Custom methods to identify conserved genetic modules applied to novel transcriptomic data from Amborella trichopoda

Ana Rivarola Sena, Amélie Andres-Robin, Aurélie Vialette, Jérémy Just, Alexandra Launay-Avon, Néro N. Borrega, Bertrand Dubreucq, C.P. Scutt

Custom methods to identify conserved genetic modules applied to novel transcriptomic data from *Amborella trichopoda*

Ana C. Rivarola Sena[1*] (anafoxi@gmail.com),

Amélie Andres-Robin[1*] (amelie.andresrobin@free.fr),

Aurelie C. Vialette[1] (aurelie.vialette@ens-lyon.fr),

Jérémy Just[1] (jeremy.just@ens-lyon.fr),

Alexandra Launay-Avon[2,3] (alexandra.launay-avon@inrae.fr),

Néro Borrega[4] (nero.borrega@inrae.fr),

Bertrand Dubreucq[4] (bertrand.dubreucq@inrae.fr), and

Charles P. Scutt[1#] (charlie.scutt@ens-lyon.fr).

***Contributed equally.**

[#]Author for correspondence.

Author

1. Laboratoire Reproduction et Développement des Plantes, Univ. Lyon, ENS de Lyon, UCB Lyon-1, CNRS, INRA, F-69342 Lyon, France
2. Université Paris-Saclay, CNRS, INRAE, Univ Evry, Institute of Plant Sciences Paris-Saclay (IPS2), 91405, Orsay, France
3. Université de Paris, CNRS, INRAE, Institute of Plant Sciences Paris-Saclay (IPS2), 91405, Orsay, France
4. UMR 1318 INRAe-AgroParisTech, Route de Saint Cyr, 78026 Versailles cedex, France

Highlight

We use Laser-Assisted Microdissection to characterize the female floral transcriptome of the early diverging angiosperm *Amborella trichopoda* and present a novel protocol to compare transcriptomic data between widely diverged plant species.

Abstract

We have devised a procedure for the inter-species comparison of transcriptomic data and used this to reconstruct the expression dynamics of major genetic modules that were present at least 149 million years ago in the most recent common ancestor (MRCA) of living angiosperms. We begin by using Laser-Assisted Microdissection to generate novel transcriptomic data from female flower tissues of *Amborella trichopoda,* the likely sister to all other living angiosperms. We then employ a gene-expression clustering method, followed by a custom procedure to compare genetic modules on the basis of gene-orthology between *Amborella* and the molecular-genetic model angiosperm *Arabidopsis thaliana* (Arabidopsis). Using this protocol, we have succeeded in identifying nine major genetic modules that appear to have conserved their expression dynamics from an early stage in angiosperm evolution. The genes of these modules, representing over 5000 orthogroups, include around a third of those known to control female reproductive development in Arabidopsis. Our study constitutes a proof-of-concept for the comparison of transcriptomic data between widely diverged plant species and represents a first step in the large-scale analysis of gene expression dynamics in a macro-evolutionary context.

Keywords: *Amborella trichopoda,* angiosperm, *Arabidopsis thaliana,* carpel, evo-devo, flower, Laser-Assisted Microdissection, ovule, transcriptomics, WGCNA

Introduction

A major branch of evolutionary-developmental biology (evo-devo) uses inter-species comparisons within an established phylogenetic framework to reconstruct the evolution of the gene regulatory networks (GRNs) that control organismal development. In parallel, gene clustering methods have been widely used to define GRNs within individual species, and in some cases to make comparisons of networks between closely related species (Morandin *et al.*, 2016; Muntane *et al.*, 2017; Yu *et al.*, 2020). However, few studies to date have used clustering methods to compare gene expression between highly diverged species, as is often necessary in evo-devo studies that aim to answer macro-evolutionary questions.

The origin of the angiosperm flower, and of its numerous unique organ systems and processes, is one of the major unresolved questions of macro-evolutionary biology. To address this question from a molecular angle, several studies have compared, on a gene-by-gene basis, flower-development regulators from model angiosperms such as *Arabidopsis thaliana* (Arabidopsis) with their orthologs from members of the three earliest-diverging "ANA-grade" angiosperm orders: Amborellales, Nymphaeales and Austrobaileyales (reviewed by Scutt, 2018). Indeed, many of these studies have focused on *Amborella trichopoda* (hereafter referred to as *Amborella)*, the only living representative of Amborellales and probable sister to all other living flowering plants (Stevens, 2001). *Amborella* is a scrambling dioecious shrub, endemic to the sub-tropical rainforests of New Caledonia. Flowers of female *Amborella* individuals contain spirally arranged organs that make up a perianth of typically 7-8 tepals, surrounding a gynoecium of 5(-6) unfused carpels (Endress & Igersheim, 2000). One or two staminodes (sterile, stamen-like organs) may also be present, inserted between the tepals and carpels. The undifferentiated perianth of the *Amborella* flower is a probable pleisiomorphic feature of angiosperms, while its gynoecium also contains a number of likely pleisiomorphies, including the presence of a compitum, generated through physical contact between the stigmatic surfaces of adjacent carpels, which facilitates the exchange of pollen tubes and thus increases the efficiency of fertilization. *Amborella* carpels also show the probable pleisiomorphies of an ascidiate (vase-like) shape; ridged, multicellular protrusions on the stigmatic surface; a secretion-filled apical canal for pollen tube growth; and a single, pendent, bitegmic ovule. In contrast to *Amborella,* the most recent common ancestor (MRCA)

of living flowering plants, estimated to have lived at least 149 million years ago (MYA; Barba-Montoya *et al.*, 2018), is likely to have possessed bisexual flowers (Sauquet *et al.*, 2017). Whether the floral phyllotaxy of this ancestral species was spiral (as in *Amborella*), whorled, or a mixture of both types, remains a topic of hot debate (Sauquet *et al.*, 2017, 2018; Sokoloff *et al.*, 2018; De-Paula *et al.*, 2018; Ruempler & Theissen, 2019).

The *Amborella* nuclear genome measures approximately 810 Mb (*Amborella* Genome Project, 2013). Its detailed analysis reveals clear traces of the At-ε whole genome duplication (WGD), which is believed to have preceded the radiation of living flowering plants (Jiao *et al.*, 2011), but provides no evidence of any subsequent WGD events along the *Amborella* lineage. The availability of flower tissue-specific transcriptomic resources in *Amborella* has until now been limited, though a detailed study of gene expression in the *Amborella* egg apparatus was recently published, which also included data from mature tepals, leaves and roots (Flores-Tornero *et al.*, 2019). Transcriptomic datasets from young, male and female *Amborella* flower buds have also recently been generated (Käfer *et al.,* 2022).

Here, we contribute the first detailed transcriptomic study of *Amborella* female flower development, performed using Laser Assisted Microdissection (LAM). We use the data obtained to address the two following questions. Firstly, can transcriptomic data be used to identify entire genetic modules whose expression in plant tissues has been conserved over deep evolutionary time? Secondly, to what extent can transcriptomic datasets, generated in independent studies, be usefully compared in plant evo-devo analyses? As a pilot study, we have applied Weighted Gene Co-expression Network Analysis (WGCNA; Langfelder & Horvath, 2008) to our own and published transcriptomic datasets from *Amborella,* and to comparable published datasets from the model angiosperm *Arabidopsis thaliana*. We have succeeded in identifying nine pairs of genetic modules that show both highly significant intersections in orthogroup-content and congruent expression dynamics between *Amborella* and Arabidopsis, strongly suggesting their conservation since the MRCA of living angiosperms.

Materials and methods

Plant material

Seedlings of *Amborella trichopoda* (Bail.) were generously provided by Gildas Gâteblé and Bruno Fogliani (IAC, St Michel, New Caledonia) and grown to maturity in Lyon in a greenhouse under conditions of ~70% relative humidity, 18-30°C, and natural daylight attenuated by artificial shade-covering. Flower buds for LAM procedures were harvested from five-year-old plants during peak flowering season in the autumn. These materials were used to initiate the workflow indicated in Fig. 1.

Laser-Assisted Microdissection

Amborella female flower buds were fixed, embedded, sectioned and subjected to LAM procedures on a PALM Micro-Beam system (Zeiss) as described by Sakai *et al.* (2018). Approximately ten tissue sections were combined for each sample. RNA was extracted from these using an Arcturus Pico RNA Extraction Kit (Thermo Fisher Scientific). RNA integrity was assessed using a 2100 BioAnalyzer (Agilent Technologies) and samples showing RNA Integrity Numbers (RINs) of >7.0 were processed for library construction.

Library construction, sequencing, read mapping and data normalization

cDNA syntheses were performed from *Amborella* RNA samples using the SMARTer Ultra Low Input RNA Kit for Sequencing (Clontech) and libraries were prepared according to Illumina DNA Sample Preparation kit instructions. Libraries were checked for quality using a 2100 Bioanalyzer and pooled in groups of five using a barcoding system to be processed in three runs of Hi-seq2000 (Illumina) paired-end high-throughput sequencing. Subsequent steps were performed on the data obtained from these sequencing runs, on datasets from *Amborella* leaves, roots and mature tepals (Flores-Tornero *et al.*, 2019), and on datasets from young *Amborella* flower buds (Käfer *et al.,* 2022). Read quality was assessed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adapters, sequencing

artifacts and low quality regions of reads were trimmed using Cutadapt v1.16 (Martin, 2011). SNPs between the reference genome and reads were identified using STAR mapper v2.7.3a (Dobin *et al.*, 2013) and reads were then mapped to the *Amborella* V1 complete genome sequence (*Amborella* Genome Project, 2013) using HISAT2 v2.2.1 (Kim *et al.*, 2019) and the previously identified SNPs. Splicing sites were first identified by mapping all reads, and then used for per-sample mapping. Raw reads-per-gene were counted using FeatureCounts v1.5.1 (Liao *et al.*, 2014). Subsequent steps were performed on the *Amborella* read-count data obtained as described above, and on published read-count data obtained in duplicate from Arabidopsis flower and vegetative tissues (Klepikova *et al.*, 2016). Arabidopsis datasets were chosen to be broadly equivalent to the *Amborella* datasets included in our analyses, including leaf, root and flower tissues, with particular emphasis on female reproductive organs. Data from both species were normalized using DEseq2 (Love *et al.*, 2014) and subjected to batch-effect removal using limma (Ritchie *et al.*, 2015). The R codes used for these data normalization steps are given in Supplementary Protocol S1.

Gene expression clustering

Gene clustering was performed in parallel on *Amborella* and Arabidopsis datasets using the WGCNA package (Langfelder & Horvath, 2008). All datasets were used in clustering analyses, except for five male and four female *Amborella* flower-bud datasets, and one dataset from each of two *Amborella* laser-microdissected sample-types, as these did not group with the majority of the corresponding replicates during the WGCNA data-cleaning procedures (Supplementary Figs S1-S2). Clustering was performed using the WGCNA automated method using default settings, except that a signed (instead of the default unsigned) network topology was chosen, and the maximum block size parameter was set to 30 000, i.e. greater than the number of genes under analysis in each species. Gene expression modules of a minimum size of 30 genes were generated from dendrograms at a cut-height of 0.35. The R-codes used for these WGCNA data cleaning and clustering steps are given in Supplementary Protocol S1.

Inter-species comparison of gene-expression modules

A global comparison of gene orthology between the *Amborella* and Arabidopsis genomes (Supplementary Table S1 at *JXB* online) was downloaded from BIOMART in PHYTOZOME (https://PHYTOZOME.jgi.doe.gov/pz/portal.html). This comparison classes genes from each species into orthogroups, each of which has a unique identifier and is putatively derived from a single ancestral gene present in the MRCA of *Amborella* and Arabidopsis (which is also the probable MRCA of all living angiosperms).

The *Amborella* and Arabidopsis modules generated using WGCNA were processed into return-separated lists of gene-identifiers using Perl-Code-1 (Supplementary Protocol S1), and these were then converted to orthogroup identifiers using Perl-Code-2 (Supplementary Protocol S1), which also takes as input a mapping file in .csv format, derived from the gene-orthology relationships obtained using PHYTOZOME (Supplementary Tab. S1). Orthogroup members in *Amborella* and Arabidopsis modules were then compared to provide input for a series of 234 hypergeometric tests to determine the significance of all possible inter-species module intersections (for 13 *Amborella* x 18 Arabidopsis modules). Lists of orthogroups, generated using Perl-Code-2, were first compared between species using Perl-Code-3 (SI Methods), which lists and counts the *Amborella* orthogroup members present in each module intersection (= sample successes, *q*). Perl-Code-4 (SI Methods) was then used to obtain the number of successes obtained when the entire population of *Amborella* orthogroups is compared to each Arabidopsis module in turn (= population successes, *m*), and also the total number of *Amborella* orthogroup members (*t*), permitting the calculation of population failures (*n* = *t*-*m*). Perl-Code-5 (SI Methods) was used to count the number of *Amborella* orthogroup member in each module (= number of trails, *k*). Upper-tail p-values for each pairwise module comparison were then calculated by the hypergeometric test using the R code:

>phyper(q-1, m, n, k, lower-tail=FALSE);

The *Amborella* and Arabidopsis orthologs of the nine most significant inter-species module intersections were then compiled in lists, together with their orthogroup IDs, using Perl-Code-6 (SI Methods). Gene annotation data were added to these lists for

the first-listed Arabidopsis gene in each orthogroup from V9 of the Arabidopsis genome, obtained from TAIR (https://www.arabidopsis.org/). GO-term gene-function analyses were then performed on these lists using PANTHER, accessed via TAIR. As a complement to GO-term analyses, gene annotations were also searched semantically using a custom list of search terms (Perl-Code-7).

The full list of Arabidopsis genes from the nine most highly conserved module intersections was also searched using Perl-Code-8 (SI Methods) for a list of 92 genes of particular interest, known to control aspects of female reproductive development.

Additional microscopy

Images of female and male mature *Amborella* flowers were obtained using Keyence VHX-900F and Leica MZ12 binocular microscopes, respectively. Scanning electron microscope (SEM) images of whole flower buds were obtained using a HIROX 400 environmental SEM at a tube voltage of 10 kV and a stage temperature of -20°C.

Results

Laser-Assisted Microdissection proves an efficient method to define the floral transcriptome of the early-diverging angiosperm *Amborella trichopoda*

We applied, to the dioecious early-diverging angiosperm *Amborella trichopoda* (Fig. 2A-C), LAM and RNA-seq procedures to generate transcriptomic datasets from two stages of female flower development. The developmental stages sampled corresponded to Stage 5-6, and late Stage 7 as defined by Buzgo *et al.* (2004). At Stage 5-6, the young tepals are expanding laterally and the carpels are developing from a solid primordium into a hollow structure that will later enclose the ovule. At this stage, young carpels are clearly distinguishable from any staminodes which may be present. Ovule initiation occurs at early Stage 7, and by late Stage 7, ovule

development is well underway, while distinct regions of the carpel corresponding to the ovary wall and stigma/carpel apex are also clearly apparent.

We sampled two tissues from Stage 5-6, corresponding to young tepals and carpels (Fig. 2D-F). From late Stage 7, we sampled three tissues corresponding to the ovule, ovary wall and stigma/carpel apex (Fig. 2G-I). All samples were obtained in triplicate, making a total of 15 samples for library construction and paired-end read sequencing. Between 89% and 96% of reads generated from the 15 samples could be mapped to the *Amborella* genome sequence (*Amborella* Genome Project, 2013), giving an average of 2.62 x 10^7 mapped reads per sample (Supplementary Tab. S2). These data, which constitute the first detailed transcriptomic study of *Amborella* female flower buds, add considerably to the transcriptomic resources available from this species, the probable sister to all other living angiosperms.

Comparisons between *Amborella* and Arabidopsis identify nine major gene-expression modules conserved since the MRCA of living flowering plants

Gene-clustering analyses using WGCNA from normalized read counts (Supplementary Tabs S3 and S4) yielded totals of 19 and 16 expression modules (AtrME1-13 and AthME1-19) from *Amborella* and Arabidopsis transcriptomic datasets, respectively (Supplementary Figs S3 and S4, and Tab. S5). We compared the orthogroup-content of these modules pairwise between species (Supplementary Tab. S6) and tested the significance of all possible module intersections using the hypergeometric test (Supplementary Tab. S7). The nine most significant of these intersections, which we term Conserved Modules A to I (CM-A to -I), contain between 205 and 1324 orthogroups, with p-values ranging from 5.97x10^{-13} to 1.50x10^{-89} (Tab.1).

We then used the WGCNA package to obtain eigengene values that characterized the expression of the 13 *Amborella* and 18 Arabidopsis modules identified (Supplementary Tab. S8). Module eigengene values are defined as the first right-singular vector of the standardized module expression data (Langfelder and Horvath, 2007), and can take positive or negative values representing highest-to-lowest relative expression levels in each experimental sample or replicate. Effectively, the

set of eigengene values associated with a given module describes the expression of the theoretical gene which perfectly represents the expression dynamics of that module. We averaged module eigengene values over replicates, and in some cases over closely similar samples (see Supplementary Tab. S8), divided these into categories of high (ME>0.1) medium (ME=0-0.1) and low (ME<0) relative expression, and combined these data with the results from the hypergeometric test (Tab. 1, Supplementary Tab. S7). The cut-off values of ME=0 and ME=0.1 were fixed arbitrarily to divide the tables of ME values obtained (Supplementary Tab. S8) into three categories of similar overall sizes, such that high-, medium- and low-expression conditions were visible across the range of modules identified and biological tissues/stages sampled. We then noted plant tissues/developmental stages in which eigengene values correlated between the most highly significant intersecting modules (i.e. in CM-A to -I), suggesting the expression profiles of these modules to have been conserved since the radiation of living angiosperms.

According to this analysis, the results of which are shown in Fig. 3, CM-A, -G and -H were all highly upregulated in leaf tissues. These modules showed, however, differences in other tissues as CM-A was highly upregulated in young tepals and moderately upregulated in the ovary wall, whereas CM-G was moderately upregulated in roots and mature tepals, and CM-H was highly upregulated in mature tepals. CM-B showed a very specific expression pattern as it was highly upregulated in roots and low-expressed in all other tissues. CM-C, -D and -I all showed peaks of expression in female reproductive tissues. Accordingly, CM-C was highly expressed in young carpel and ovules and moderately expressed in the ovary wall, while CM-D showed a more specific upregulation in ovules, though was also moderately expressed in roots. CM-I, showed the clearest upregulation in female reproductive tissues, this module being highly expressed in young carpels, ovules and the ovary wall, and moderately expressed in stigmatic tissues. CM-E and –F showed less clearly tissue-specific patterns than most of the remaining modules, CM-E being moderately upregulated in leaves and roots and low-expressed elsewhere, while CM-F showed an unresolved expression level in leaves, and only moderate upregulation in roots and mature tepals.

Gene-ontogeny (GO) and semantic analyses reveal tissues-specific roles for expression modules in the MRCA of living angiosperms

We compiled lists of the Arabidopsis and *Amborella* genes in each highly significant module intersection, together with their orthogroup identifiers (Supplementary Tab. S9). We then performed GO-term enrichment analyses on the Arabidopsis genes from these conserved modules, the results of which are shown in Supplementary Tab. S10. Lists of tissue-/process-specific GO-terms were extracted from these data by searching using the key-words listed in Supplementary Tab. S10. The results of this analysis, shown in Fig. 4, indicate clear tissue-related specializations for seven of the nine ancient modules identified, in good agreement with the tissues in which these modules were putatively upregulated in the MRCA of living angiosperms (Fig. 3). Accordingly, over-represented genes in CM-A and –H appear to function principally in photosynthesis, in agreement with the strong upregulation of these two modules in leaf tissues. CM-C, and -D appear to make the most significant contribution to reproductive processes, while these two modules, together with CM-I, also show the greatest levels of enrichment for genes involved in reproductive development. CM-E is also moderately enriched in genes that function in reproduction and reproductive development. CM-G and –I appear to play major roles in shoot and leaf development, with CM-G additionally enriched in genes involved in root development, in close agreement with its expression profile (from Fig. 3).

As CM-B, which was highly upregulated in roots (Fig. 3), and CM-F, which showed a less specific expression profile than most of the other conserved modules identified (Fig.3), did not show any clear tissue-related functions in our GO-term analyses (Fig. 4), we directly searched the annotation data of the *Arabidopsis* genes from these modules (Supplemental Table 10, Column E) using a list of tissue-/process-related terms (Fig. 5). This custom semantic analysis showed a very clear relationship between CM-B and root-related processes, in agreement with CM-B's very clear expression profile. Semantic analysis of genes from CM-F also revealed likely functions in roots and petals, in agreement with the moderate upregulation of this module in both perianth and root tissues. For the remaining seven conserved modules, semantic analyses (Fig. 5) confirmed the functional insights obtained using GO-term analyses (Fig. 4), and were also in close agreement with the reconstructed ancestral expression patterns presented in Fig. 3. However, semantic analyses (Fig.

5) tended to reveal broader functions of conserved modules than did GO-term analyses (Fig. 4). For example, GO-term analyses showed CM-H to be very highly enriched in genes related to photosynthesis, while semantic searching of annotation data revealed moderately high occurrences of terms suggesting functions for this module in embryos, seeds and petals, as well as in leaves, cotyledons and guard cells.

Key regulators of female reproductive development occur in five ancient conserved angiosperm modules

The novel transcriptome analyses performed in this work were focused particularly on female flower tissues, and so we verified the positions, among the nine ancient conserved modules identified, of genes known to control female reproductive development. We first compiled from published sources (Colombo *et al.*, 2008; Kelley *et al.*, 2009; Ferrandiz *et al.*, 2010; Drews *et al.*, 2011; Pinto *et al.*, 2019; Reyes-Olalde & de Folter, 2019) a list of 92 genes involved in Arabidopsis female reproductive development (Supplementary Tab. S11). Sixty-one of these genes proved to be represented in our gene-orthology analysis performed using PHYTOZOME (Supplementary Tab. S1). Of these 61 genes, 22 grouped within the nine most highly conserved modules identified in this work (Tab. 2). Eighteen of these genes occurred in modules CM-C, -D and –I, which all showed upregulation in female reproductive tissues (Fig. 3), while the remaining four genes occurred in the modules CM-G and –H, which showed less strongly tissue-specific profiles. Thus, the data presented here suggest that at least a third of a sample of key female developmental regulators in Arabidopsis have conserved major elements of their expression patterns and functions since the MRCA of living flowering plants.

Discussion

A proof-of-concept for the reconstruction of ancient gene-expression modules in plants

In this work, we show that it is possible to meaningfully compare the results of gene-expression clustering analyses performed independently between plant species that diverged at least 149 MYA (Barba-Montoya *et al.*, 2018). By comparing expression modules on the basis of gene orthology, we have been able to identify nine highly significant intersections between modules whose eigengenes show largely congruent expression profiles between the species under comparison. Accordingly, the members of the >5000 orthogroups (Tab. 1) contained in these nine intersections can be considered to have conserved major aspects of their expression profiles in both the *Amborella* and Arabidopsis lineages since the MRCA of living angiosperms. These genes include around a third of a sample of 61 genes known to be involved in the control of female reproductive development in Arabidopsis (Tab. 2, Supplementary Tab. S11). These results represent an initial proof-of-concept, giving a positive answer to the first of the questions we asked at the outset of this work concerning the feasibility of comparing gene-expression clustering data over vast phylogenetic distances.

A combination of GO-term and custom semantic analyses showed clear functional specializations for seven of the conserved modules identified, while CM-E and -F appeared to show somewhat less-specific roles and expression patterns (Figs 3 and 5, Supplementary Tab. S10). We observed that female reproductive processes in flower tissues appear to rely mainly on genes from the three ancient modules CM-C, -D and -I, all of which also contribute significantly to the development of vegetative aerial tissues (Figs 4 and 5). These results underline the nature of plant reproductive organs, whose development is regulated in part by genetic pathways that control the development of all aerial lateral organs, and in part by pathways that are specific to reproductive development.

For example, all aerial lateral organs in plants, including leaves and floral organs, are believed to be based on a common ground-plan, and their development accordingly involves a common set of mechanisms for the definition of abaxial-adaxial polarity and lateral organ expansion. These mechanisms involve several genes, including

KANADI1 (KAN; Eshed *et al.*, 2001) and *AUXIN RESPONSE FACTOR4 (ARF4;* Pekker *et al.*, 2005), which our study shows to be present in CM-H, indicated by our analyses to be expressed (Fig. 3) and to function in (Fig. 5) both vegetative and reproductive tissues. However, lateral organ expansion in the carpel and outer ovule integument also involve the tissue-specific YABBY genes *CRABS CLAW (CRC)* and *INNER-NO-OUTER (INO),* respectively (reviewed by Ferrandiz *et al.*, 2010). These two genes were accordingly found in our study in the conserved module CM-I, which is highly expressed in female reproductive tissues, and contains numerous other reproductive regulators such as *AGAMOUS (AG),* which directly controls carpel and stamen identity. *CRC, INO* and *AG* are among the genes whose expression patterns have previously been studied in both Arabidopsis and *Amborella* using RNA *in-situ* hybridization and other methods (Fourquin *et al.*, 2005; Kim *et al.*, 2005; Arnault *et al.*, 2018). The results of these published studies are in agreement with the conclusion from the present work that *CRC, INO* and *AG,* along with the genes of the other 353 orthogroups in CM-I (Tab. 1, Fig. 3, Supplementary Tab. S6), have been highly expressed in female reproductive tissues since the MRCA of living angiosperms.

Analyses of genome structure indicate the *Amborella* genome to have undergone no whole genome duplication since the At-ε event, which occurred before the radiation of living angiosperms (Jiao *et al.*, 2011). In the Arabidopsis genome, by contrast, there is evidence of one genome triplication (At-γ) and two genome duplications (At-α and At-β) that have taken place more recently. These multiple, large-scale duplication events in the Arabidopsis lineage inevitably render gene-orthology relationships with *Amborella* more complex than with a much more closely related species, such as another member of Brassicaceae. Nonetheless, the use of a genome-wide analysis of orthology relationships has permitted, in the present study, a meaningful comparison of global transcriptome data between species whose lineages diverged at the dawn of the angiosperm clade.

How useful are datasets derived from independent transcriptomic analyses?

Clearly, it is very convenient for practitioners of evo-devo to be able to use published datasets from other workers, in addition to their own novel data. Accordingly, the

second question we asked at the outset of this study concerned the usefulness of comparing datasets from different sources, which by definition were not generated as part of a unified experimental design. The answer to this question, from the results presented here, seems to be that such datasets can be usefully employed, at least if care is taken to match tissues and developmental stages as closely as possible.

Several of the nine most highly significant module-intersections identified in our study showed deviations from a simple 1:1 relationship between species. For example, the Arabidopsis module AthME1 showed highly significant intersections with both AtrME1 and AtrME4 from *Amborella* (Tab. 1). Such deviations from a 1:1 relationship almost certainly result, at least in part, from differences in the tissues/stages included in gene clustering analyses. Indeed, the exact matching of tissues/stages between widely diverged species is often impossible due to problems both of accurately assessing organ homology and of heterochrony (Roux *et al.*, 2015). However, by comparing modules generated through gene-clustering in two species, we would argue that a large part of the dependency of the results obtained on the precise experimental design is likely to be overcome, leading to the accurate identification of conserved modules that were present in the MRCA of the species under comparison.

Future prospects to integrate genetic and morphological modules in plant evo-devo

Modularity is a key concept of evolutionary-developmental biology (evo-devo). However, the definition given of a module may not always correspond closely between evolutionary morphologists and developmental biologists (Mabee, 2006). Evolutionary morphologists tend to think of modules as elements of an organism in which internal connectivity is greater than external connectivity. Such modules, which might for example correspond to a particular organ system or tissue, may be able to undergo evolutionary change without substantial change to the rest of the organism. Developmental biologists, by contrast, think increasingly in terms of GRNs, in which a module might represent a particular set of genes that can be regulated together to produce a given developmental effect, and which might

contribute to the formation of several different organs or tissues in different parts of the organism.

Results from the present study underline the lack, in many cases, of a simple correspondence between molecular and morphological modules. For example, the development of female reproductive organs such as carpels and ovules clearly integrates genes from several ancient modules, some of which are also highly expressed in vegetative organs (Tabs 1 and 2, Fig. 3).

Although many plant transcriptomes have now been assembled (Leebens-Mack et al., 2019), the availability of large-scale gene-expression datasets that focus on distinct organs, tissues and developmental stages continues to lag considerably behind in all but a few model species. It is likely, however, with techniques such as the LAM method employed in the present work, and also the very promising single-cell RNA-sequencing technologies (e.g. Ryu *et al.*, 2019), that detailed large-scale expression data will soon become available from a much wider range of plant species.

We are only just beginning to probe the complex and layered relationship between molecular and morphological modules in plant development, but this endeavor will clearly be facilitated by the increasing availability of detailed transcriptomic data, generated using such methods as LAM and single-cell technologies. It is consequently becoming increasingly important to devise methods to compare global expression data between species. In the present work, we perform a pilot analysis for two species whose lineages separated at the base of the living angiosperm clade as a first step in the comparison of transcriptomic data with an established phylogenetic context. Further work will be necessary to expand this method to accommodate multiple species within a given phylogenetic framework.

Supplementary data

The following supplementary data are available at *JXB* online.

Supplementary Protocol S1. R code blocks used for data normalization, cleaning and clustering, and custom Perl Codes 1-8 for the inter-species comparison of gene-expression clustering results.

Supplementary Table S1. Gene orthology between *Amborella* and Arabidopsis, generated in PHYTOZOME.

Supplementary Table S2. Read counts for *Amborella* samples.

Supplementary Table S3. Normalized *Amborella* data used for input to WGCNA.

Supplementary Table S4. Normalized Arabidopsis data used for input to WGCNA.

Supplementary Table S5. WGCNA gene expression clustering results.

Supplementary Table S6. Lists of orthogroups returned from PERL-CODE-3 (Supplementary Protocol S1), the numbers of which correspond to sample successes for input into the hypergeometric test.

Supplementary Table S7. Hypergeometric test of all possible module intersections, results ranked by p-value.

Supplementary Table S8. Eigengene values of *Amborella* and Arabidopsis modules.

Supplementary Table S9. Reconstructed lists of orthogroup contents from the nine most significant module intersections.

Supplementary Table S10. Results of nine separate GO-term analyses on the Arabidopsis genes from Conserved Modules A to I.

Supplementary Table S11. Regulators of female reproductive development in Arabidopsis.

Supplementary Figure S1. *Amborella* samples used in WGCNA clustering analyses, following removal of outliers.

Supplementary Figure S2. Arabidopsis samples used in WGCNA clustering analyses.

Supplementary Figure S3. *Amborella* module dendrograms.

Supplementary Figure S4. Arabidopsis module dendrograms.

Acknowledgements

We are grateful to Gildas Gâteblé and Bruno Fogliani (IAC, St Michel, New Caledonia) for generously supplying plant material and to Marc Robinson-Rechavi (University of Lausanne) for advice on the comparison of clustering data. We thank the PSMN (Pôle Scientifique de Modélisation Numérique) of ENS-Lyon for the use of computing resources. This paper is dedicated to the memory of Michael W. Frohlich (Royal Botanic Gardens, Kew), a pioneer in Plant Evo-Devo and an exceptional friend and colleague, who passed away suddenly in December 2021.

Author contributions

ACRS performed transcriptomic analyses and general microscopy, AAR performed LAM work, ACV helped with transcriptomics and figure preparation, JJ supervised transcriptomics, ALA performed library construction and high-throughput sequencing, BD and NB supervised LAM work, and CPS coordinated the project and contributed to bioinformatics.

Conflicts of interest

No conflicts to declare.

Funding

This work was supported by ANR grant “ORANGe” (2013-18) and through core funding from CNRS, INRAe, ENS-Lyon and Université Claude-Bernard Lyon-1 to the RDP Laboratory. ACRS was supported by a BECAL studentship (Paraguay).

Data availability

RNA-seq datasets generated in this work are available from the NCBI Sequence Read Archive (SRA) under BioProject ID PRJNA788337 (http://www.ncbi.nlm.nih.gov/bioproject/788337).

References

***Amborella* Genome Project. 2013**. The *Amborella* Genome and the Evolution of Flowering Plants. *Science* **342**: 1467-+.

Arnault G, Vialette ACM, Andres-Robin A, Fogliani B, Gateble G, Scutt CP. 2018. Evidence for the Extensive Conservation of Mechanisms of Ovule Integument Development Since the Most Recent Common Ancestor of Living Angiosperms. *Frontiers in Plant Science* **9**: 1352.

Barba-Montoya J, dos Reis M, Schneider H, Donoghue PCJ, Yang Z. 2018. Constraining uncertainty in the timescale of angiosperm evolution and the veracity of a Cretaceous Terrestrial Revolution. *New Phytologist* **218**: 819–834.

Buzgo M, Soltis PS, Soltis DE. 2004. Floral developmental morphology of Amborella trichopoda (Amborellaceae). *International Journal of Plant Sciences* **165**: 925–947.

Colombo L, Battaglia R, Kater MM. 2008. Arabidopsis ovule development and its evolutionary conservation. *Trends in Plant Science* **13**: 444–450.

De-Paula OC, Assis LCS, de Craene LPR. 2018. Unbuttoning the Ancestral Flower of Angiosperms. *Trends in Plant Science* **23**: 551–554.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Drews GN, Wang D, Steffen JG, Schumaker KS, Yadegari R. 2011. Identification of genes expressed in the angiosperm female gametophyte. *Journal of Experimental Botany* **62**: 1593–1599.

Endress PK, Igersheim A. 2000. Reproductive structures of the basal angiosperm Amborella trichopoda (Amborellaceae). *International Journal of Plant Sciences* **161**: S237–S248.

Eshed Y, Baum SF, Perea JV, Bowman JL. 2001. Establishment of polarity in lateral organs of plants. *Current Biology* **11**: 1251–1260.

Ferrandiz C, Fourquin C, Prunet N, Scutt CP, Sundberg E, Trehin C, Vialette-Guiraud ACM. 2010. Carpel Development. In: Kader JC, Delseny M, eds. Advances in Botanical Research, Vol 55. 1–73.

Flores-Tornero M, Proost S, Mutwil M, Scutt CP, Dresselhaus T, Sprunck S. 2019. Transcriptomics of manually isolated Amborella trichopoda egg apparatus cells. *Plant Reproduction* **32**: 15–27.

Fourquin C, Vinauger-Douard M, Fogliani B, Dumas C, Scutt CP. 2005. Evidence that CRABS CLAW and TOUSLED have conserved their roles in carpel development since the ancestor of the extant angiosperms. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 4649–4654.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, *et al.* 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**: 97-U113.

Käfer J, Bewick A, Andres-Robin A, Lapetoule G, Harkess A, Caïus J, Fogliani B, Gâteblé G, Ralph P, dePamphilis CW, *et al.* 2022. A derived ZW chromosome system in Amborella trichopoda, representing the sister lineage to all other extant flowering plants. *New Phytologist* **233:** 1636-1642.

Kelley DR, Skinner DJ, Gasser CS. **2009**. Roles of polarity determinants in ovule development. *Plant Journal* **57**: 1054–1064.

Kim S, Koh J, Yoo MJ, Kong HZ, Hu Y, Ma H, Soltis PS, Soltis DE. 2005. Expression of floral MADS-box genes in basal angiosperms: implications for the evolution of floral regulators. *Plant Journal* **43**: 724–744.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. **2019**. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**: 907–915.

Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA. **2016**. A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *Plant Journal* **88**: 1058–1070.

Langfelder P, Horvath S. 2007. Eigengene networks for studying the relationships between co-expression modules. *BMC Systems Biology* **1**:54.

Langfelder P, Horvath S. **2008**. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 559.

Liao Y, Smyth GK, Shi W. **2014**. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923–930.

Love MI, Huber W, Anders S. **2014**. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**: 550.

Mabee PM. **2006**. Integrating evolution and development: The need for bioinformatics in evo-devo. *Bioscience* **56**: 301–309.

Martin M. **2011**. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 3.

Morandin C, Tin MMY, Abril S, Gomez C, Pontieri L, Schiott M, Sundstrom L, Tsuji K, Pedersen JS, Helantera H, *et al.* 2016. Comparative transcriptomics reveals the conserved building blocks involved in parallel evolution of diverse phenotypic traits in ants. *Genome Biology* **17**: 43.

Muntane G, Santpere G, Verendeev A, Seeley WW, Jacobs B, Hopkins WD, Navarro A, Sherwood CC. **2017**. Interhemispheric gene expression differences in

the cerebral cortex of humans and macaque monkeys. *Brain Structure & Function* **222**: 3241–3254.

Pekker I, Alvarez JP, Eshed Y. **2005**. Auxin response factors mediate Arabidopsis organ asymmetry via modulation of KANADI activity. *Plant Cell* **17**: 2899–2910.

Pinto SC, Mendes MA, Coimbra S, Tucker MR. **2019**. Revisiting the Female Germline and Its Expanding Toolbox. *Trends in Plant Science* **24**: 455–467.

Reyes-Olalde J, de Folter S. **2019**. Control of stem cell activity in the carpel margin meristem (CMM) in Arabidopsis. *Plant Reproduction* **32**: 123–136.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. **2015**. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**: e47.

Roux J, Rosikiewicz M, Robinson-Rechavi M. **2015**. What to compare and how: Comparative transcriptomics for Evo-Devo. *Journal of Experimental Zoology Part B-Molecular and Developmental Evolution* **324**: 372–382.

Ruempler F, Theissen G. **2019**. Reconstructing the ancestral flower of extant angiosperms: the 'war of the whorls' is heating up. *Journal of Experimental Botany* **70**: 2615–2622.

Ryu KH, Huang L, Kang HM, Schiefelbein J. **2019**. Single-Cell RNA Sequencing Resolves Molecular Relationships Among Individual Plant Cells. *Plant Physiology* **179**: 1444–1456.

Sakai K, Taconnat L, Borrega N, Yansouni J, Brunaud V, Paysant-Le Roux C, Delannoy E, Magniette M-LM, Lepiniec L, Faure JD, *et al.* 2018. Combining laser-assisted microdissection (LAM) and RNA-seq allows to perform a comprehensive transcriptomic analysis of epidermal cells of Arabidopsis embryo. *Plant Methods* **14**: 10.

Sauquet H, von Balthazar M, Doyle JA, Endress PK, Magallon S, Staedler Y, Schoenenberger J. **2018**. Challenges and questions in reconstructing the ancestral flower of angiosperms: A reply to Sokoloff et al. *American Journal of Botany* **105**: 127–135.

Sauquet H, von Balthazar M, Magallon S, Doyle JA, Endress PK, Bailes EJ, de Morais EB, Bull-Herenu K, Carrive L, Chartier M, *et al.* 2017. The ancestral flower of angiosperms and its early diversification. *Nature Communications* **8**: 16047.

Scutt CP. **2018**. The origin of angiosperms. *Evolutionary developmental biology: a reference guide*: 1–20.

Sokoloff DD, Remizowa MV, Bateman RM, Rudall PJ. **2018**. Was the ancestral angiosperm flower whorled throughout? *American Journal of Botany* **105**: 5–15.

Stevens PF. **2001**. Angiosperm Phylgeny Website. Version 14, July 2017. *http://www.mobot.org/MOBOT*.

Yu Y, Hu H, Doust AN, Kellogg EA. **2020**. Divergent gene expression networks underlie morphological diversity of abscission zones in grasses. *New Phytologist* **225**: 1799–1815.

Figure legends

Figure 1. Workflow summary.

Figure 2. Laser-Assisted Microdissection of *Amborella* female flower buds.

A. Young female plant. **B.** Mature female flower. **C.** Mature male flower (for comparison). **D-E.** Stage 5-6 of female flower development showing (D) SEM of whole flower bud, and (E and F) LS of flower bud before and after removal of carpel tissues by LAM. **G-I.** Late Stage 7 of female flower development showing (G) SEM of whole flower bud, and (H and I) LS of flower bud during sequential removal of ovule, ovary wall and stigmatic tissues by LAM. Key: c = carpel, sd = staminode, st = stamen, t = tepal.

Figure 3. Partially reconstructed expression patterns of nine ancient gene modules in flower, root and leaf tissues of the MRCA of living angiosperms.

Key: le = leaf, mp = mature perianth, ms = mature stamen, ova = ovary wall, ovu = ovule, ro = root, st = stigma/carpel apex, yp = young perianth, ys = young stamen.

Figure 4. Enrichement in organ-/process-related GO-terms for *Arabidopsis* genes in seven of the nine most highly conserved modules, showing clearly distinct functional specializations (coloured boxes).

Full lists of over-represented GO terms, together with the search-terms used to define the sub-lists shown here, are given in Supplementary Tab. S10. The remaining two conserved modules, CM-B and CM-F (Supplementary Tab. S10), showed enrichment only in GO terms not including the organ-/process-specific search terms used. Three statistical parameters are shown for each overrepresented GO-term in each module: the number of genes is underlined, the fold-enrichment value is given in normal text, and the $\log_{10}$ p-value is given in bold. Data are shown only for p-values <0.05 ($-\log_{10}$ p >1.30).

Figure 5. Custom semantic analyses of conserved modules CM-A t CM-I, showing some organ-related specificities that were not apparent from GO-term analyses alone.

Percentages of *Arabidopsis* gene annotations from each conserved module (from Supplementary Tab. S9, Column E) are given that match the search terms shown on the left.

Table 1. The nine most significant module intersections and their expression dynamics in Arabidopsis and *Amborella.*

Conserved module	**Module intersection**	**No. of orthogroups**	**p-value**	**Arabidopsis ME value >0.1**	***Amborella* ME value >0.1**	**Arabidopsis ME value 0-0.1**	***Amborella* ME value 0-0.1**
CM-A	AthME5∩AtrME6	**205**	1.50E-89	Leaf lamina, Young Sepal, Pedicel	Leaf, Young tepal	Petiole, Leaf vein, Mature sepal, Ovary wall	Ovary, Stigma, Young carpel
CM-B	AthME3∩AtrME3	**360**	1.39E-80	Root	Root	Petal	–
CM-C	AthME2∩AtrME2	**1249**	7.32E-69	Young carpel, Mature carpel, Ovule	Young tepal, Ovule, Ovary, Stigma, Young carpel	Root, Flowers 1-3, Flowers 4-11, Flowers 12+, Pedicel, Ovary wall	–
CM-D	AthME2∩AtrME5	**425**	2.00E-59	Young carpel, Mature carpel, Ovule	Male flower, Female flower	Root, Flowers 1-3, Flowers 4-11, Flowers 12+, Pedicel, Ovary wall	Root, Ovule
CM-E	AthME18∩AtrME1	**114**	9.17E-31	Young carpel, Ovule	Leaf, Mature tepal	Root, Petiole, Leaf vein, Mature carpel, Ovary wall	Root, Male flower, Female flower
CM-F	AthME1∩AtrME4	**422**	8.46E-30	Leaf vein	Mature tepal	Root, Petiole, Petal, Mature sepal	Root
CM-G	AthME1∩AtrME1	**1324**	6.37E-17	Leaf vein	Leaf, Mature tepal	Root, Petiole, Petal, Mature sepal	Root, Male flower, Female flower
CM-H	AthME4∩AtrME1	**670**	5.63E-16	Leaf vein, Leaf lamina, Mature sepal	Leaf, Mature tepal	Petiole, Young sepal, Pedicel	Root, Male flower, Female flower
CM-I	AthME7∩AtrME2	**354**	5.97E-13	Young carpel, mature carpel, Ovule, Ovary wall	Young tepal, Ovule, Ovary, Stigma, Young carpel	Stigma, Flowers 1-3, Flowers 4-11, Flowers 12+	–

Table 2. Regulators of Arabidopsis female reproductive organ development in highly conserved gene-expression modules.

Conseerved module	Module intersection	Arabidopsis female regulators
CM-C	AthME2∩AtrME2	*CLF, DMC1, ULT*
CM-D	AthME2∩AtrME5	*AGO4, ANT,FIE, GFA-1, KLU, PHB, YUC4*
CM-G	AthME1∩AtrME1	*WRKY28*
CM-H	AthME4∩AtrME1	*ARF4, KAN1, SEP2*
CM-I	AthME7∩AtrME2	*RBL, HEC1, HEC2, LFY, AG, CRC, PID, INO*

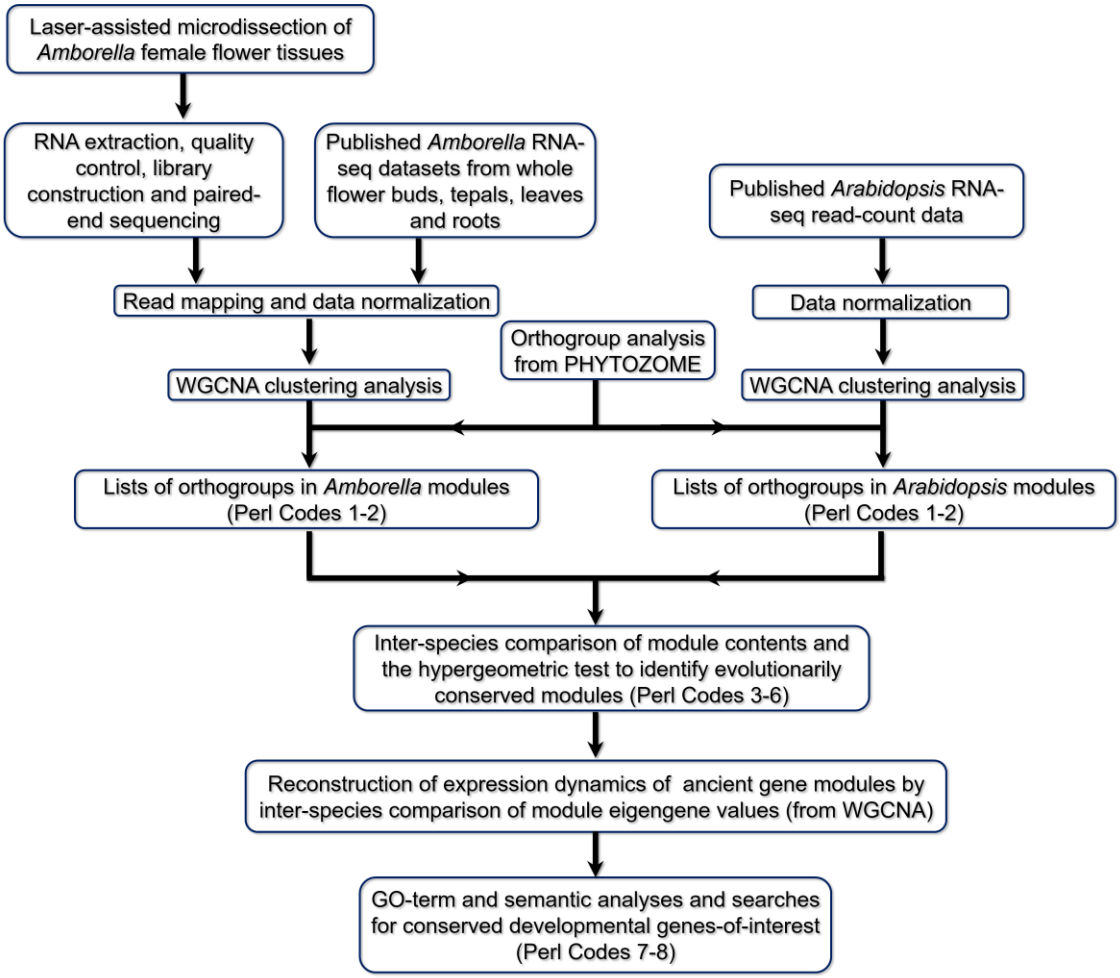

Figure 1. Workflow summary.

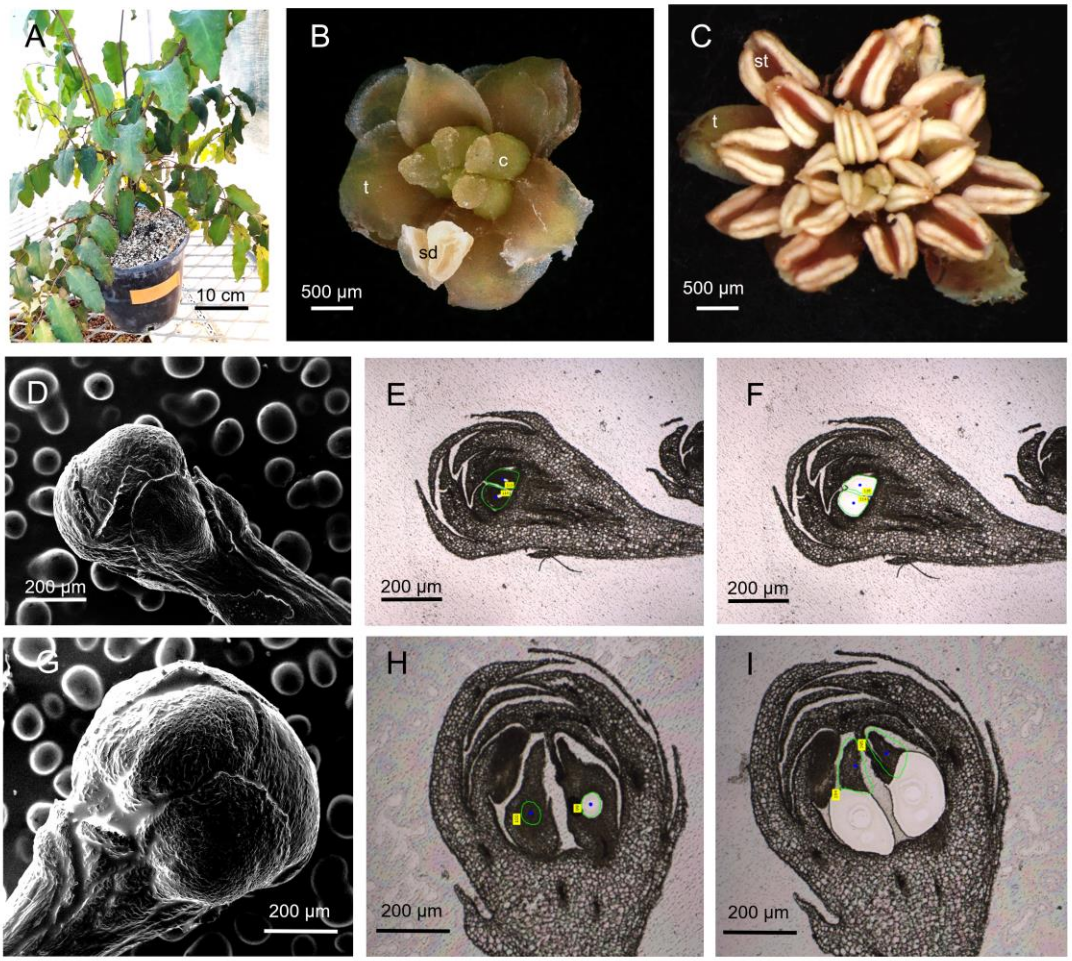

Figure 2. Laser-Assisted Microdissection of *Amborella* female flower buds.
A. Young female plant. **B.** Mature female flower. **C.** Mature male flower (for comparison). **D-E.** Stage 5-6 of female flower development showing (D) SEM of whole flower bud, and (E and F) LS of flower bud before and after removal of carpel tissues by LAM. **G-I.** Late Stage 7 of female flower development showing (G) SEM of whole flower bud, and (H and I) LS of flower bud during sequential removal of ovule, ovary wall and stigmatic tissues by LAM.
Key: c = carpel, sd = staminode, st = stamen, t = tepal.

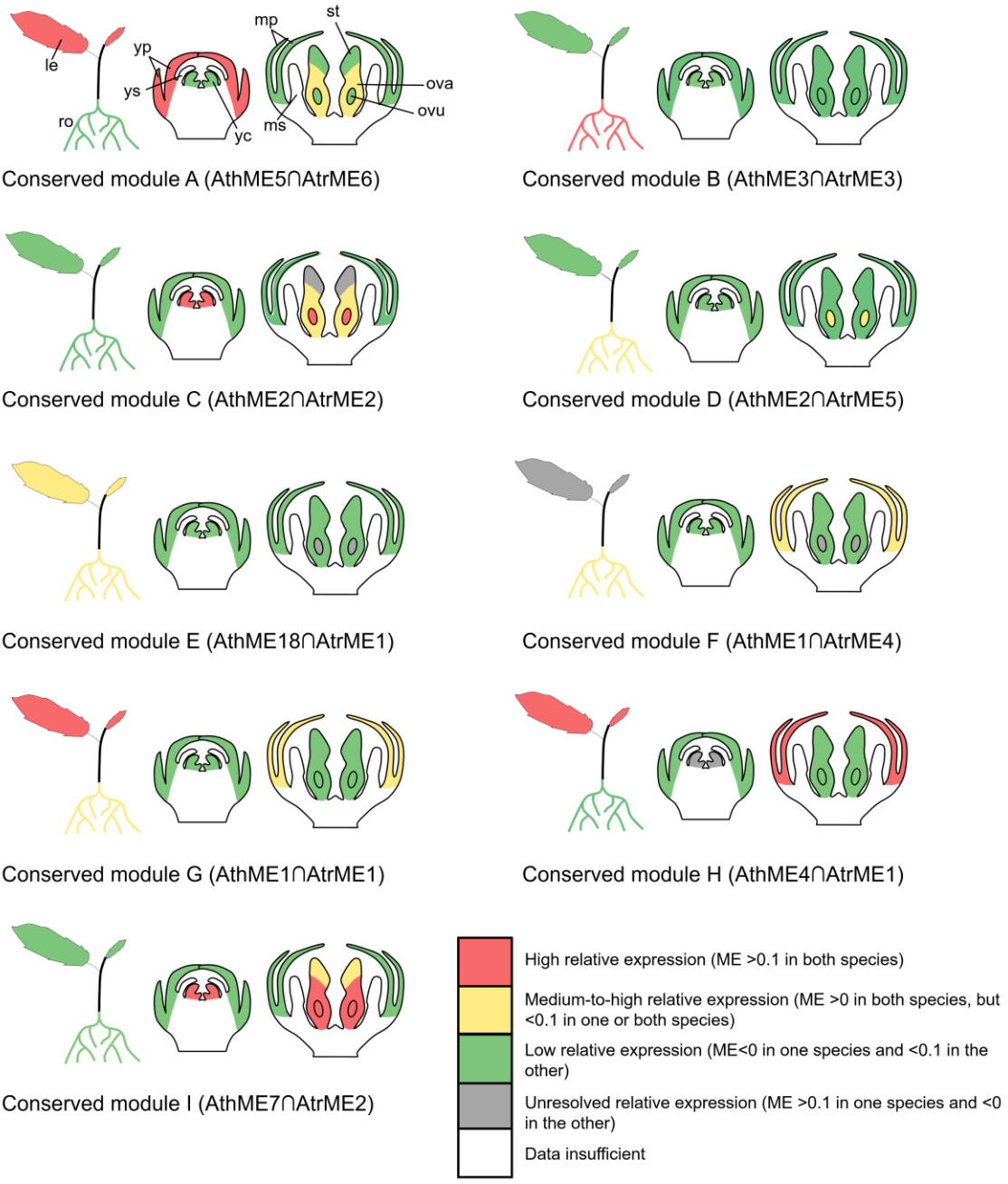

Figure 3. Partially reconstructed expression patterns of nine ancient gene modules in flower, root and leaf tissues of the MRCA of living angiosperms.
Key: le = leaf, mp = mature perianth, ms = mature stamen, ova = ovary wall, ovu = ovule, ro = root, st = stigma/carpel apex, yp = young perianth, ys = young stamen.

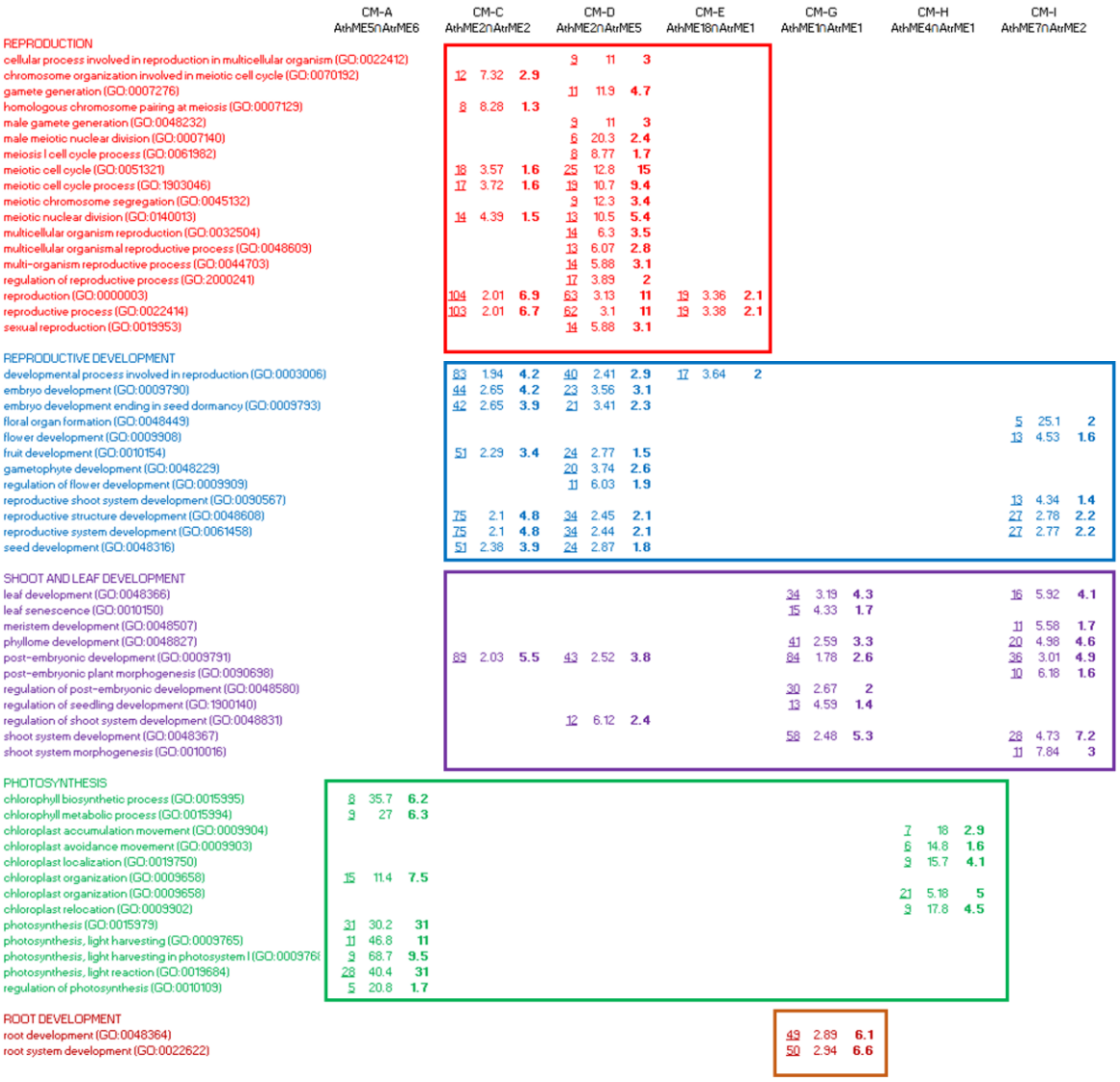

Figure 4. Enrichement in organ-/process-related GO-terms for *Arabidopsis* genes in seven of the nine most highly conserved modules, showing clearly distinct functional specializations (coloured boxes). Full lists of over-represented GO terms, together with the search-terms used to define the sub-lists shown here, are given in Supplementary Tab. S10. The remaining two conserved modules, CM-B and CM-F (Supplementary Tab. S10), showed enrichment only in GO terms not including the organ-/process-specific search terms used. Three statistical parameters are shown for each overrepresented GO-term in each module: the number of genes is underlined, the fold-enrichment value is given in normal text, and the $\log_{10}$ p-value is given in bold. Data are shown only for p-values <0.05 ($-\log_{10}$ p >1.30).

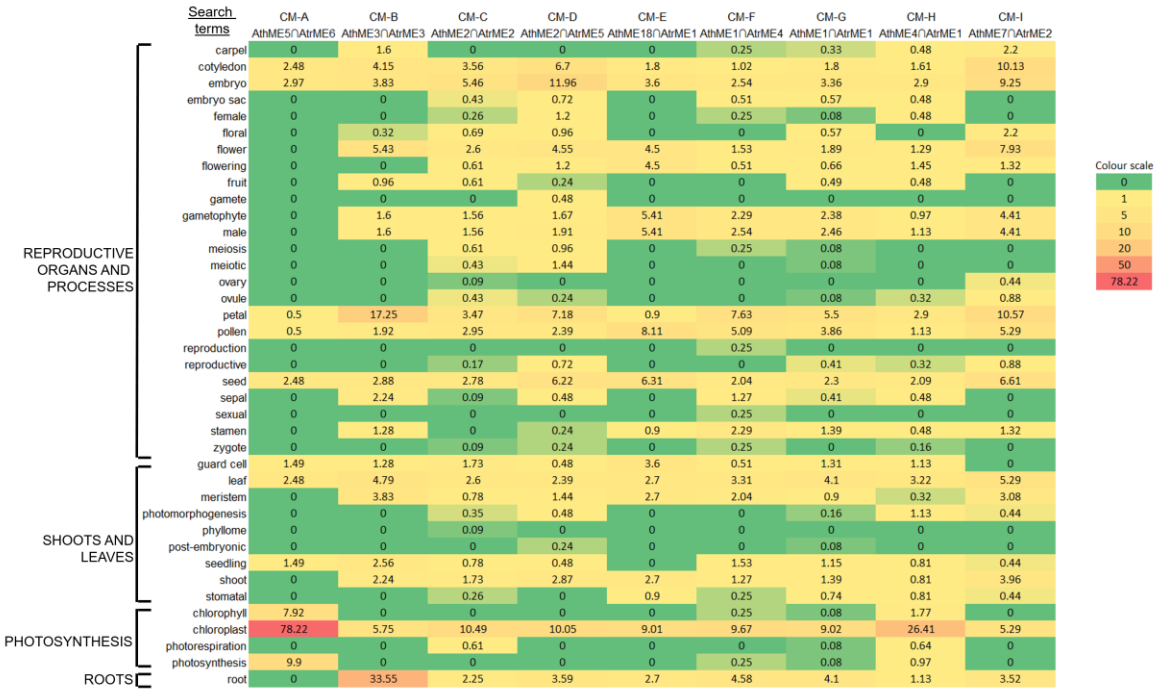

Figure 5. Custom semantic analyses of conserved modules CM-A t CM-I, showing some organ-related specificities that were not apparent from GO-term analyses alone. Percentages of *Arabidopsis* gene annotations from each conserved module (from Supplementary Tab. S9, Column E) are given that match the search terms shown on the left.