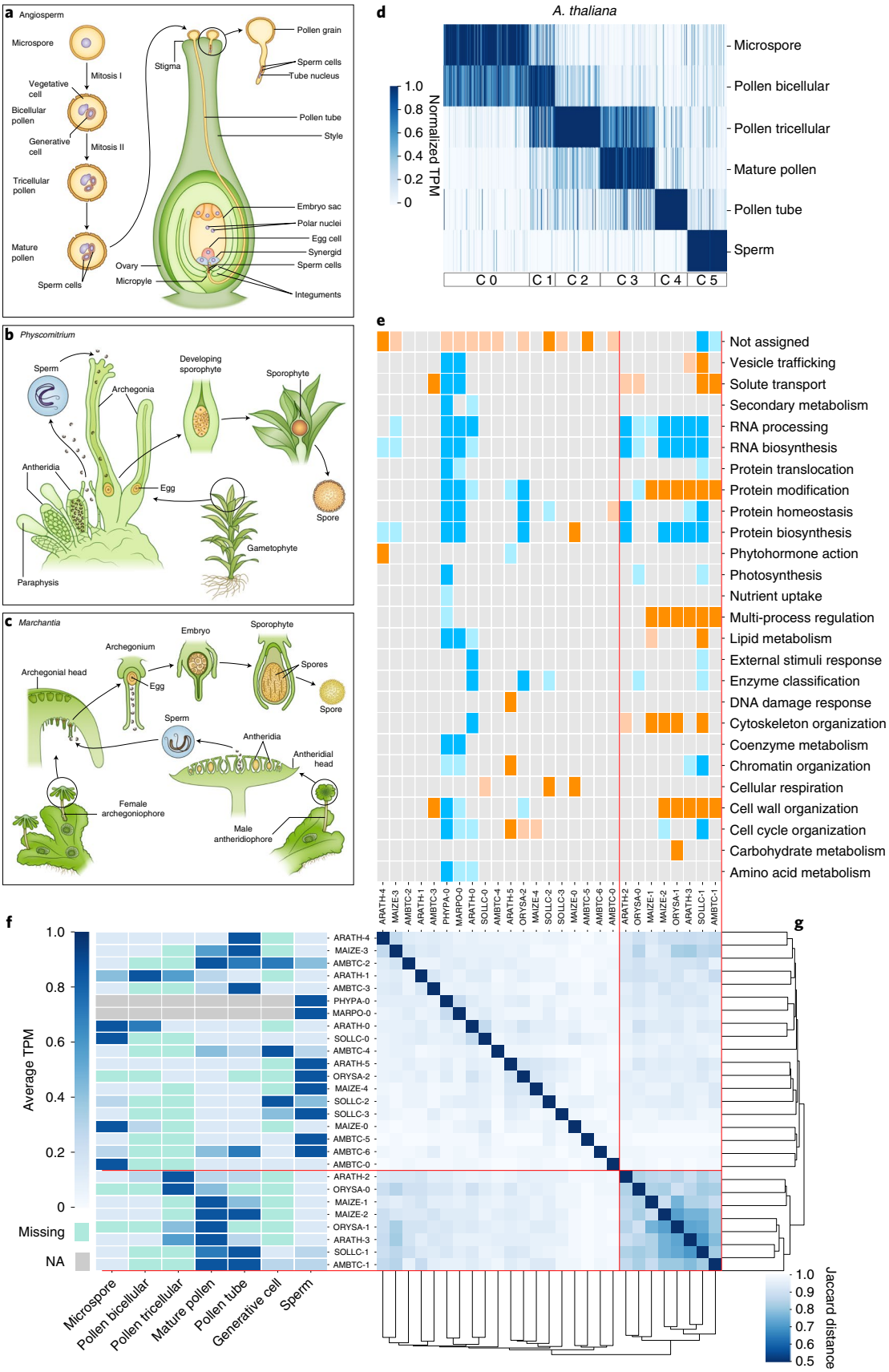
In summary, the genetic programmes for organ-specific genes are present in older nodes, before the divergence of land plants. Monocots seem to be the group with more gene family gains, which is in agreement with previous studies[45].

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 whether 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 (Methods). Interestingly, the number of duplications was much higher in orthogroups with a ubiquitous expression profile than in any other organ-specific group (Supplementary Fig. 8a). Conversely, the organ-specific orthogroups predominantly showed one or two duplications.

To test whether the organ-specific orthogroups evolved 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 (Methods). A total of 1,621 orthogroups were analysed, and the average pairwise dN, dS and dN/dS values were 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 values for ubiquitous and organ-specific orthogroups were less than 1, which suggests that there is purifying selection (Supplementary Fig. 8b), as 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 significantly lower median dN/dS values (Wilcoxon rank-sum test, $P = 1.4 \times 10^{-2}$ and 1.5×10^{-2}, respectively), and female and leaf orthogroups have significantly higher values (Wilcoxon rank sum test, $P = 3.4 \times 10^{-2}$ and 2.9×10^{-2}, respectively) (Supplementary Fig. 8b). For female and leaf orthogroups, the higher dN/dS values observed were mainly due to a significant difference in the dN rate

Fig. 5 | Comparison of male development across species. a–c, Schematic overviews of sexual reproduction in Angiosperms (**a**), *Physcomitrium* (**b**) and *Marchantia* (**c**). **d**, Heatmaps showing the expression of male samples genes for *A. thaliana*. Genes are in columns, sample names in rows. Gene expression is scaled to range between 0 and 1. Darker colour 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 colours: $P < 0.05$, dark colours: $P < 0.01$. In all cases, one-sided empirical *P* values were calculated using the functional enrichment analysis method (Supplementary Methods). 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. In **e–g**, the red horizontal and vertical lines are used to indicate the cluster containing predominantly mature pollen samples.

(Supplementary Fig. 8c), which suggests that there are higher rates of adaptive evolution. Interestingly, a recent study[48] also observed higher d*N*/d*S* values in genes expressed in style and ovules in *Solanum* species, thereby supporting our findings. However, the lower d*N*/d*S* values observed in male and leaf are mainly explained by significantly higher d*S* rates, which is a proxy for the mutation



rate and could indicate that these orthogroups are evolving faster. Other studies showed that genes expressed in pollen tend to have lower d*N*/d*S* values than genes not expressed in pollen, which is attributed to a stronger purifying selection on genes expressed in the haploid gametophyte[49]. Furthermore, high d*S* values were observed in genes predominantly expressed in the sperm and pollen tube of *Arabidopsis*[32]. We observed that male samples expressed younger transcriptomes (transcriptome age index (TAI) values; Fig. 3e), and since proteins that evolve rapidly could underestimate the phylostratigraphic age[34], we cannot exclude the possible effect of this higher evolutionary rate in male orthogroups on the TAI. However, we also observed higher d*S* rates for seeds, stems and roots (Supplementary Fig. 8d), which were 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 d*N*/d*S*, higher d*N* and lower d*S* in younger nodes (Supplementary Fig. 8e,f,g and Supplementary Table 13). Interestingly, node 14 (gymnosperms) showed the highest median d*N*/d*S*, whereas node 1 showed the lowest median value, which was significantly different from younger nodes (Supplementary Table 13). We observed that older orthogroups have significantly higher d*S* values, which points to fast evolving genes. Previous studies showed that older orthogroups have lower d*N*/d*S*, but did not observe large differences in d*S* values[46]. It is worth mentioning that monocots (node 20) seem to evolve faster than gymnosperms (node 14), and gymnosperms show significantly higher d*N*/d*S* than angiosperms (nodes 17 and 20), which can be explained by 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 transcription programmes of gametes. Sexual reproduction is a complex process that requires 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), while *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 transcription programmes of sexual reproduction are conserved in land plants, we applied *k*-means clustering on the male- and female-specific genes from 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 and Supplementary Table 15), plotted average expression profiles (Fig. 5f) and used Jaccard distances 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). The 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).

The 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*, for which the ovule was divided into two clusters (Supplementary Fig. 10 and Supplementary Tables 16 and 17). Interestingly, when we measured the Jaccard distances 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). The 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 analysed.

Analysis of signalling 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 subsamples (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 analysed whether the number of observed connections was similar to the number of expected connections (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*, for which 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), which suggests that the sperm cell transcriptome is distinctive, thereby 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, and 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 analysed 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 were described as unknown (Fig. 6c), which indicates that there are no experiments associated with these genes. It is important to note that the genes classified as 'none' have been found to have an effect in other organs, but since a pollen phenotype can be easily missed, this does not rule out the possibility that these genes are associated with male development. Also, many of these genes showed effects in

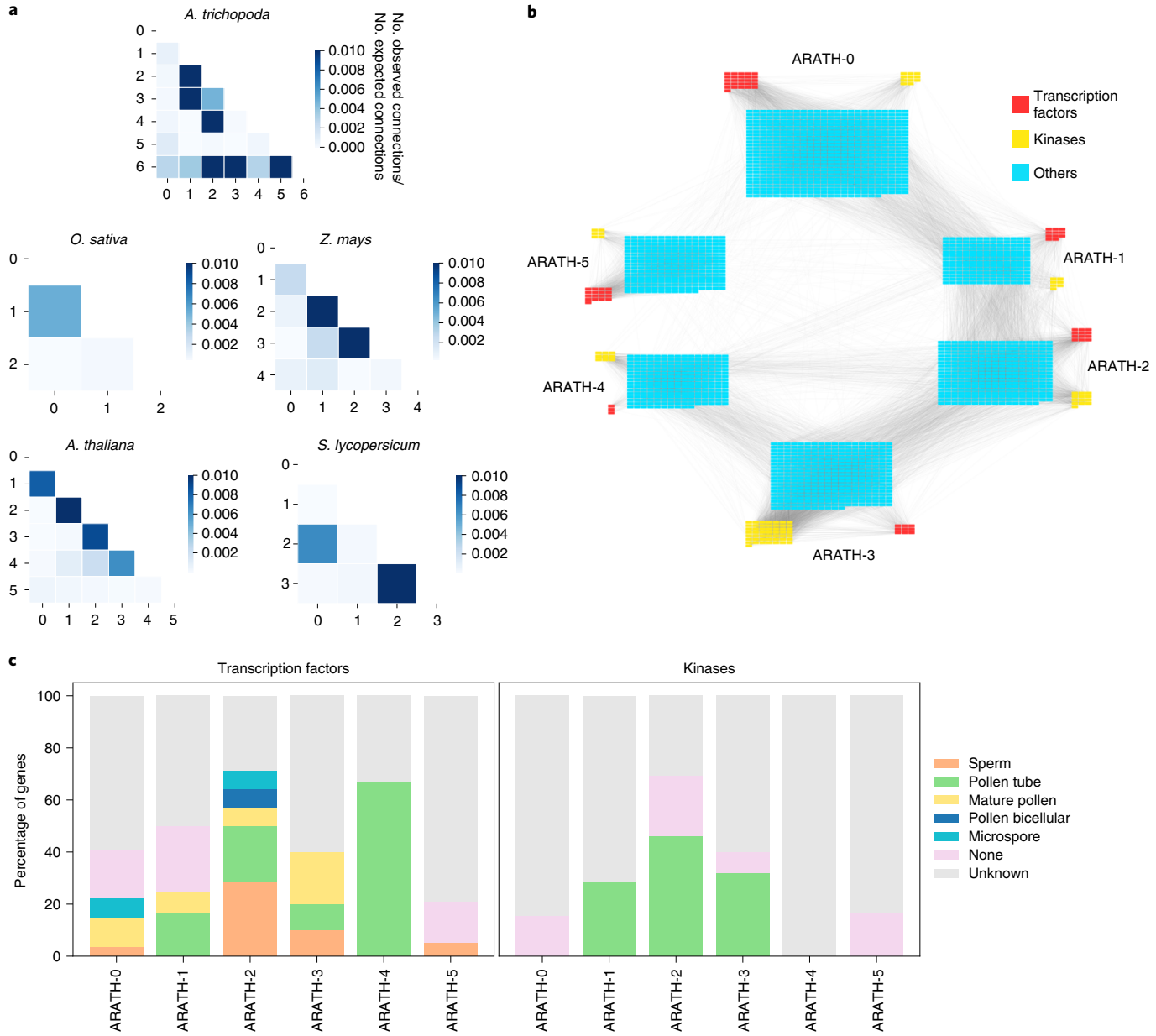


Fig. 6 | A network analysis of male clusters. a, Heatmaps showing the number of observed connections divided by the number of expected connections. Darker colours indicate more connections between clusters. **b**, *A. thaliana* co-expression network clusters showing the edges between the different clusters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

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 the 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* (grapevine, eudicot), *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, for example, >35% reads of a gene expressed in male organs for *Arabidopsis;* Supplementary Fig. 1). This can be accomplished for one (https://evorepro.sbs.ntu.

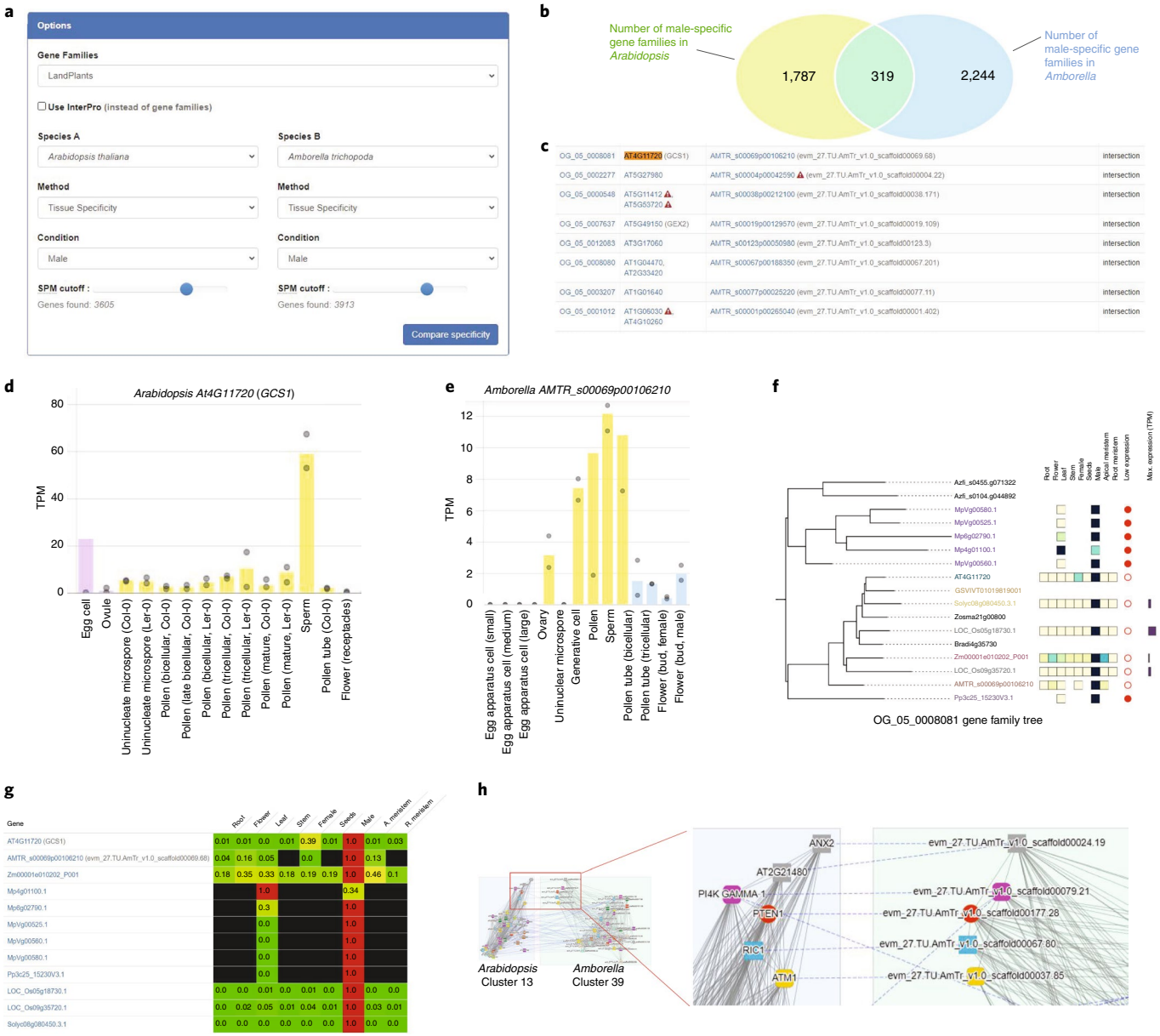


Fig. 7 | Features of the EVOREPRO database. a, Compare specificities tool. The dropdown menus allow selection of the species, orthogroups, organs, tissues, cell types and SPM cut-off values. The analysis is started by clicking on the "Compare specificity" button. **b**, The Venn diagram shows the number of unique and common orthogroups of male-specific genes in *Arabidopsis* and *Amborella*. The default SPM value cut-off 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 (<10 TPM). **d**, Expression profile of *GCS1* from *Arabidopsis*. The coloured columns indicate the average expression values in the different samples, while grey 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 grey 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 colour-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 the 'pollen tube' GO term in *Arabidopsis* (cluster 13, left) and *Amborella* (cluster 39, right). Nodes indicate genes, while solid grey 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.

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, respectively, from the drop-down menus, 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 cut-off' 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 cut-off. We left the slider at the default value (0.85) and clicked on the 'Compare specificities' button. The analysis revealed that 319 orthogroups are specifically expressed 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 and Supplementary Table 21). Interestingly, among the conserved genes, we observed *GCS1*, 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* 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 GO functional enrichment analysis of the gene[55]. As expected, the interactive, exportable expression profiles confirmed that the *Arabidopsis GCS1* and the *Amborella* orthologue (Fig. 7e; https://evorepro.sbs.ntu.edu.sg/sequence/view/45084) are male-specific, with the highest expression in sperm and pollen. Clicking on the gene family identifier (OG_05_0008081) redirects to the gene family page (https://evorepro.sbs.ntu.edu.sg/family/view/139708), which, among others, contains an interactive phylogenetic tree (Fig. 7f; https://evorepro.sbs.ntu.edu.sg/tree/view/88288) and heatmap (Fig. 7g; https://evorepro.sbs.ntu.edu.sg/heatmap/comparative/tree/88288/row) showcasing the male-enriched expression profiles for most of the genes in this family. Therefore, this approach can be used to identify conserved, organ-specific genes across two species and to study family-wide expression patterns.

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

Furthermore, a table labelled 'Similar clusters' reveals the identity of similar (defined by the Jaccard index, see Methods) co-expression clusters in other species, which can be used to rapidly identify functionally equivalent clusters across species. To exemplify this, we first clicked on the 'Jaccard index' table header to sort the similar clusters and clicked on the 'Compare' link next to Cluster 39 from *Amborella* (https://evorepro.sbs.ntu.edu.sg/graph_comparison/cluster/113/769/1). This redirected us to a co-expression network page showing the genes (nodes), co-expression relationships (grey edges) and orthologous genes (coloured 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* (ref. [56]), *BUPS2* (*At2g21480*)[57], *PI4K Gamma-1* (ref. [58]), *PTEN1* (ref. [59]), *RIC1* (ref. [60]) and *ATM1* (ref. [61]). To conclude, this approach can be used to uncover functionally equivalent, conserved transcription programmes.

Discussion

To study the evolution of plant organs and gametes, we generated and analysed gene expression profiles for ten land plants, comprising representatives of bryophytes, lycophytes, gymnosperms, sister to all angiosperms, monocots and eudicots. The main advantage of our analysis 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), which suggests 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 the different representation of the various tissues found in the organs (for example, buds, stamen filaments and carpels in *Arabidopsis* versus whole flowers in tomato; Table 1), we observed a significant and robust conservation of the transcription programmes of the analysed organs. With the surprising exception of female gametes, the corresponding transcriptomes tended to be more similar across the analysed samples (Fig. 2b and Supplementary Figs. 3 and 4). As also observed in previous studies, roots, male and seeds express conserved expression programmes[42,62]. Another exception is seen in the leaf-like organs of bryophytes (leaflets and thallus for *Physcomitrium* and *Marchantia*, respectively), which indicates that these organs have evolved independently from the leaves of flowering plants or that they have substantially diverged since the last common ancestor of flowering plants and bryophytes.

Next, we examined the expression patterns of expressed orthogroups as a function of their age. We report a clear trend of older orthogroups having more ubiquitous (that is, 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 a specialized function in a plant organ, tissue or cell type, rather than being integrated into fundamental biological pathways. As expected, male gametes showed the highest expression of the youngest genes (Fig. 3d,e and 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 observed a significant enrichment for orthogroups that appeared long before the corresponding organ (Fig. 4a), which shows that the establishment of organs relies heavily on the co-option of existing genetic material, as previously suggested[20,36,41]. Flowers (appearance in angiosperms), stems (appearance in vascular plants) and roots (appearance in vascular/seed plants) showed 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). This suggests that during monocot evolution, organ-specific transcriptomes were enriched with novel functions. Surprisingly, eudicots showed 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), which suggests 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 further investigation and analysis, which are made possible by identifying the corresponding orthogroups (Supplementary Table 11) and genes (Supplementary Table 8).

Our comparative analysis of male gamete development revealed that transcription programmes of mature pollen form well-defined clusters and are therefore conserved across species (Fig. 5f,g). The mature pollen clusters were enriched for processes related to signalling (protein modification comprising protein kinases) and cell wall remodelling (Fig. 5e), which are probably 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), which suggests that the processes taking place in the early stages of pollen development are yet to be uncovered. Furthermore, the female gametes showed poor clustering, which indicates that there is overall low conservation of the transcription programmes 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 analyses, which is now made possible by our identification of these genes (Supplementary Tables 14 and 16). Of particular interest are the male-specific transcription factors and kinases that we identified (Fig. 6c), which are presumably 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 plant 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 noncoding DNA (for example, *cis*-regulatory elements) and noncoding RNA (for example, long noncoding RNAs and microRNAs).

Methods

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

Compiling gene expression atlases. RNA data of different samples from nine species (*P. patens*, *M. polymorpha*, *G. biloba*, *P. abies*, *A. trichopoda*, *O. sativa*, *Z. mays*, *A. thaliana* and *S. lycopersicum*) were grouped in ten different classes (organs) (flower, female, male, seeds, spore, leaf, stem, apical meristem, root meristem and root) (Table 1 and Supplementary Table 1). For male and female reproductive organs, we also included different subsamples (female: egg cell, ovary and ovule; male: microspore, bicellular pollen, tricellular pollen, mature pollen, pollen tube, generative cell and sperm) for each species (Table 1 and Supplementary Table 1). A total of 4,806 different RNA-seq 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 data were downloaded from ENA (https://www.ebi.ac.uk/ena/browser/home). For more details, see the Supplementary Methods.

Identifying organ-specific genes. Organ-specific genes based on expression data were detected by calculating the SPM, using a similar method as described in ref. [65]. For each gene, we calculated the average TPM value in each sample (for example, root, leaf and 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 the SPM value of a gene 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 the 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 values were 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 the organ-specific transcriptome of a species was significantly similar to a corresponding sample in the other species (for example, *Arabidopsis* root versus rice root or tomato root), we tested whether the d_j values comparing the same sample were smaller (that is, more similar) than d_j values comparing the sample to the other samples (for example, *Arabidopsis* root versus rice flower, rice leaf, tomato flower or tomato leaf). We used Wilcoxon rank-sum to obtain the *P* values, which were adjusted using false discovery rate correction[66] with a cut-off of 0.05.

Phylogenomic and phylostratigraphic analysis. We used proteomes of 23 species representing key phylogenetic positions in the plant kingdom (Supplementary Table 20) to construct orthologous gene groups (orthogroups) with Orthofinder (v.2.4.0)[29]. A species tree, of the 23 individuals, based on a recent phylogeny including more than 1,000 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 (v.3.0)[68]. For more details, see the Supplementary Methods.

TAI calculation. TAI is the weighted mean of phylogenetic ranks (phylostrata) and we calculated it for every sample[63]. We used the species tree from the One Thousand Plant Transcriptomes Initiative, 2019 (ref. [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 (v.3.0)[68] 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 (v.3.0)[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 TPM value of the gene 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 v.2.0 (https://www.plabipd.de/portal/web/guest/mercator4/-/wiki/Mercator4/recent_changes). Transcription factors and kinases were predicted using iTAK (v.1.7a)[69]. Additional transcription factors were identified using the online tool PlantTFDB v.5.0 (http://planttfdb.cbi.pku.edu.cn/prediction.php)[70]. For more details, see the Supplementary 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 Mercator4 v.2.0. Briefly, for a group of *m* genes (for example, genes specifically expressed in *Arabidopsis* root), we first counted the number of Mapman bins present in the group, and then evaluated whether 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 the Supplementary Methods.

Identification of orthogroup expression profiles. To analyse the expression profiles at the phylostrata level, orthogroups were classified as 'organ-specific', 'ubiquitous' or 'not conserved'. Organ-specific orthogroups are orthogroups containing organ-specific genes and can be subclassified according to the organ (flower-, female-, male-, seed-, spore-, leaf-, apical meristem-, stem-, root meristem-, root-specific). Ubiquitous are orthogroups that are expressed in different organs for each species; that is, they do not show an 'organ-specific' expression profile. Not conserved are orthogroups that have different organ-specific expression profiles in different species (for example, 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 analysed orthogroups that fulfilled the following criteria: (1) species-specific with transcriptome data or (2) contained at least two species with transcriptome data. To identify organ-specific orthogroups, we required (3) >50% of genes of the orthogroup should support the expression profile and (4) ≥50% of the species with transcriptome data present in the node should support the expression profile.

Gene enrichment analysis per phylostrata. 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 and root), we counted the number of genes present in each node of

the species tree and then evaluated whether 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 node by calculating the difference between 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 and root) and ubiquitous expression profile, 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 were computed using the approach described in ref. [71] with ETE (v.3.0)[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 orthologues 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 by 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, for each expression profile (ubiquitous, flower, female, male, seeds, spore, leaf, stem, apical meristem, root meristem and root), a graphical representation of the different clades showing the nodes with an intensity of colour proportional to the normalized values of gains and losses was plotted using ETE (v.3.0)[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 (v.2.1.2)[73]. For more details, see the Supplementary Methods.

Identification of gamete-specific transcription profiles by clustering analysis. We analysed the male and female organ-specific genes and their different subsamples (Supplementary Table 1) to identify transcription profiles by clustering analysis. For the clustering analysis, we only included species with at least two subsamples (*A. trichopoda*, *O. sativa*, *Z. mays*, *A. thaliana* and *S. 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 value in each subsample was calculated, and the average TPM values were scaled by dividing by 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. Co-expression 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 (for example, cluster 1 and cluster 2) by the expected value (the product of the number of genes in cluster 1 × the number of genes in cluster 2). These values were used to perform a Pearson's correlation analysis and the results were presented in heatmaps. The networks present in the male clusters were visualized using Cytoscape (v.3.8.0)[74]. The network files are available at https://evorepro.sbs.ntu.edu.sg/species/.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

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.

Received: 30 October 2020; Accepted: 2 June 2021;
Published online: 12 July 2021

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**, 20150490 (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 O_2 and CO_2. *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 CO_2 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 of *Arabidopsis* 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. in *Annual Plant Reviews* (eds Roberts, J. A. et al.) 1–50 (John Wiley & Sons, 2018).
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**, 20160496 (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 d*N*/d*S* 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*. *Front. Plant Sci.* **12**, 635990 (2021).
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 generations in *Arabidopsis*. *eLife* **10**, e61894 (2021).
51. Rao, X. & Dixon, R. A. Co-expression networks for plant biology: why and how. *Acta Biochim. Biophys. Sin. (Shanghai)* **51**, 981–988 (2019).
52. Borges, F. et al. Comparative transcriptomics of *Arabidopsis* sperm cells. *Plant Physiol.* **148**, 1168–1181 (2008).
53. Becker, J. D., Takeda, S., Borges, F., Dolan, L. & Feijó, J. A. Transcriptional profiling of *Arabidopsis* root hairs and pollen defines an apical cell growth signature. *BMC Plant Biol.* **14**, 197 (2014).
54. von Besser, K., Frank, A. C., Johnson, M. A. & Preuss, D. *Arabidopsis HAP2* (*GCS1*) is a sperm-specific gene required for pollen tube guidance and fertilization. *Development* **133**, 4761–4769 (2006).
55. Proost, S. & Mutwil, M. CoNekT: an open-source framework for comparative genomic and transcriptomic network analyses. *Nucleic Acids Res.* **46**, W133–W140 (2018).
56. Boisson-Dernier, A. et al. Disruption of the pollen-expressed FERONIA homologs ANXUR1 and ANXUR2 triggers pollen tube discharge. *Development* **136**, 3279–3288 (2009).
57. Zhu, L. et al. The *Arabidopsis* CrRLK1L protein kinases BUPS1 and BUPS2 are required for normal growth of pollen tubes in the pistil. *Plant J.* **95**, 474–486 (2018).
58. Alves-Ferreira, M. et al. Global expression profiling applied to the analysis of *Arabidopsis* stamen development. *Plant Physiol.* **145**, 747–762 (2007).
59. Gupta, R., Ting, J. T. L., Sokolov, L. N., Johnson, S. A. & Luan, S. A tumor suppressor homolog, AtPTEN1, is essential for pollen development in *Arabidopsis*. *Plant Cell* **14**, 2495–2507 (2002).
60. Zhou, Z. et al. *Arabidopsis* RIC1 severs actin filaments at the apex to regulate pollen tube growth. *Plant Cell* **27**, 1140–1161 (2015).
61. Liang, Y. et al. MYB97, MYB101 and MYB120 function as male factors that control pollen tube–synergid interaction in *Arabidopsis thaliana* fertilization. *PLoS Genet.* **9**, e1003933 (2013).
62. Szövényi, P., Waller, M. & Kirbis, A. Evolution of the plant body plan. *Curr. Top. Dev. Biol.* **131**, 1–34 (2019).
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. *J. R. Stat. Soc. Ser. B (Methodol.)* **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–D560 (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).

Acknowledgements

I.J. is supported by Singaporean Ministry of Education grant MOE2018-T2-2-053, while M.M. is supported by a NTU Start-Up Grant. ERA-CAPS EVO-REPRO I2163 and FWF grant P30802 were awarded to F.B.; FCT ERA-CAPS-0001-2014 and PTDC-BIA-FBT-28484-2017 to J.D.B.; and ERA-CAPS EVO-REPRO DR 334/12-1 to S.S. and T.D. D. Hackenberg was supported by ERA-CAPS UK Biotechnology and Biological Research Council grant BB/N005090 awarded to D.T.; 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. Honys were supported by the Czech Ministry of Education, Youth and Sport (LTC18034 and LTAIN19030) through the European Regional Development Fund-Project "Centre for Experimental Plant Biology" number CZ.02.1.01/0.0/0.0/16_019/0000738. The Genomics Unit of Instituto Gulbenkian de Ciência was partially supported by the ONEIDA Project (LISBOA-01-0145-FEDER-016417) co-funded by FEEI–'Fundos Europeus Estruturais e de Investimento' from the '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 the FCT (PD/BD/114362/2016) under the Plants for Life PhD Program. J.D.B. received salary support from the FCT through an 'Investigador FCT' position. M.J. and J.G. were supported by a US National Science Foundation grant (IOS-1540019). Help with sample generation was provided by L. Z. Drábková and D. Reňák. *Marchantia* growth was performed by the Plant Sciences Facility at the Vienna BioCenter Core Facilities GmbH (VBCF), member of the Vienna BioCenter (VBC), Austria. M. Weigend, C. Löhne and B. Reinken (Botanical Garden of the University of Bonn, Germany) are acknowledged for providing *A. trichopoda* plant material. D. Shivhare is acknowledged for a preliminary analysis of *Physcomitrium* RNA-seq data. We thank D. Maizels (http://www.scientific-art.com/) for the illustrations in Figs. 1 and 5.

Author contributions

J.D.B. and M.M. conceived and designed the analysis. A.-C.L., M.F.-T., S.G.P., C.S.M., J.G., M.J., I.J., L.S., C.M., D. Honys and D. Hackenberg collected the data. F.B., M.B., S.S., T.D., T.K. and D.T. contributed data or analysis tools. I.J., C.F., S.P., A.-C.L. and M.M. performed the analyses. I.J., J.D.B. and M.M. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-021-00958-2.

Correspondence and requests for materials should be addressed to J.D.B. or M.M.

Peer review information *Nature Plants* thanks Jan de Vries, Dabing Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2021