
Supplementary information

Insights into angiosperm evolution, floral development and chemical biosynthesis from the *Aristolochia fimbriata* genome

In the format provided by the
authors and unedited

1 **Supplementary Information**

2 **Insights into angiosperm evolution, floral development and chemical biosynthesis** 3 **from the *Aristolochia fimbriata* genome**

4 Liuyu Qin^{1,2#}, Yiheng Hu^{1,2#}, Jinpeng Wang^{1,2,3#}, Xiaoliang Wang^{1,2#}, Ran Zhao^{1#},
5 Hongyan Shan¹, Kunpeng Li^{1,2}, Peng Xu^{1,2}, Hanying Wu¹, Xueqing Yan^{1,2}, Lumei
6 Liu^{1,2}, Xin Yi¹, Stefan Wanke⁴, John E. Bowers^{5,6}, James H. Leebens-Mack⁵, Claude
7 W. dePamphilis⁷, Pamela S. Soltis⁸, Douglas E. Soltis^{8,9}, Hongzhi Kong^{1,2}, Yuannian
8 Jiao^{1,2*}

9 ¹State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany,
10 the Chinese Academy of Sciences, Beijing 100093, China.

11 ²University of Chinese Academy of Sciences, Beijing 100049, China.

12 ³School of Life Sciences, and Center for Genomics and Computational Biology,
13 North China University of Science and Technology, Tangshan, Hebei 063000, China.

14 ⁴Institute of Botany, Dresden University of Technology, Dresden 01062, Germany.

15 ⁵Department of Plant Biology, University of Georgia, Athens, Georgia 30602, USA.

16 ⁶Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia 30602,
17 USA.

18 ⁷Department of Biology and Huck Institutes of the Life Sciences, The Pennsylvania
19 State University, University Park, Pennsylvania 16802, USA.

20 ⁸Florida Museum of Natural History, University of Florida, Gainesville, Florida
21 32611, USA.

22 ⁹Department of Biology, University of Florida, Gainesville, Florida 32611, USA.

23 #These authors contributed equally: Liuyu Qin, Yiheng Hu, Jinpeng Wang, Xiaoliang
24 Wang, Ran Zhao.

25 *e-mail: jiaoyn@ibcas.ac.cn.

26 **This PDF file includes:**

27 Supplementary Notes

28 Supplementary Figs. 1.1-6.9

29 Supplementary References

30 **Addational Supplementary Files for this manuscript include the following:**

31 Supplementary Tables 1.1-6.7

32

33	Supplementary Notes and Figures	3
34	1 Introduction to <i>Aristolochia</i> and the genome sequencing.....	3
35	1.1 Genome survey and genome size estimation.....	3
36	1.2 Oxford Nanopore Technologies (ONT), optical maps, and Hi-C sequencing .	4
37	1.3 Nuclear and chloroplast genome assemblies	5
38	1.4 Genome quality assessments	9
39	2 Genome annotation and gene family classification.....	10
40	2.1 Repeat annotation and TE analyses.....	10
41	2.2 Transcriptome sequencing by Illumina and PacBio Iso-Seq.....	11
42	2.3 Gene prediction and functional annotation.....	13
43	2.4 Gene family classification and comparison.....	14
44	3 Comparative genomics and whole-genome duplication (WGD) events in	
45	magnoliids.....	15
46	3.1 Comparative genomic analyses indicate no WGD in <i>A. fimbriata</i> since the	
47	origin of extant angiosperms	15
48	3.2 Three rounds of WGDs detected in black pepper genome using the reference	
49	genome of <i>A. fimbriata</i>	19
50	3.3 Timing of the previously identified WGDs in magnoliids.....	22
51	3.4 Genome structure comparisons identified genomic rearrangement events	
52	shared by magnoliids and monocots	26
53	4 Phylogenomic investigation of the relationship among magnoliids, monocots, and	
54	eudicots	42
55	4.1 Phylogenomic analyses of strictly and mostly single-copy gene families.....	42
56	4.2 The impact of taxon sampling on the resulting phylogenetic topology	53
57	4.3 Phylogenetic analysis of chloroplast genes.....	55
58	4.4 Codon usage bias and the most plausible phylogenetic topology.....	58
59	4.5 Molecular dating and gene family evolution	63
60	5 High-specialized flower morphology and floral development in <i>A. fimbriata</i>	65
61	5.1 Phylogeny of the MADS-box gene family	66
62	5.2 Alternative splicing of MADS-box genes in <i>A. fimbriata</i>	69
63	5.3 Expression patterns of floral organ identity genes	70
64	5.4 Genetic basis of the floral organ fusion and the bilaterally symmetry of	
65	flowers	72
66	5.5 Floral color and trichome formation genes in <i>A. fimbriata</i>	75
67	6 Metabolic gene clusters in <i>A. fimbriata</i> and the biosynthesis of terpenoid and	
68	aristolochic acids.....	79
69	6.1 Identification of metabolic gene clusters in the <i>A. fimbriata</i> genome	79
70	6.2 Floral scent and terpenes	87
71	6.3 LC-MS-based metabolomic analysis and reconstruction of the aristolochic	
72	acid I biosynthesis pathway.....	89
73	6.4 Identifying the key candidate genes based on phylogenetic classification and	
74	the sequence conservation of key residues	91
75	Supplementary References.....	100

76 **1 Introduction to *Aristolochia* and the genome sequencing**

77 *Aristolochia*, a genus in the magnoliid order Piperales with approximately 450
78 species, has long been famous for its complicated flower morphology, specialized
79 pollination, and enriched alkaloid chemistry¹⁻⁴. The flowers of *Aristolochia* species
80 consist of a monosymmetric, often tubular, dull purple-brown, and insect-trapping
81 perianth, a gynostemium formed by the congenital fusion between stamens and the
82 stigmatic region of the carpels, and elaborated inner epidermis (Fig. 1a and Extended
83 Data Fig. 1)⁵. The flowers often emit unusual flower scent, which together with the
84 peculiar flower morphology, enables virtually all *Aristolochia* species to exhibit
85 deceptive pollination strategies that include attraction, imprisonment, and release of
86 specific pollinators^{3,6-8}. In addition, *Aristolochia* has been widely used in traditional
87 pharmacopeias⁹. However, a class of nitrophenanthrene carboxylic acids called
88 aristolochic acids (AAs) are naturally synthesized in *Aristolochia*; these compounds
89 have been demonstrated to be highly nephrotoxic and carcinogenic to humans¹⁰⁻¹³.
90 Therefore, many herbal medicines are banned or restricted due to their AA content.

91 *Aristolochia fimbriata* is a typical member of *Aristolochia*, it has special features that
92 differ from closely related *Aristolochia* species including the development of fimbriae
93 and papillae. *A. fimbriata* has two genotypes (VL and NV) which possess a number of
94 readily discernible traits (e.g., leaf variegation, perianth shape and color)⁵.
95 Importantly, as we proposed previously⁵, because of its short life cycle, ease of large-
96 scale cultivation, and small genome size, *A. fimbriata* could become a fantastic model
97 system in magnoliids to facilitate comparative studies and functional investigations on
98 the evolution of angiosperms and flowers. Therefore, here we selected *A. fimbriata* as
99 the first *Aristolochia* species to decipher the nuclear genome sequence.

100 **1.1 Genome survey and genome size estimation**

101 **Materials and Methods**

102 The *A. fimbriata* (VL genotype) used for sequencing has been propagated via selfing
103 for approximately 20 years in the greenhouse of Pennsylvania State University, USA.
104 We raised new individuals from cuttings of a single plant in Pennsylvania State
105 University, and cultivated them in the greenhouse of Institute of Botany, Chinese
106 Academy of Sciences in Beijing. Young leaves from one individual were collected for
107 flow cytometry and genome survey sequencing. For flow cytometry experiment,
108 nuclei were released by chopping the young leaves and analyzed with the Moflo XDP
109 Cell Sorter (Beckman-Coulter) according to the manufacturer's instructions. The
110 genome size was then estimated by Summit software using *Arabidopsis thaliana* (125
111 Mb/2C) as a reference.

112 We also estimated the genome size of *A. fimbriata* using a *k*-mer-based approach.
113 DNA was extracted from young leaves of *A. fimbriata* for the genome survey. Paired-
114 end (PE) libraries with an insert size of 500 bp were constructed and sequenced using
115 the Illumina platform (Supplementary Table 1.1). The *k*-mer distribution was then
116 investigated using Jellyfish v2.2.10 with the parameters of ‘-t 10 -C -m 19 -s 4G’¹⁴.
117 Genome size (G) was estimated by $G = k_{\text{num}}/k_{\text{depth}}$, where the k_{num} represents the total
118 number of *k*-mers, and the k_{depth} denotes the *k*-mer depth of the peak frequency of *k*-
119 mer distribution. We also used GenomeScope¹⁵ to estimate the overall genome
120 characteristics (heterozygosity rate, haploid genome length, and unique content
121 length) from the Illumina data using Jellyfish v2.2.10¹⁴ with the parameters of ‘-t 10 -
122 C -m 19 -s 4G’.

123 **Results and Discussion**

124 The flow cytometry experiments found the mean peak fluorescence of *A. thaliana*
125 (R1) at 26.99 and *A. fimbriata* (R2) at 62.51 (Extended Data Fig. 2a); the genome size
126 of *A. fimbriata* was estimated to be approximately 289.50 Mb. We generated ~31 Gb
127 of paired-end clean reads from the Illumina genome survey (Supplementary Table
128 1.1). Using the *k*-mer approach, we estimated the genome size as approximately 251-
129 271 Mb when using different *k*-mer lengths from 17 to 31 (Supplementary Table 1.2).
130 If we set the *k*-mer length to 17, we found the total number of *k*-mer = 20,853,344,487
131 and the *k*-mer depth = 83, and the estimated genome size as 251 Mb (Extended Data
132 Fig. 2b and Supplementary Table 1.2). Therefore, based on estimations from the
133 above two approaches, we infer that the genome size of *A. fimbriata* is about 251-290
134 Mb. In addition, the sequenced individual of *A. fimbriata* has extremely low
135 heterozygosity (~0.07%) (Fig. 1d), following ~20 years of inbreeding in cultivation.

136 **1.2 Oxford Nanopore Technologies (ONT), optical maps, and Hi-C sequencing**

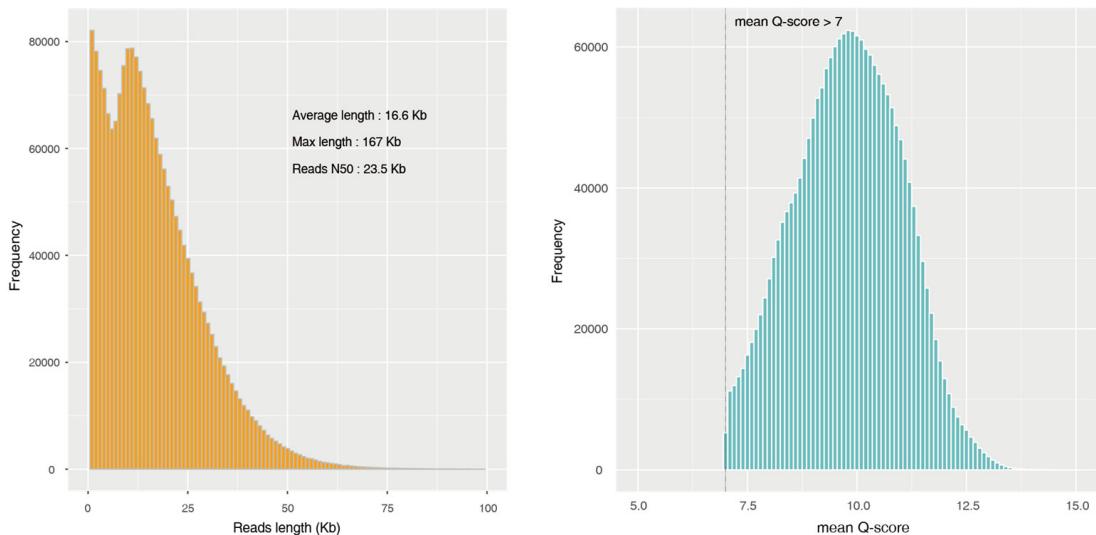
137 **Materials and Methods**

138 The same individual of *A. fimbriata* used for the flow cytometry and genome survey
139 was used for the following three different types of sequencing. For ONT sequencing,
140 DNA was extracted from young leaves using QIAGEN® Genomic Kits, and its
141 quality was then assessed by 0.75% agarose gel and Nanodrop. Libraries with an
142 insert size of 20-40 Kb were then prepared and sequenced on a GridION X5
143 instrument. For optical maps, DNA was extracted from young leaves according to a
144 modified Bionano genomics protocol¹⁶. The long high-quality DNA was labeled by
145 enzyme Nt.BspQI, and then loaded into the Saphyr chip for scanning. The output of
146 the Bionano single-molecule map files in bnx format was subjected to an initial
147 quality check to obtain high-quality molecule maps. To collect sufficient material for

148 Hi-C sequencing, we cultivated the seedlings by tissue culture using stem cuttings
149 from the same individual used for the above sequencing. The samples were processed,
150 and the DNA was extracted and crosslinked using the standard protocol. The Hi-C
151 libraries were then amplified and sequenced with 150 bp paired-end reads using
152 Illumina HiSeq.

153 **Results and Discussion**

154 We achieved about 34 Gb of Nanopore long reads after filtering out the runs with the
155 mean Q-scores less than 7, and the final length of the clean reads is about 120 \times
156 coverage of the estimated genome size of *A. fimbriata* (Supplementary Table 1.3).
157 The maximum length of the long reads is about 167 Kb, and the reads N50 and
158 average length are 23.5 Kb and 16.6 Kb, respectively (Supplementary Fig. 1.1). For
159 the Bionano optical maps, we obtained 404,934 molecules with minimum length of
160 180 Kb, and a total length of ~142 Gb (552 \times coverage of the genome)
161 (Supplementary Table 1.4). For the Hi-C sequencing, there is 45 Gb valid data, about
162 180 \times coverage of the genome (Supplementary Table 1.5).



163
164 **Supplementary Fig. 1.1 | The frequency of ONT long reads length (left) and Q-**
165 **score (right).** The lowest mean Q-score value was setting to 7.

166 **1.3 Nuclear and chloroplast genome assemblies**

167 **Materials and Methods**

168 **Nuclear genome *de novo* assembly.** The passed reads from GridION X5 runs were
169 used to generate overlap maps using minimap2 v2.15-r914¹⁷ with the parameter setting
170 of '-x ava-ont'. The graphical fragment assembly format (GFA) file was then generated

171 by miniasm v0.3¹⁸ with default parameters and was further converted into unitig
172 sequences. We also performed three rounds of consensus correction using racon
173 v1.3.3¹⁹ based on minimap overlaps, and the consensus sequences were polished by
174 Illumina reads using bwa-mem v0.7.12-r1039²⁰ and Pilon v1.22²¹ with default
175 parameters (Fig. 2b).

176 **Bionano optical maps and scaffold construction.** Low-quality optical molecules
177 with length < 180 Kb or the molecule label number < 9 were removed first. The
178 optical map assembly for *A. fimbriata* was generated using the Bionano Solve
179 Pipeline v 3.3 with the parameter settings of ‘-i 0 -V 0 -A -z -u -m’
180 (<https://bionanogenomics.com/support/software-downloads/>). The first assembly was
181 reassembled by launching the settings of ‘-y -r (rough assembly cmap) -V 0 -m’.
182 Hybrid scaffolds were generated by aligning the optical maps to ONT assembled
183 genomic contigs using Bionano’s hybrid-scaffold software
184 (<https://bionanogenomics.com/support/software-downloads/>). Conflicting sites were
185 further separated for both the contig sequences and optical maps, and the resulted
186 maps and contigs were assembled again with the hybrid-scaffold parameters of ‘-B 2 -
187 N 2’.

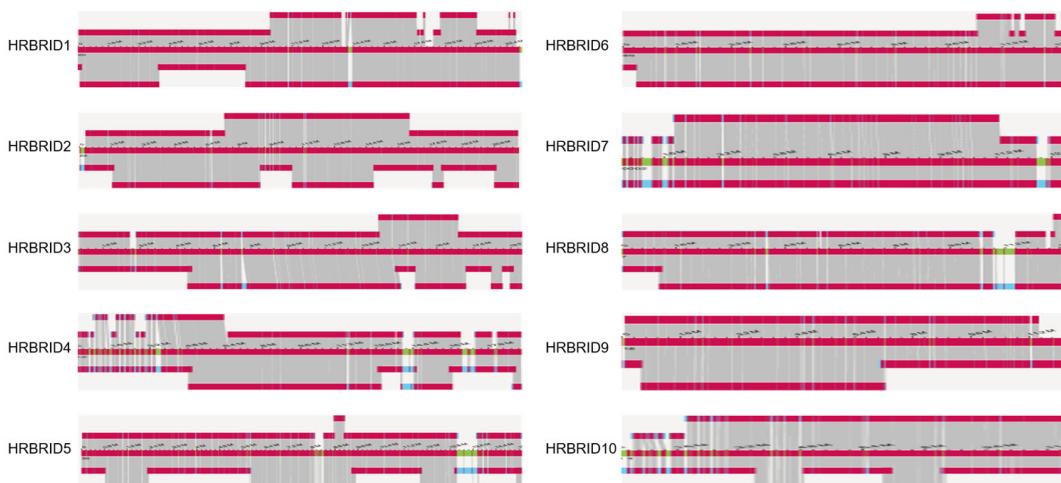
188 **Anchor scaffolds to pseudo-chromosomes using Hi-C.** Hybrid scaffolds with
189 length > 100 Kb were used to construct pseudomolecules. A Hi-C contact map
190 between genomic loci was computed using Juicer v1.7.6²² with parameters of “-s
191 MboI -r bwa mem” and visualized using Juicebox v1.8.8.8²³. 3D-DNA v180114 was
192 used to anchor and orient scaffolds²⁴ with following parameters: -m haploid -r 2. We
193 also manually corrected the order or orientation of several misassembled scaffolds
194 based on the Hi-C contact frequency and assembled the pseudo-chromosomes based
195 on the Rabl configuration of the telomere-to-telomere contact enrichment using
196 Juicebox Assembly Tools (JBAT version 1.8.8)²³.

197 **Chloroplast genome assembly.** To extract the reads from the chloroplast genome, we
198 trimmed raw ONT reads and Illumina reads using Canu v1.8²⁵ and Trimmomatic
199 v3.8²⁶ respectively, and then mapped them to the chloroplast genome of the
200 *Aristolochia debilis* (downloaded from GenBank, accession NC_036153) using
201 minimap2 v2.16-r922¹⁷ and bowtie2 v2.3.4.1²⁷, respectively. These reads that mapped
202 to the chloroplast genome were then extracted using samtools v1.9²⁸, and further *de*
203 *novo* assembled with SPAdes v3.11.1²⁹. The assembled scaffolds were aligned to the
204 chloroplast genome of *A. debilis* using Geneious v8.0.2 to determine the order of the
205 scaffolds³⁰, and MITObim v1.9 was used to fill gaps between the scaffolds based on
206 the Illumina short reads³¹. Lastly, the final circular chloroplast genome was plotted
207 using OGDRAW³².

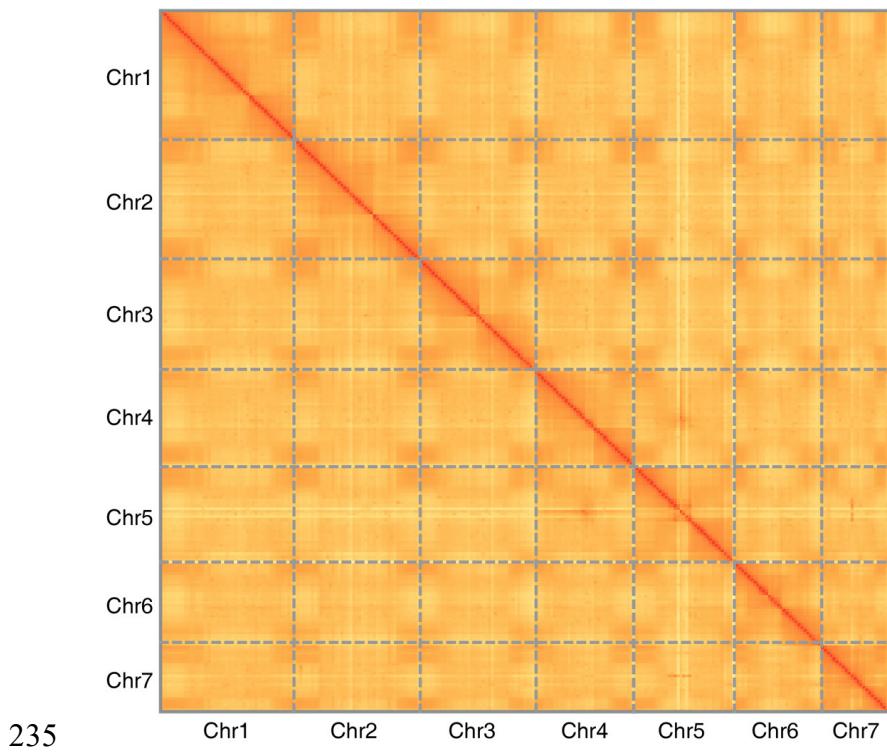
208 **Results and Discussion**

209 We assembled the 120 \times ONT clean long reads into 378 contigs (contig N50 = 5.16
210 Mb), which have better contiguity than the other published magnoliid genomes (Fig.
211 2b and Supplementary Tables 1.6 and 1.7). The consensus genomic sequences of the
212 contigs were then corrected using 100 \times Illumina reads and further used to assemble
213 scaffolds by integrating the 351 \times valid optical map data, with 89.3% of the total
214 length of the contig sequences and 97.4% of the genome maps incorporated
215 (Supplementary Fig. 1.2 and Supplementary Table 1.8). We generated a hybrid
216 assembly of 258 Mb of sequence in 283 scaffolds with an N50 of 12.89 Mb
217 (Supplementary Table 1.6). The scaffolds were further clustered and ordered using
218 Hi-C and assembled into 7 pseudo-chromosomes that covered ~95% of the assembled
219 genome (Supplementary Fig. 1.3).

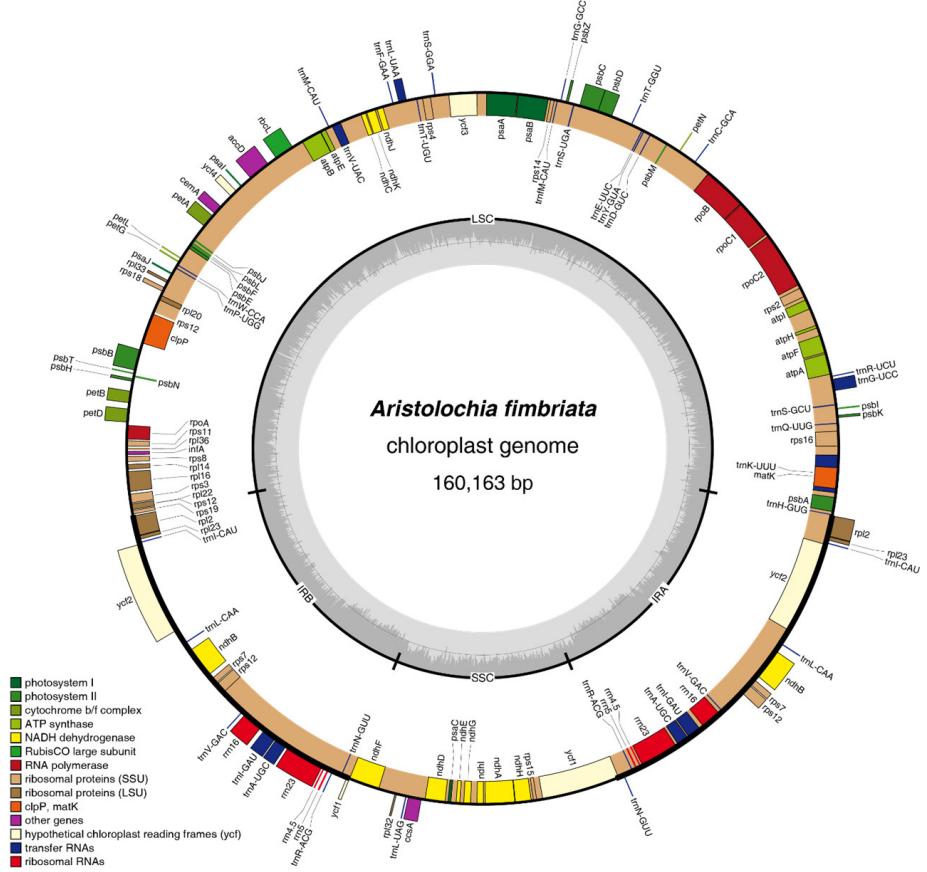
220 We also assembled the complete chloroplast genome of *A. fimbriata*, with length of
221 160,529 bp and an average GC content of 38.5%. This assembled circular molecule
222 has a typical quadripartite structure containing a large single-copy region of 90,080
223 bp, an inverted region A (IRa) of 25,433 bp, a short single-copy region of 19,583 bp,
224 and an inverted region B (IRb) of 25,433 bp. In total, 131 genes were annotated,
225 including 86 protein-coding genes, 37 transfer RNA (tRNA), and 8 ribosomal RNA
226 (rRNA) genes (Supplementary Fig. 1.4).



227
228 **Supplementary Fig. 1.2 | Alignments of the *A. fimbriata* Bionano optical maps**
229 **against in-silico maps of the largest 10 hybrid scaffolds.** Alignment of hybrid
230 scaffolds showed ONT contigs ordered and oriented after mapping to the Bionano
231 optical maps. The ONT contigs are on the top track; the hybrid scaffolds are in the
232 middle track; Bionano maps are on the bottom track. Collinear Nt.BspQI markers on
233 the two maps are linked with gray lines. The green blocks mark the gap regions in the
234 hybrid scaffolds.



236 **Supplementary Fig. 1.3 | Hi-C contact matrices of the seven pseudo-**
237 **chromosomes of the final *A. fimbriata* assembly.**



Supplementary Fig. 1.4 | The assembled circular chloroplast genome of *A. fimbriata*. In total, 131 genes were annotated, including 86 protein-coding genes, 37 transfer RNA (tRNA), and 8 ribosomal RNA (rRNA) genes.

242 **1.4 Genome quality assessments**

243 **Materials and Methods**

244 The quality and completeness of the *A. fimbriata* genome assembly were assessed
245 from four aspects. First, we evaluated the mapping rates of the clean raw reads from
246 transcriptomes and genomic DNA by TopHat2³³ and BWA-MEM²⁰ with default
247 parameters, respectively. We further used the “--vcf” option in Pilon v1.23²¹ to call
248 SNPs from the Illumina genomic reads with parameters of “--diploid --fix all --
249 mindepth 0.1”. Second, we investigated the Benchmarking Universal Single Copy
250 Orthologs (BUSCO) genes from Embryophyta in the final assembly³⁴. Third, we
251 employed the LTR Assembly Index (LAI) to infer the assembly continuity³⁵ with
252 parameters of “-step 50000 -t 8”. Finally, we aligned Bionano molecules back to the
253 final *A. fimbriata* genome assembly to check the consistency between Bionano
254 molecules and the final genome assembly using the RefAligner tool
255 (<https://bionanogenomics.com/support/software-downloads/>) with default parameters.
256 In addition, we also checked the consistency of the Bionano assembly consensus
257 genome maps (CMAP) and the in-silico maps of the *A. fimbriata* genome assembly.

258 **Results and Discussion**

259 First, the average overall mapping rate of the nine transcriptomes was about 93.2%,
260 and the average rate of the properly mapped reads was approximately 89%
261 (Supplementary Table 1.9). The overall mapping rate of the 100× Illumina PE reads
262 was 99.77%, and the rate of properly mapped reads was 95.96% (Supplementary
263 Table 1.10). Very low heterozygosity (~0.07%) was identified based on the SNPs
264 calling (Fig. 1d). The median depth of coverage of the ONT clean raw reads was
265 about 80×, and only limited gaps were observed across the seven assembled pseudo-
266 chromosomes (Extended Data Fig. 3a). Second, we identified 96.8% of the Plantae
267 BUSCO genes in the assembled *A. fimbriata* genome; the other published magnoliid
268 genomes identified 86.3% to 96.1% of the BUSCO genes (Supplementary Tables 1.11
269 and 1.12). Third, the LAI of the *A. fimbriata* assembly is about 21, similar to those of
270 well assembled *Oryza sativa* (MSU7) and *Zea mays* (B73v4) genomes (Extended
271 Data Fig. 3b,d), suggestive of high-quality assembly (Fig. 1c)³⁵. Finally, we observed
272 that the Bionano molecules uniformly and consistently mapped to the assembled
273 genome (Extended Data Fig. 3c). Moreover, 97.5% of the Bionano consensus genome
274 maps (CMAP) were uniquely aligned to the seven in-silico maps of the *A. fimbriata*
275 reference assembly (Supplementary Table 1.13). We also successfully assembled the
276 telomeric tandem repeat arrays (CCCTAAA/TTTAGGG repeats) at both distal ends
277 of chromosomes 2, 3, 5, and 6, and at one distal end of chromosomes 1 and 7
278 (Supplementary Table 1.14).

279 **2 Genome annotation and gene family classification**

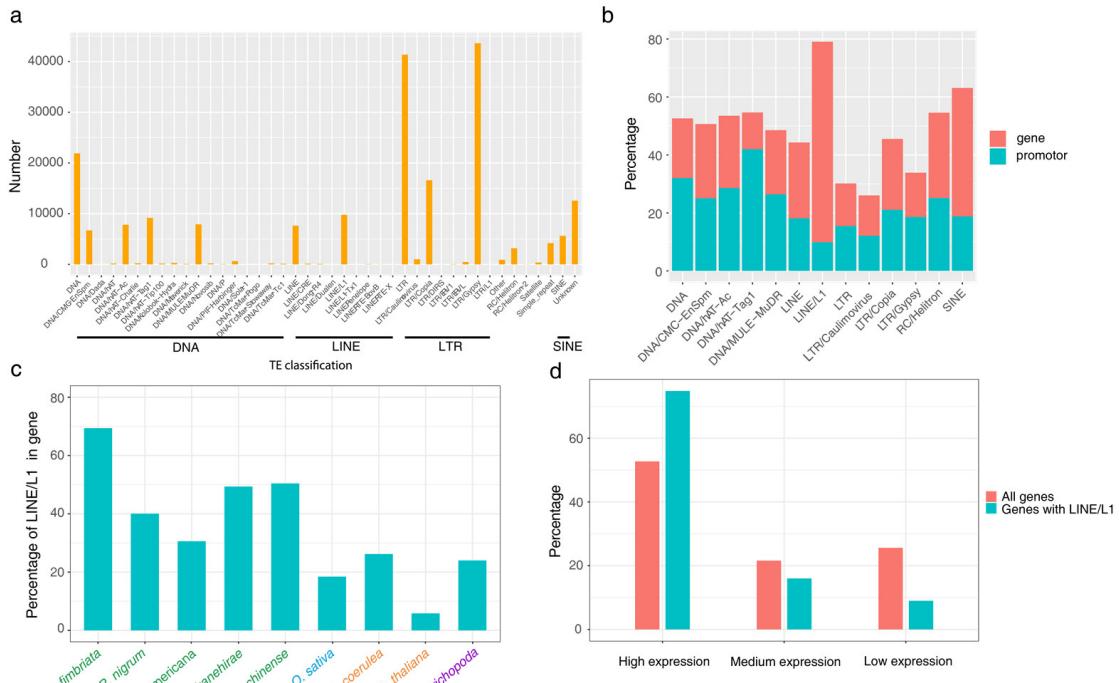
280 **2.1 Repeat annotation and TE analyses**

281 **Materials and Methods**

282 The repetitive sequences were identified using a combination of repeat similarity
283 searching and *ab initio* prediction approaches. Repbase v20.05³⁶ was employed to
284 search against the *A. fimbriata* genome using RepeatMasker v4.0.7³⁷ with default
285 parameters. For *ab initio* prediction, a consensus sequence library was built using
286 RepeatModeler v1.0.10 (<http://repeatmasker.org/RepeatModeler/>) with the parameter
287 ‘-engine ncbi’. Then, LTRharvest v1.5.10³⁸, LTR_FINDER v1.05³⁹, and
288 LTR_retriever v1.8.0⁴⁰ were used to build an LTR library with default parameters.
289 These two libraries were used to annotate the *A. fimbriata* genome using
290 RepeatMasker, and the detected transposable elements (TEs) were then combined to
291 obtain the final TE annotation. We also downloaded the genomic sequences of
292 *Amborella trichopoda*, four other magnoliids (*Piper nigrum*, *Cinnamomum kanehirae*,
293 *Liriodendron chinense*, *Persea americana*), two eudicots (*Aquilegia coerulea*, *A.*
294 *thaliana*), and one monocot (*O. sativa*) (Supplementary Table 2.1), and performed TE
295 annotation for these genomes based on the same processes for comparison.

296 **Results and Discussion**

297 We found that 52.1% of the *A. fimbriata* genome sequence is composed of
298 transposons (Supplementary Table 2.2). Long terminal repeat (LTR) retrotransposons
299 are the most abundant transposable elements (Supplementary Fig. 2.1a), of which
300 *Ty3/Gypsy*- and *Ty1/Copia-like* DNA sequences account for 21.3% and 4.6% of the
301 genome assembly, respectively (Supplementary Table 2.2). The fractions of different
302 types of TEs in the *A. fimbriata* genome are similar to those of other published
303 magnoliid genomes and *A. trichopoda* (Supplementary Table 2.3). We plotted the
304 density of the main TE types along each chromosome (Fig. 1d). The identified
305 *Ty3/Gypsy* LTRs are clustered around the centromeric region of all chromosomes,
306 while the LINE/L1 TEs tend to be located outside of the centromeric regions (Fig.
307 1d). Notably, the highest percent of LINE/L1 TEs occur in the introns of genic
308 regions in *A. fimbriata*, compared to several other genomes, and genes with LINE/L1
309 TEs insertions tend to have relatively high expression levels (Supplementary Fig. 2.1).
310 Furthermore, we also found that DNA transposon types *MULE-MuDR* and *CMC-*
311 *EnSpm* are localized to centromeric regions but absent from the rest of the genome
312 (Fig. 1d).



313
314 **Supplementary Fig. 2.1 | TE annotation in the genome of *A. fimbriata*.** (a) Number
315 of different types of TEs identified in the *A. fimbriata* genome. (b) Percentage of
316 different types of TEs located in gene or promoter regions. (c) Comparison of the
317 percentage of LINE/L1 TEs located in gene regions of the *A. fimbriata* and eight other
318 angiosperm genomes. (d) Expression levels of genes with LINE/L1 insertions in the
319 intron regions. High expression: $\text{TPM} \geq 10$; Medium expression: $1 \leq \text{TPM} < 10$; Low
320 expression: $\text{TPM} < 1$.

321 2.2 Transcriptome sequencing by Illumina and PacBio Iso-Seq

322 Materials and Methods

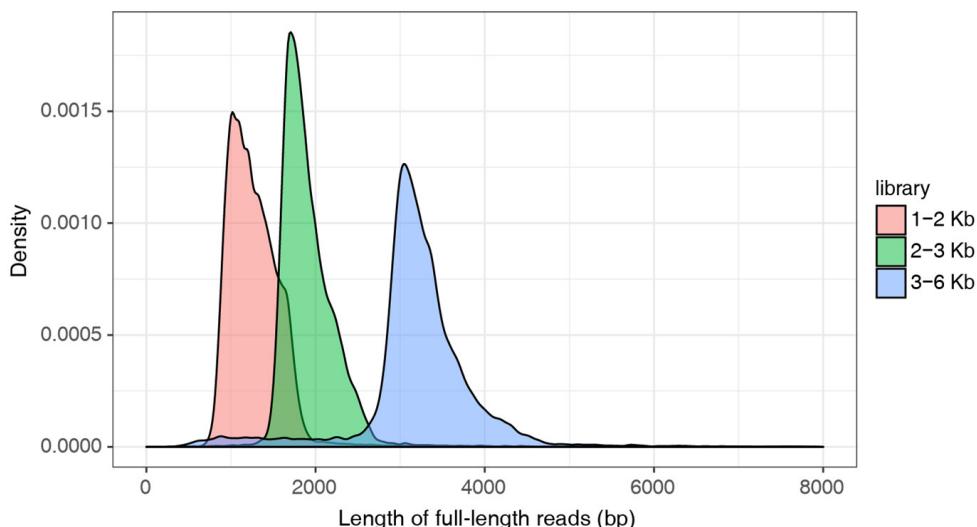
323 **Illumina RNA-Seq data.** Total RNAs were separately extracted and processed from
324 leaves, seedlings, and five different floral organs (limb, tube, utricle, gynostemium,
325 and ovary) using Trizol reagent (Invitrogen, CA, USA) following the manufacturer's
326 procedure. The paired-end cDNA libraries with insert size of 150 bp were constructed
327 and sequenced using Illumina HiSeq4000 instrument. Low-quality reads were filtered
328 out using Trimmomatic²⁶ with the parameters of 'PE -phred33 ILLUMINACLIP:
329 Trimmomatic-0.36/adapters/TruSeq3-PE.fa: 2:30:10 LEADING: 20 TRAILING: 20
330 SLIDINGWINDOW: 4:20 MINLEN: 50'.

331 **PacBio Iso-Seq data.** We sequenced a library of full-length transcripts using PacBio
332 Iso-Seq. Samples from anthetic flowers, seedlings under normal growth conditions,
333 seedlings treated with low temperature (4°C) for 9 hours, and roots were collected,

334 and the extracted RNAs from the four samples were mixed together in equal amount
335 to obtain transcriptomes from various plant tissues and treatments. The cDNA
336 libraries were constructed using the SMARTer™ PCR cDNA Synthesis Kit. The full-
337 length cDNA fragments were screened using a BluePippin instrument to construct
338 cDNA libraries of different sizes (1-2 Kb; 2-3 Kb; and 3-6 Kb) (Supplementary Fig.
339 2.2). The libraries were sequenced on a PacBio RS II instrument, and the raw reads
340 were processed using SMRT Link 5.0 software. First, the circular consensus
341 sequences (CCSs) were generated from the subreads BAM files with parameters of “--
342 minLength=300 --minPasses=1 minPredictedAccuracy=0.8”. Next, all the CCSs were
343 further classified into full-length non-chimeric (FLNC) and non-full-length (nFL)
344 transcript sequences based on whether the 5'-primers, 3'-primers, and poly(A) tail
345 could be detected. To improve consensus accuracy, we clustered and polished the FL
346 sequences using an isoform-level clustering algorithm, iterative clustering for error
347 correction (ICE), and the Quiver tool in the SRMT Link software. The FL reads were
348 further corrected using RNA-Seq reads using LoRDEC⁴¹ with the parameters of ‘-k
349 19 -s 3 -T 4’, and redundancy was removed using Cd-hit⁴² with the parameters of ‘-c
350 0.99 -T 10 -G 0 -aL 0 -aS 0.99 -AS 30 -d 0 -p 1’.

351 Results and Discussion

352 For Illumina RNA-Seq, we obtained approximately 6 Gb clean data on average for
353 each sample (Supplementary Table 2.4). For PacBio Iso-Seq, there were 4,418,800
354 raw subreads in total from the three libraries of different sizes, and we obtained
355 46,287 FLNC transcripts after multi-step processing (Supplementary Fig. 2.2). These
356 transcriptome data were used for genome annotation.



357
358 **Supplementary Fig. 2.2 | The length distribution of the full-length non-chimeric**
359 **transcripts of PacBio Iso-Seq.** The output results are consistent with the three
360 constructed cDNA libraries.

361 **2.3 Gene prediction and functional annotation**

362 **Materials and Methods**

363 The protein-coding genes were predicted using the well-developed combination
364 strategies of transcriptome, homology-based annotation, and *ab initio* gene prediction.
365 For the *ab initio* prediction, the repeat-masked scaffolds were annotated using
366 Fgenesh⁴³ and AUGUSTUS⁴⁴ with default parameters. For the homology-based
367 prediction, we used the inferred amino acid sequences from the *A. coerulea*, *Ananas*
368 *comosus*, *A. thaliana*, *A. trichopoda*, *Papaver somniferum*, and *C. kanehirae*
369 genomes. GeneWise⁴⁵ and GeMoMa⁴⁶ were used to annotate the gene models using
370 alignments from amino acid sequence similarity against the *A. fimbriata* assembled
371 sequences. For transcriptome-based prediction, PASA⁴⁷ and GMAP⁴⁸ were used to
372 predict the gene models. If the transposable domain occupied more than 60% of the
373 predicted gene length, the gene was removed using TransposonPSI
374 (<http://transposonpsi.sourceforge.net>). Finally, the results from the three approaches
375 were integrated to generate EVidenceModeler (EVM)⁴⁹ gene models to obtain the
376 final annotated protein-coding gene set.

377 The putative functions of the genes were predicted by searching the best matched
378 proteins in SwissProt (https://web.expasy.org/docs/swiss-prot_guideline.html), non-
379 redundant (Nr) (<https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/>) and Eukaryotic
380 Orthologous Groups (KOG)
381 (<https://hsls.pitt.edu/obrc/index.php?page=URL1144075392>) databases using
382 BLASTP (*E*-value $\leq 10^{-6}$). Gene Ontology (GO) terms were also assigned to the genes
383 by combining the results from Blast2GO v5.2.5⁵⁰ and eggNOG-mapper v22⁵¹
384 annotations. We also used the KEGG database (<https://www.genome.jp/kegg/>) to
385 obtain KEGG orthologs to infer putative gene pathways.

386 **Results and Discussion**

387 For the *ab initio* gene annotation, we predicted 24,596 genes using AUGUSTUS⁴⁴,
388 and 21,774 genes using Fgenesh⁴³. For the protein similarity-based prediction, we
389 obtained 14,868 to 18,677 genes using GeMoMa⁴⁶ and 17,504 genes using
390 GeneWise⁴⁵. For transcriptome-based predictions, we obtained 11,787 genes from the
391 Illumina RNA-Seq and 12,091 genes from PacBio Iso-Seq. We combined these
392 results together, and annotated 21,751 genes in total for the *A. fimbriata* genome,
393 19,582 of which were classified as high-confidence genes based on whether they have
394 support from the aforementioned transcriptomes and whether they exhibit overlapping
395 with TEs. The annotated genes were further compared among *A. fimbriata*, *A.*
396 *trichopoda*, and representative genomes of magnoliids, eudicots, and monocots
397 (Supplementary Table 2.6). In addition, we were able to functionally annotate 17,966

398 (82.6%) genes using the databases of SwissProt, KOG, KEGG, GO, and Nr. The
399 remaining 3,785 genes were treated as “unknown function” (Supplementary Table
400 2.7).

401 **2.4 Gene family classification and comparison**

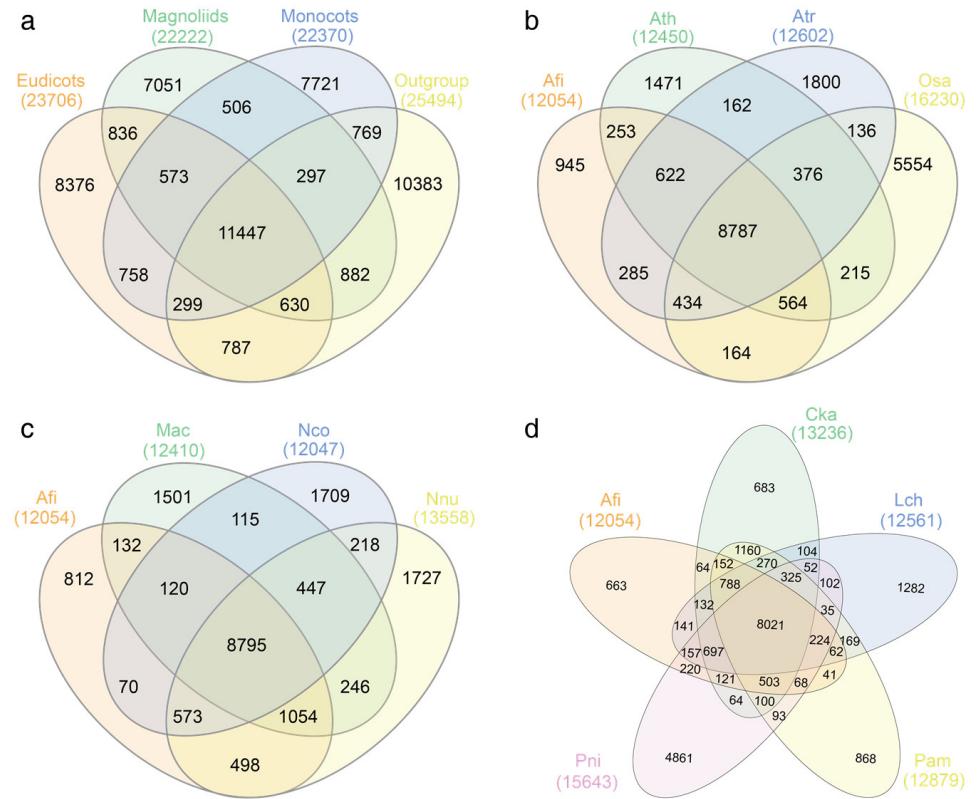
402 **Materials and Methods**

403 We selected 22 species to construct putative gene families, including five monocots
404 (*O. sativa*, *Sorghum bicolor*, *Musa acuminata*, *Spirodela polyrhiza*, and *Phalaenopsis*
405 *equestris*), seven eudicots (*Vitis vinifera*, *A. thaliana*, *Populus trichocarpa*, *Solanum*
406 *lycopersicum*, *P. somniferum*, *A. coerulea*, and *Nelumbo nucifera*), five magnoliids
407 (*L. chinense*, *C. kanehirae*, *P. americana*, *P. nigrum*, and *A. fimbriata*), two
408 representatives of the ANA grade (*A. trichopoda* and *Nymphaea colorata*), two
409 gymnosperms (*Ginkgo biloba* and *Picea abies*), and one lycophyte (*Selaginella*
410 *moellendorffii*) (Supplementary Table 2.8). The longest transcript isoform for each
411 locus was selected for all-vs-all BLASTP⁵² with an *E*-value cut-off setting of 10^{-5} .
412 OrthoMCL v2.0.9⁵³ was used to identify gene clusters of putative gene families, and
413 the inflation parameter was set to 1.5 in the mcl process⁵⁴. The output from
414 OrthoMCL was summarized using a custom Python script to obtain the number of
415 genes from each species belonging to the orthogroups (OGs). Venn diagrams of the
416 selected taxa were generated using InteractiVenn (<http://www.interactivenn.net/>).

417 **Results and Discussion**

418 In total, 577,294 genes (77.1% of the total sequences) from the selected 22 species
419 were clustered into 51,315 OGs (Supplementary Table 2.9). 3,327 OGs contained
420 genes from all 22 species, and 7,051 OGs were magnoliid-specific (Supplementary
421 Table 2.10). For the annotated 21,751 genes in *A. fimbriata*, 17,995 were clustered
422 into 12,054 OGs, 335 of which contained only *A. fimbriata* genes and could be
423 species-specific (Supplementary Tables 2.9 and 2.10). The numbers of gene families
424 shared among different clades of angiosperms, and among representative species, are
425 shown in the Venn diagrams (Supplementary Fig. 2.3).

426



427 **Supplementary Fig. 2.3 | Venn diagrams showing the numbers of shared and**
428 **species/clade specific OGs in the 22 selected species.** (a) The common and clade-
429 specific OGs among magnoliids, eudicots, monocots, and outgroups (two
430 gymnosperms and a lycophyte). (b) Venn diagram of OGs shared by *A. fimbriata*
431 (Afi), *A. thaliana* (Ath), *O. sativa* (Osa), and *A. trichopoda* (Atr). (c) Venn diagram of
432 OGs shared by *A. fimbriata* (Afi), *N. nucifera* (Nnu), *M. acuminata* (Mac), and *N.*
433 *colorata* (Nco). (d) Venn diagram of OGs shared by the five magnoliid species, Afi,
434 *P. nigrum* (Pni), *P. americana* (Pam), *L. chinense* (Lch), and *C. kanehirae* (Cka).
435 Numbers in parentheses are the total number of OGs in that species or clade; numbers
436 in the intersections show the number of OGs shared by the species or clade.

437 3 Comparative genomics and whole-genome duplication (WGD) events in 438 magnoliids

439 3.1 Comparative genomic analyses indicate no WGD in *A. fimbriata* since the 440 origin of extant angiosperms

441 Materials and Methods

442 Inter- and intragenomic comparisons were conducted to infer the WGD history in *A.*
443 *fimbriata*, sister to all other extant angiosperms⁵⁵⁻⁵⁸, and water lily (*N.*
444 *colorata* of Nymphaeales, the subsequent sister to all remaining angiosperms)⁵⁹ were

445 used as references to infer if any additional WGD events had occurred in the
446 evolutionary history of *A. fimbriata*. It has been widely accepted that no additional
447 WGDs occurred in the *Amborella* lineage after its split from the rest of extant
448 angiosperms⁵⁵. The *N. colorata* genome study revealed a recent WGD after the origin
449 of angiosperms⁵⁹. Both the *Amborella* and *Nymphaea* genomes, with their clear WGD
450 history, can serve as important references to infer the evolutionary history of
451 *Aristolochia*.

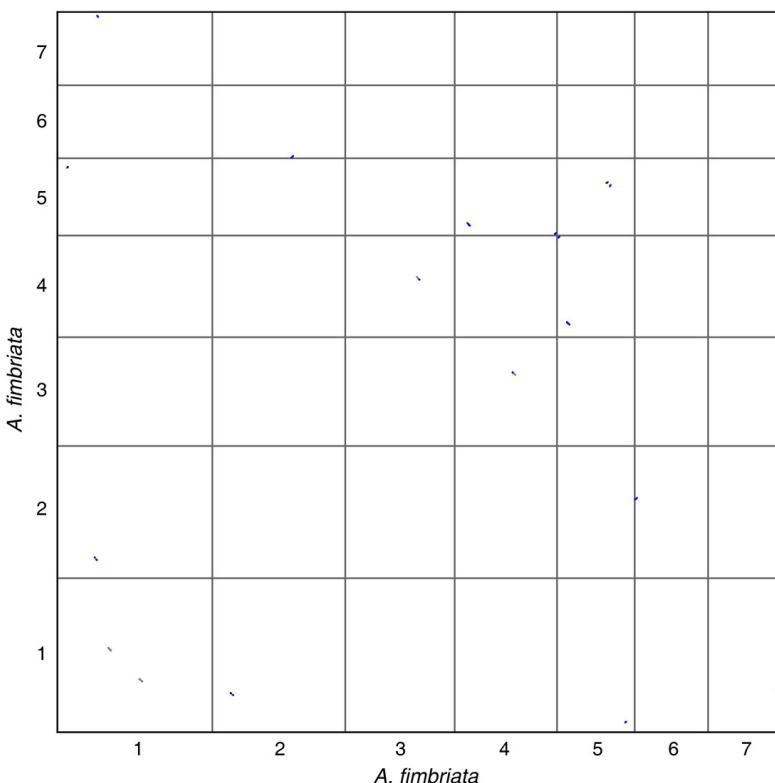
452 For the genome structural comparisons, we investigated the intra- and intergenomic
453 syntenic blocks that were then shown in dotplots. First, we performed all-against-all
454 BALSTP (*E*-value < 10⁻⁵, and Score > 100)⁵² within and between genomes. Then, the
455 top ten BLAST matches are selected for inferring syntenic blocks within or between
456 genomes. We employed MCScanX⁶⁰ to identify syntenic blocks by setting the
457 maximum gap between the anchor genes to 25, as commonly implemented in other
458 studies^{61,62}. We further plotted the syntenic gene pairs according to their genomic
459 locations in dotplots, and used different color-coded dots to distinguished whether the
460 anchor gene pairs are the best BLAST hit within/among the genomes. Finally, we
461 inferred the WGD history by investigating the syntenic depth ratios within and among
462 genomes.

463 The median synonymous substitution rate (*Ks*) values of syntenic anchor gene pairs
464 were further employed to determine the divergence degree of the identified syntenic
465 blocks. First, *Ks* was estimated using the Nei–Gojobori approach⁶³ implemented in
466 the Bioperl statistical module. Then, we adopted a kernel function analysis to obtain
467 the *Ks* distribution, which was then simulated as a mixture of multiple normal
468 distributions by the kernel smoothing density function (*Ks* density, width was set to
469 0.05). Lastly, we performed the Gaussian multi-peak fitting of the curve by using the
470 gaussian approximation function Gaussian in the fitting toolbox cftool. We set the R-
471 squared above 95% which is a parameter to evaluate the fitting level. The smallest
472 number of normal distributions was used to represent the multiple peaks of the *Ks*
473 distribution.

474 Results and Discussion

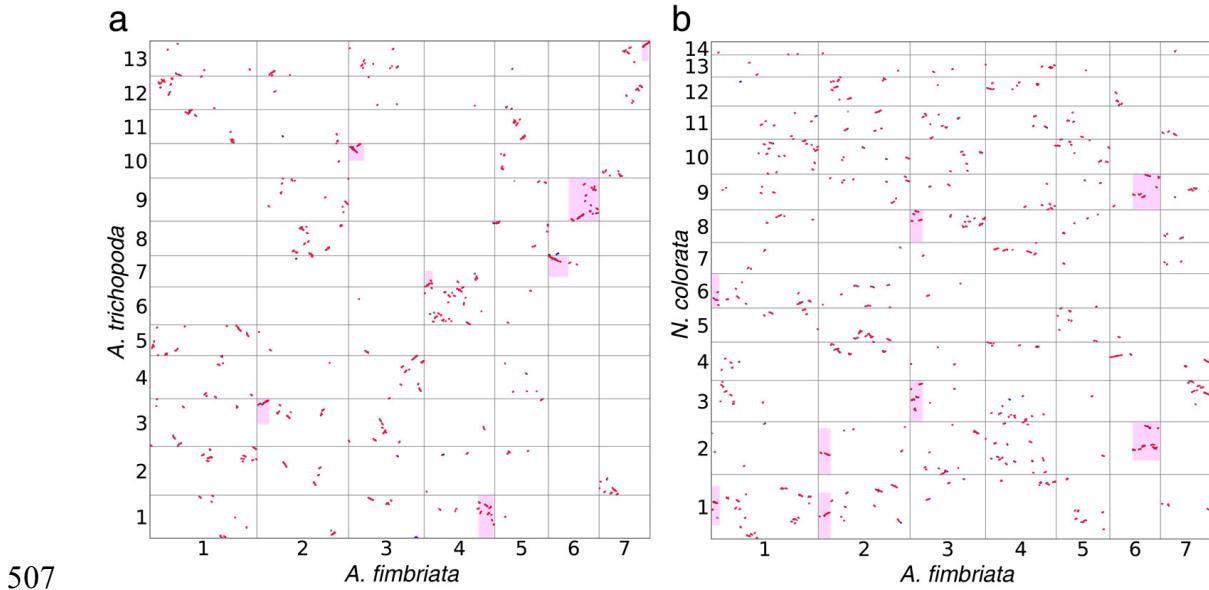
475 Intra-genomic comparison of *A. fimbriata* identified very sparse and weak syntenic
476 blocks (Supplementary Fig. 3.1), which were potentially retained from the previously
477 identified WGD characterizing all living angiosperms⁶⁴. In the comparison of the *A.*
478 *fimbriata* and *Amborella* genomes, we identified 450 microsyntenic blocks
479 comprising 6,378 anchor genes in each genome, in which 207 and 10 syntenic blocks
480 have more than 10 and 50 anchor gene pairs, respectively (Supplementary Fig. 3.2a
481 and Supplementary Table 3.1). The longest syntenic block, which is between *A.*
482 *fimbriata* chromosome 3 and *Amborella* chromosome 4, contains 77 syntenic anchor

483 gene pairs. We found that each *Amborella* genomic region only matched a single
 484 region in *A. fimbriata* (Supplementary Fig. 3.2a). The results clearly showed that the
 485 syntenic depth ratio between *Amborella* and *A. fimbriata* is 1:1, indicating that, as in
 486 *Amborella*, no independent WGD has occurred in *A. fimbriata* since the shared WGD
 487 that occurred in the common ancestor of all extant angiosperms. Moreover,
 488 intergenomic comparison between *A. fimbriata* and *N. colorata* revealed that their
 489 corresponding syntenic depth ratio is 1:2 (Supplementary Fig. 3.2b), which is
 490 consistent with an independent WGD having occurred in the *N. colorata* lineage⁵⁹.
 491 Additionally, only 3/567 syntenic regions with more than 50 anchored gene pairs
 492 were detected (Supplementary Table 3.1), which again suggests extensive
 493 chromosomal rearrangements following each round of WGD⁶⁵. Taking one
 494 microsyntenic comparison as an example, we identified the region of Chr1:6.61-7.12
 495 Mb in *A. fimbriata* matching with two genomic regions, Chr8:32.66-33.38 Mb and
 496 Chr10:7.79-8.76 Mb, in *N. colorata*, and one region, Chr4:52.21-48.06 Mb, in
 497 *Amborella* (Supplementary Fig. 3.3). Therefore, our results clearly indicate that no
 498 WGD occurred in the history of *A. fimbriata* since its shared ancestry with *Amborella*.
 499 Importantly, the high degree of conservation between *A. fimbriata* and *Amborella*
 500 enables the construction of the ancestral angiosperm genome and also indicates that
 501 the *A. fimbriata* genome would be an exceptional evolutionary reference genome for
 502 flowering plants.

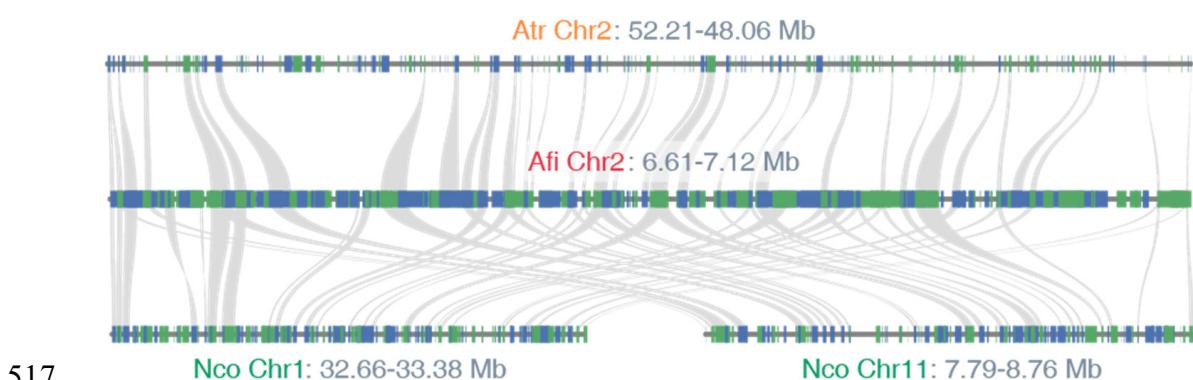


503

504 **Supplementary Fig. 3.1 | Synteny of the self-genomic comparison in *A. fimbriata*.**
 505 Genomic synteny blocks (≥ 5 anchor gene pairs) inferred from MCScanX were
 506 shown in dotplot according to their genomic locations in *A. fimbriata*.



508 **Supplementary Fig. 3.2 | Intergenomic comparative analyses.** (a) Syntenic dotplot
 509 between the *A. fimbriata* and *A. trichopoda* genomes. (b) Syntenic dotplot between
 510 the *A. fimbriata* and *N. colorata* genomes. Genomic synteny blocks (≥ 5 anchor gene
 511 pairs) were shown in dotplots according to their genomic locations in *A. fimbriata*, *A.*
 512 *trichopoda*, and *N. colorata*. If the anchor gene pairs are the best BLAST hit among
 513 the genomes, they are plotted as red dots; otherwise, they are shown in blue dots.
 514 Highlighted purple boxes indicate the selected orthologous regions between *A.*
 515 *fimbriata* and *A. trichopoda* with syntenic depth ratio of 1:1, and the syntenic regions
 516 between *A. fimbriata* and *N. colorata* with syntenic depth ratio of 1:2.



518 **Supplementary Fig. 3.3 | Representative microsynteny among the genomes of *A.***
 519 ***trichopoda* (Atr), *A. fimbriata* (Afi), and *N. colorata* (Nco).** Rectangles represent
 520 predicted genes with orientation on the same strand (blue) or reverse strand (green).
 521 Gray lines link gene pairs with synteny relationship. The synteny depth ratio is
 522 clearly shown as 1 (Atr) : 1 (Afi) : 2 (Nco).

523 **3.2 Three rounds of WGDs detected in the black pepper genome using the**
524 **reference genome of *A. fimbriata***

525 **Materials and Methods**

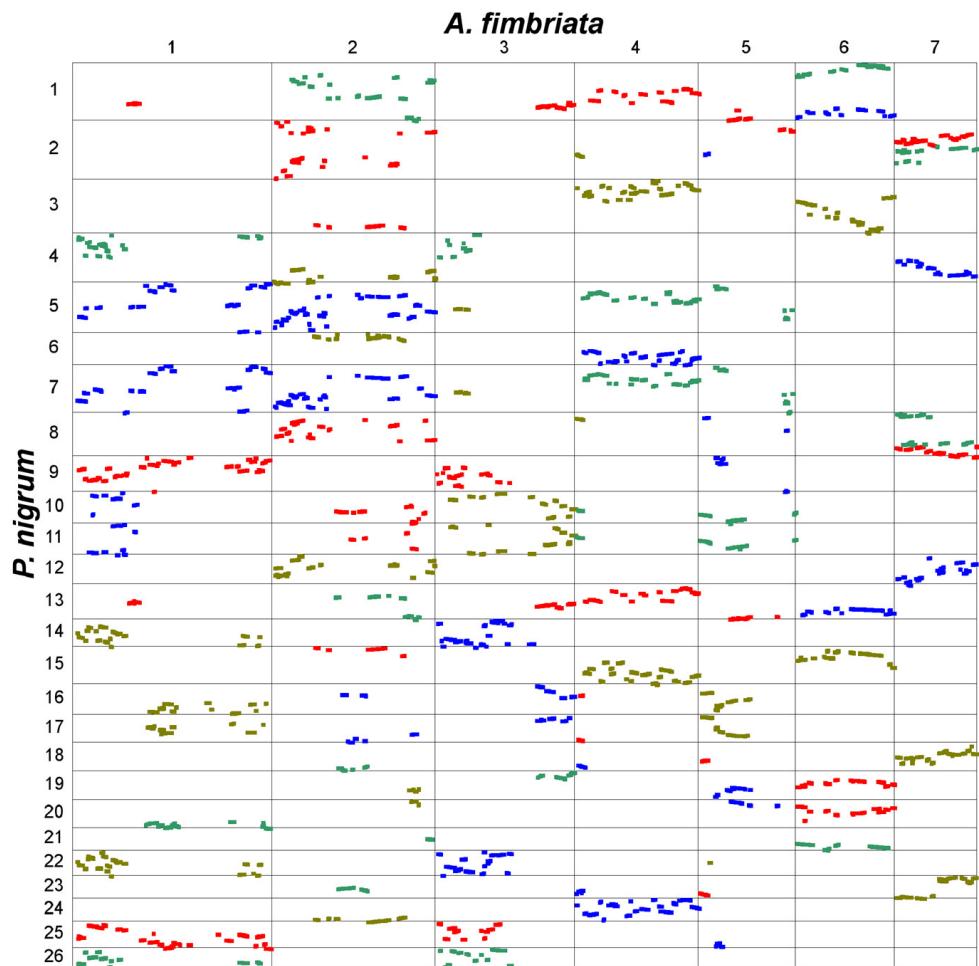
526 The high-quality genome assembly of *A. fimbriata*, lacking any additional WGD since
527 the origin of extant angiosperms, is an excellent reference for comparative genomics
528 in magnoliids, and for all angiosperms more generally. Using the *A. fimbriata* genome
529 as a reference, we compared the genome of black pepper (*P. nigrum*) using
530 MCScanX⁶⁰ with the same parameters described in methods Supplementary Note 3.1
531 and investigated the syntenic depth ratio to infer the extent of WGD in *P. nigrum*. For
532 black pepper syntenic blocks, we also calculated the *Ks* values to identify their kernel
533 density.

534 **Results and Discussion**

535 Comparing the genomes of *A. fimbriata* and *P. nigrum*, we revealed well-preserved
536 intergenomic homologous regions, with one *A. fimbriata* chromosomal region
537 matching eight homologous regions in the black pepper genome (Fig. 2b,c and
538 Supplementary Fig. 3.4). In total, 1,399 intergenomic synteny regions were identified
539 (107/1,399 synteny blocks with >20 synteny gene pairs), which contain 15,365
540 anchor gene pairs with 6,756 *A. fimbriata* genes and 14,080 black pepper genes. We
541 further performed synteny analysis of the self-comparison for black pepper
542 (Supplementary Fig. 3.5), and estimated the median *Ks* values for the paralogous
543 synteny blocks. Three different duplication periods were identified with *Ks* around
544 0.11, 0.69, and 0.91, respectively (Supplementary Fig. 3.6). Our results suggested that
545 three independent polyploidization events (Pn- α , Pn- β , and Pn- γ) occurred in the
546 evolutionary history of black pepper, and the more ancient two events (Pn- β and Pn- γ)
547 seem to have been overlooked in a previous study⁶⁶.

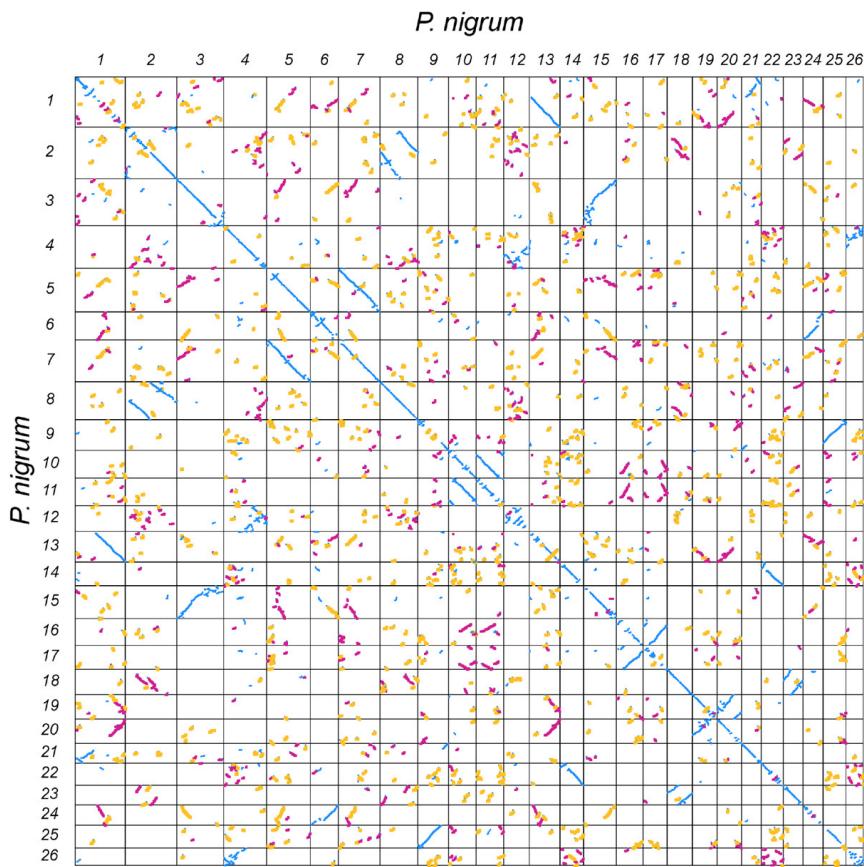
548 To further classify the synteny blocks into three events, we carefully investigated the
549 density and sequence similarity of the inter- and intragenomic synteny. For example,
550 the *A. fimbriata* chromosome 6 matched to eight orthologous regions in the *P. nigrum*
551 genome that could be further classified into four groups based on the block length and
552 their median *Ks* values (Fig. 2b and Supplementary Fig. 3.4). The two regions in
553 group_A (Pn1p-Pn13, *Ks* ~0.11, colored blue), group_B (Pn3-Pn15, *Ks* ~0.15, colored
554 orange), group_C (Pn1q-Pn21, *Ks* ~0.14, colored yellow), and group_D (Pn19-Pn20,
555 *Ks* ~0.15, colored green) were duplicated from the most recent WGD Pn- α (Fig. 2b),
556 and they also show great synteny length and density (Supplementary Fig. 3.5). The
557 median *Ks* value of collinear genes between the group_A and group_B, and between
558 the group_C and group_D are 0.71 and 0.64, suggesting they were duplicated from
559 the Pn- β event. The *Ks* value between groups_A/B and groups_C/D is about 0.89,

560 indicating that they were duplicated from the most ancient Pn- γ event (Supplementary
 561 Fig. 3.6). Therefore, we found that three rounds of WGDs differentiate black pepper
 562 from *A. fimbriata* (and from the common ancestor of all extant angiosperms) instead
 563 of one WGD reported for black pepper previously⁶⁶. Three WGDs in black pepper are
 564 also consistent with its high chromosome number of $2n = 52$, which is 4x based on a
 565 base chromosome number for *Piper* of $x = 13$ ⁶⁷, a number which itself can now be
 566 inferred to be polyploid.



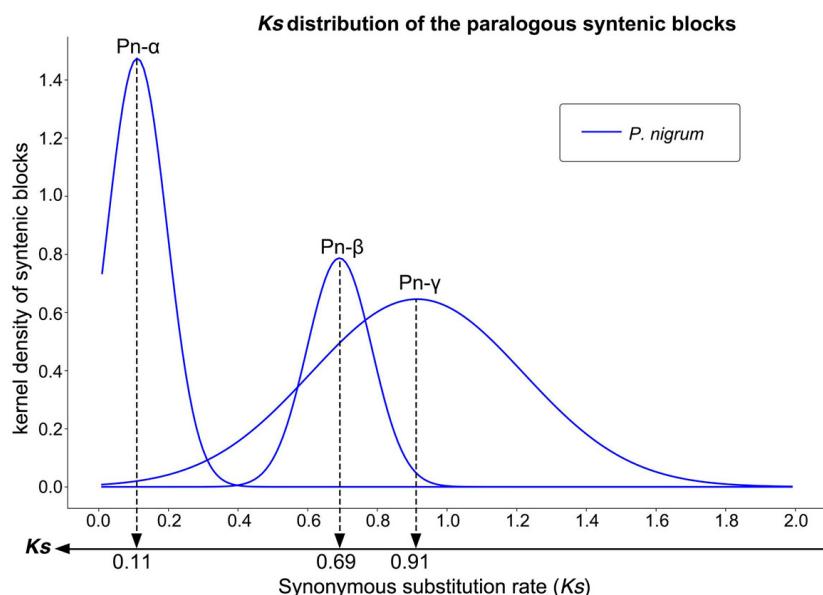
567

568 **Supplementary Fig. 3.4 | Syntenic blocks between *A. fimbriata* and *P. nigrum***
 569 **genomes.** Dotplot shows the orthologous regions detected by MCScanX, and the
 570 same color-coded regions denote duplicated regions from Pn- α in *P. nigrum*.



571

572 **Supplementary Fig. 3.5 | Intron genomic syntenic analysis in *P. nigrum*.** The median
 573 *Ks* values of the paralogous syntetic blocks were used to infer their duplication
 574 timing, and then the syntetic blocks were color-coded according to the median *Ks*
 575 value of the anchor gene pairs in the block. Blue means the *Ks* value ≤ 0.3 ; purple
 576 represents $0.3 < Ks \leq 0.8$; and orange indicates *Ks* > 0.8 .



577

578 **Supplementary Fig. 3.6 | *Ks* distribution of the identified syntetic blocks in *P.***
 579 *nigrum*. Pn- α , Pn- β , and Pn- γ represent the three rounds of WGDs detected here in *P.*
 580 *nigrum*.

581 **3.3 Timing of the previously identified WGDs in magnoliids**

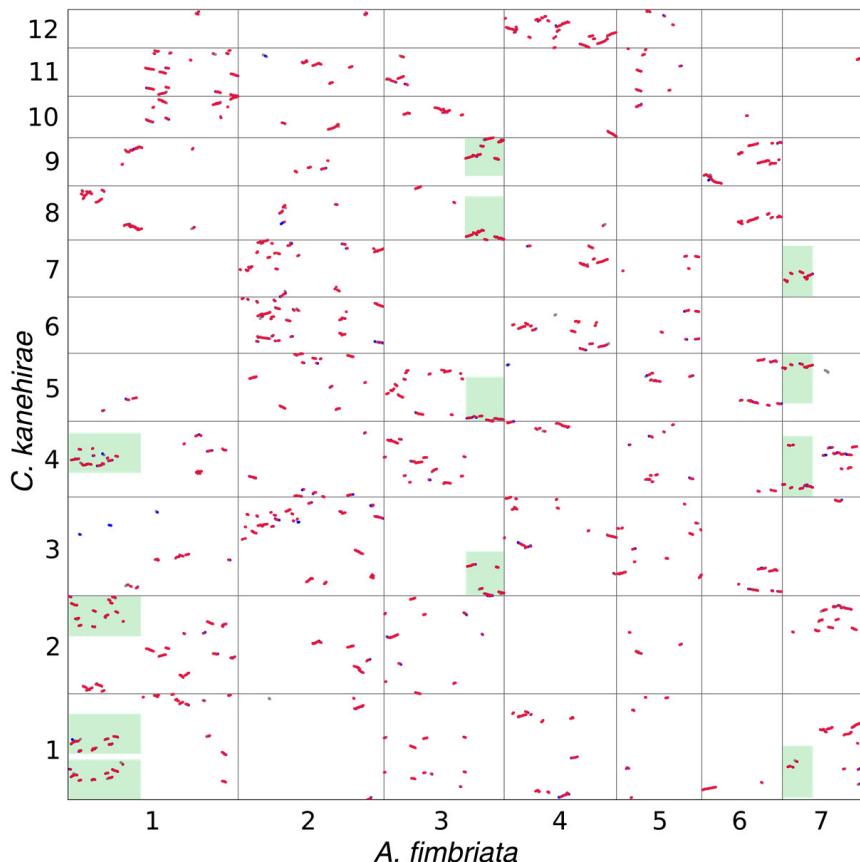
582 **Materials and Methods**

583 Genomic studies of *C. kanehirae* and *P. americana* (Lauraceae, Laurales) have
584 proposed two WGD events occurred in their evolutionary history, and *L. chinense*
585 (Magnoliaceae, Magnoliales) has experienced one ancient WGD⁶⁸⁻⁷⁰. Interestingly,
586 isozyme data suggested ancient WGD events in both Laurales and Magnoliales⁷¹. On
587 the basis of EST analysis, Cui *et al.* (2006) determined that the WGD detected in *P.*
588 *americana* and *Liriodendron tulipifera* was not shared⁷². Recently, the *Ks* analyses in
589 the *Litsea* (Lauraceae) genome suggested that an ancient WGD occurred just before
590 the divergence of Laurales and Magnoliales⁷³. However, the wintersweet
591 (*Chimonanthus praecox*; Calycanthaceae, Laurales) genome study proposed no WGD
592 event was shared by *Cinnamomum* and *Liriodendron* using *Ks* approach dating these
593 WGDs⁷⁴. As yet, it remains unclear if these sister clades (Laurales, Magnoliales) in
594 fact share any ancient WGD events. We employed the integrated approaches of
595 synteny, *Ks*, and phylogenomic analyses to investigate the timing of these WGDs. *Ks*
596 values were estimated using the Nei-Gojobori approach⁶³, and the *Ks* distribution was
597 fitted as a mixture of multiple normal distributions. We also adopted a kernel function
598 analysis to further classify the syntenic blocks based on the *Ks* values of anchor
599 genes. Here, *Ks* correction was applied by using grape (*V. vinifera*) as a comparing
600 reference to make its divergence (*Ks*) similar to the studied magnoliid genomes.
601 Detailed methods were described in the previous studies^{61,75}.

602 **Results and Discussion**

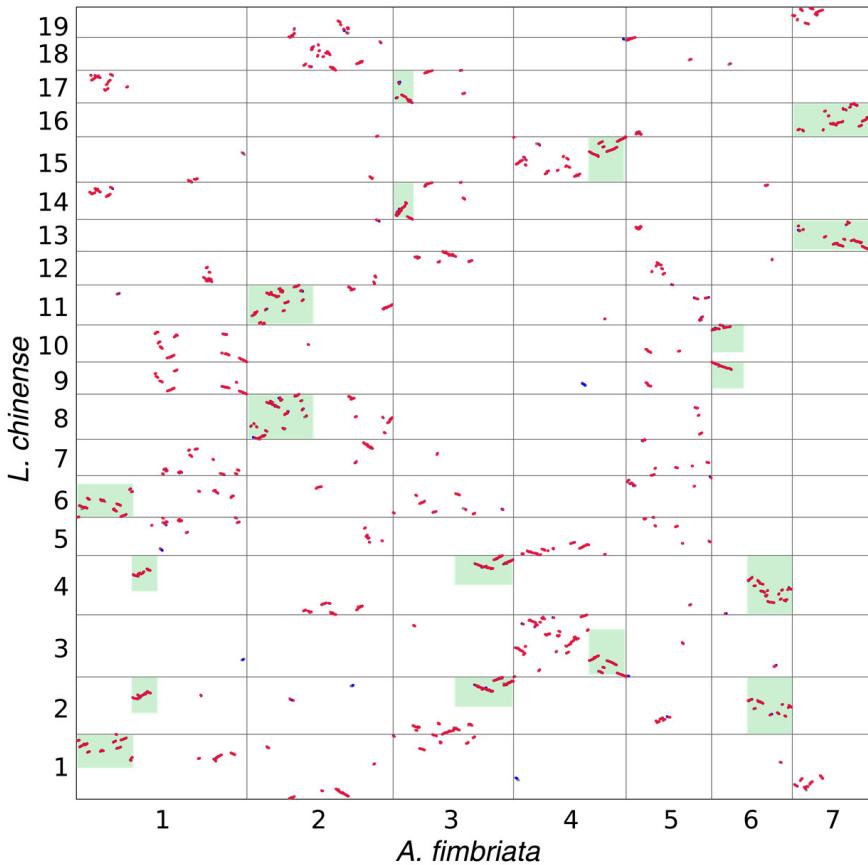
603 The intergenomic syntenic depth ratio between *A. fimbriata* and *L. chinense* is 1:2,
604 and the ratio between *A. fimbriata* and *C. kanehirae* is 1:4 (Fig. 2d and
605 Supplementary Figs. 3.7 and 3.8). These results are consistent with the previously
606 identified two WGDs in *C. kanehirae*⁷⁰ and one WGD in *L. chinense*⁶⁹. We performed
607 *Ks* analyses of the syntenic anchor gene pairs in these genomes. In *C. kanehirae*
608 paralogous regions, the *Ks* distribution of anchor genes showed a clear bimodal
609 pattern with *Ks* peaks at 0.535 (± 0.06) and 0.858 (± 0.09), which is consistent with the
610 two WGDs (Supplementary Fig. 3.9a and Supplementary Table 3.2). The *Ks*
611 distribution of *L. chinense* showed one major peak at 0.679 (± 0.07) (Supplementary
612 Fig. 3.9a and Supplementary Table 3.2). The *Ks* values of the orthologous gene pairs
613 between *C. kanehirae* and *L. chinense* were also calculated to estimate the divergence
614 between these genomes and the relative timing of their split. After evolutionary rate
615 correction, the larger *Ks* peak of the *C. kanehirae* genome shows a little larger than
616 the peak of the *C. kanehirae* - *L. chinense* divergence, but the evidence from the *Ks*
617 analyses seems hard to clarify if the ancient WGD in *C. kanehirae* was shared by *L.*
618 *chinense* (Supplementary Fig. 3.9b and Supplementary Table 3.3).

619 We further performed integrated phylogenomic and synteny analyses for the syntetic
 620 anchor gene pairs as described before^{76,77}; 67.33% (1,350/2,005) of *C. kanehirae* gene
 621 pairs in 264 syntetic blocks were duplicated before the split of *C. kanehirae* and *P.*
 622 *americana*, and 20.45% (410/2,005) were duplicated before the divergence of *C.*
 623 *kanehirae* and *L. chinense* (Supplementary Table 3.4). Similarly, 70.33% (410/583) of
 624 the *P.americana* anchor gene pairs in 228 syntetic blocks were duplicated before the
 625 divergence of *C. kanehirae* and *P. americana* (Supplementary Table 3.5) and 13.55%
 626 (79/583) of the anchor gene pairs were duplicated before the divergence of *L.*
 627 *chinense* and *P.americana* (Supplementary Table 3.5). For 35 syntetic blocks of *L.*
 628 *chinense*, we found that 54.5% (462/848) of the anchor genes were duplicated before
 629 the split of *C. kanehirae* and *L. chinense* (Supplementary Table 3.6). These results
 630 strongly suggest that the identified WGD in *L. chinense* (named LCT) was shared
 631 with *C. kanehirae* and *P. americana*, and one additional WGD event (named CCT)
 632 occurred after the *C. kanehirae-L. chinense* divergence, and before the split of *P.*
 633 *americana* and *C. kanehirae*. Using the corrected *Ks* values, we inferred that the LCT
 634 event occurred approximately 74.09-83.75 million years ago (Ma), and the more
 635 recent CCT event occurred approximately 46.31-52.35 Ma (Supplementary Fig. 3.9b).



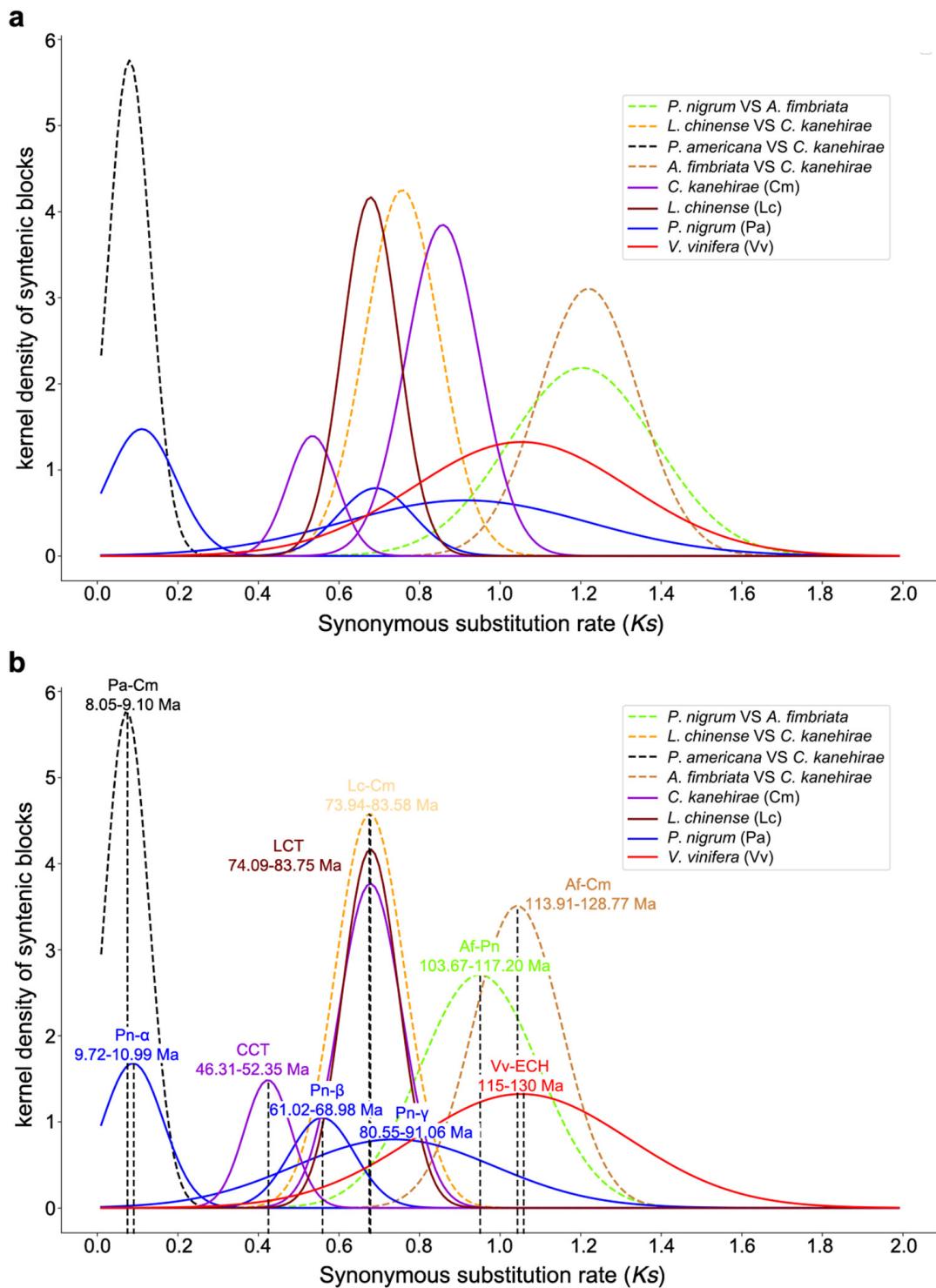
636
 637 **Supplementary Fig. 3.7 | Synteny analysis between the *A. fimbriata* and *C.***
 638 ***kanehirae* genomes.** Genomic syntetic blocks (≥ 5 anchor gene pairs) were shown in
 639 dotplots according to their genomic locations in *A. fimbriata* and *C. kanehirae*. If the
 640 anchor gene pairs are the best BLAST hit among the genomes, they are plotted as red

641 dots; otherwise, they are shown in blue dots. Highlighted boxes indicate the selected
642 orthologous regions between *A. fimbriata* and *C. kanehirae* with syntenic depth ratio
643 of 1:4.



644
645 **Supplementary Fig. 3.8 | Synteny analysis between the *A. fimbriata* and *L.***
646 ***chinense* genomes.** Syntenic blocks (≥ 5 anchor gene pairs) were shown in dotplots
647 according to their genomic locations in *A. fimbriata* and *L. chinense*. If the anchor
648 gene pairs are the best BLAST hit among the genomes, they are plotted as red dots;
649 otherwise, they are shown in blue dots. Highlighted boxes indicate the selected
650 orthologous regions between *A. fimbriata* and *L. chinense* that clearly show syntenic
651 depth ratio of 1:2.

652



653

654 **Supplementary Fig. 3.9 | Original (a) and corrected (b) synonymous substitution**
 655 **rate (K_s) analyses.** K_s correction was applied by using the *V. vinifera* as a reference
 656 to standardize its divergence from magnoliid genomes; details of the K_s values are
 657 summarized in Supplementary Tables 3.2 and 3.3.

658

659 **3.4 Genome structure comparisons identified genomic rearrangement events**
660 **shared by magnoliids and monocots**

661 **Materials and Methods**

662 The reference genome of *A. fimbriata* may be used to illustrate large genome
663 structural rearrangements during the evolutionary history of angiosperms. In
664 comparisons of the *A. fimbriata* genome with those of *A. trichopoda* and *N. colorata*,
665 we identified several chromosome-level rearrangements between the former and these
666 two angiosperms from the ANA grade, however, the timing of their occurrence
667 requires further exploration. Comparisons among genomes of *A. fimbriata* and other
668 magnoliids (*P. nigrum*, *L. chinense*, *Magnolia biondii*, *C. kanehirae*, and *Litsea*
669 *cubeba*), the eudicots (*V. vinifera*, *Acer yangbiense*, *Tetracentron sinense*, and *A.*
670 *coerulea*) as well as the monocots (*S. polystachya*, *A. comosus*, *Asparagus setaceus*, and
671 *Elaeis guineensis*) could reveal whether or not the identified chromosomal
672 rearrangements are shared among certain major groups of mesangiosperms.

673 Tracking the evolutionary history of the genomic rearrangement events include the
674 following steps: 1) generating synteny dotplots between the compared genomes and
675 identifying orthologous genomic regions using aforementioned methods
676 (Supplementary Note 3.1); 2) defining the involved regions of the genomic
677 rearrangements and revealing the connection pattern of these orthologous regions in
678 each studied genome; 3) reconstructing the ancestral connection pattern of these
679 involved regions for the major clades of angiosperms based on orthologous regions in
680 living species; 4) comparing the ancestral pattern of each clade with the predicted
681 pattern of the most common ancestor of extant angiosperms and identifying the shared
682 genomic rearrangements of major clades that potentially occurred before their
683 divergence.

684 **Results and Discussion**

685 By comparing the *A. fimbriata* genome to the *A. trichopoda* and *N. colorata* genomes,
686 we identified 1:1 and 1:2 synteny depth ratios, respectively. We found that the
687 chromosome 6 of *A. fimbriata* (Af6) separately matches chromosomes 7 and 9 of *A.*
688 *trichopoda*, and chromosomes 4 and 12 and chromosomes 2 and 9 of *N. colorata*
689 (Supplementary Fig. 3.10). The *N. colorata* genomic regions of chromosomes 4 and
690 12, as well as chromosomes 2 and 9, were generated by an independent WGD⁵⁹.
691 Similarly, we also found that chromosome 7 of *A. fimbriata* (Af7) has two separated
692 orthologous synteny regions in *A. trichopoda* and multiple regions in *N. colorata*
693 (Fig. 3a and Supplementary Fig. 3.10). Given the *A. trichopoda* and *N. colorata* are
694 two successively early diverging angiosperms, these synteny results indicated that
695 chromosomes 6 and 7 of *A. fimbriata* might be formed via two fusion events in an

696 ancestor of *A. fimbriata*, respectively. The entire Af6 and Af7 match with eight
697 integrated orthologous regions in *P. nigrum*, respectively (Extended Data Fig. 4a). It
698 confirms the correctness of the assembly of Af6 and Af7 in *A. fimbriata*, and further
699 suggests the fusions occurred, at least, before the divergence of *A. fimbriata* and *P.*
700 *nigrum*.

701 To get insights into the evolutionary history of these genomic rearrangements, we
702 further compared the *A. fimbriata* genome with other representatives in the major
703 clades of angiosperms. To better illustrate the genomic rearrangements related to
704 chromosome 6 of *A. fimbriata*, we defined the L and R regions for the Af6 (Af6L and
705 Af6R) according to the synteny relationship between *A. fimbriata* and *Amborella*
706 (Supplementary Fig. 3.10a). When comparing the Af6 to the genomes of *C. kanehirae*
707 and *L. cubeba* (Laurales), we identified two orthologous regions for the Af6L and
708 four orthologous regions for the Af6R, in which only the syntenic region in the
709 chromosome 9 of *C. kanehirae* (Supplementary Fig. 3.11a), as well as chromosome
710 11 of *L. cubeba*, shows an integrated L-R orthologous pattern (Supplementary Fig.
711 3.11c). When comparing to the genomes of *L. chinense* and *M. biondii* (Magnoliiales),
712 both of the orthologous regions for the Af6L and Af6R are located in separated
713 chromosomes (Extended Data Fig. 4b,c). Given the WGD generating the two
714 orthologous regions in the *L. chinense* genome shared with Laurales, we predict that
715 the ancestral orthologous regions of the Af6L and Af6R might be separated in the
716 common ancestor of Magnoliiales and Laurales, and the integrated orthologous region
717 in chromosome 9 of *C. kanehirae* might be resulted from a Laurales-specific fusion
718 event. Interestingly, we also revealed that the entire Af6 matched the integrated
719 orthologous regions in the studied four monocot genomes (*A. comosus*, *A. setaceus*, *S.*
720 *polyrhiza* and *E. guineensis*) (Extended Data Fig. 5 and Supplementary Fig. 3.12). For
721 example, the entire Af6 matches with chromosomes 4, 11, 12 and 16 in *E. guineensis*,
722 which derived from the two WGDs⁷⁸⁻⁷⁹. However, when comparing to eudicot
723 genomes, we found the orthologous region of Af6L and Af6R are located in different
724 set of chromosomes, in which the syntenic relationships exhibit the Af6L matching to
725 Vvi2, Vvi15, and Vvi16 and the Af6R matching to Vvi3, Vvi4, and Vvi18 in the *V.*
726 *vinifera* (Vvi) genome (Supplementary Fig. 3.13a). Similarly in other studied eudicot
727 genomes (*T. sinense*, *A. yangbiense*, and *A. coerulea*), the orthologous regions of
728 Af6L and Af6R are also located in separated chromosomes (Supplementary Figs. 3.13
729 and 3.14). In summary, the orthologous regions of Af6L and Af6R are in different set
730 of chromosomes in *Amborella*, *Nymphaea*, Laurales, Magnoliiales, and eudicots, and
731 the breakpoints between Af6L and Af6R are likely in similar genomic locations
732 (Af6:7.22 Mb) when examined by microsynteny analyses (Supplementary Figs. 3.15
733 and 3.16). The integrated regions of L and R were revealed in Piperales and
734 monocots, suggesting either a fusion event connected the L and R regions in a
735 common ancestor of monocots and magnoliids (with a further fission event in the
736 common ancestor of Laurales and Magnoliiales) or a paralleling evolution led to the

737 “L-R” fusion in the Piperales and monocots (Extended Data Figs. 4a and 5 and
738 Supplementary Fig. 3.12).

739 Similarly, for the chromosome 7 of *A. fimbriata*, we defined the ordered regions of
740 E[A1]-A2-B1-B2 for the Af7 and as shown in the Fig. 3a, where E[A1] represents A1
741 region overlapped with the E region. Comparison between *A. fimbriata* and *L.*
742 *chinense* (Lch), the orthologous regions of the A1-A2-B1-B2 of Af7 are integrated in
743 *L. chinense* genome which located in Lch13 and Lch16 (Fig. 3a and Extended Data
744 Fig. 4b), while the orthologous regions of the E part of the Af7 (Af7E) located in
745 other set of chromosomes which are Lch1 and Lch19 (Fig. 3a and Extended Data Fig.
746 4b). Comparison between *A. fimbriata* and *C. kanehirae* (Ck) identified four
747 orthologous regions of the A1-A2-B1-B2 in *C. kanehirae* (two on Ck1, Ck2, and Ck4,
748 these generated from the two WGDs) and three orthologous regions of Af7E located
749 in Ck4, Ck5, and Ck7 (Supplementary Fig. 3.11a and Extended Data Fig. 6). In
750 addition, the same fused syntenic pattern of A1-A2-B1-B2 of Af7 also present in *M.*
751 *biondii* and *L. cubeba* genomes, whereas the orthologous regions of Af7E are located
752 in different chromosomes (Extended Data Fig. 4c and Supplementary Fig. 3.11c).
753 Therefore, these results indicated that the fusion of A1-A2 and B1-B2 regions of Af7
754 occurred at least before the divergence of magnoliids, while a translocation event
755 involving Af7E seems to be Piperales-specific.

756 Further integrating syntenic evidence from comparisons of the *A. fimbriata* genome
757 with the two representatives of the ANA grade (*A. trichopoda* and *N. colorata*) and
758 with other magnoliids, we found that the fusion event of A1-A2-B1-B2 was
759 associated with two other genomic regions on Af1 (defined regions of D1, D2, and
760 C2) and Af3 (defined C1 region) as illustrated in Fig. 3a. In *A. trichopoda* genome,
761 the orthologous regions of A1, A2, and C1 of *A. fimbriata* are located on chromosome
762 10, while the orthologous regions of B1 and C2 of *A. fimbriata* are located on the
763 chromosome 12 (Fig. 3a and Supplementary Fig. 3.10a). In the *L. chinense* and *M.*
764 *biondii* (Magnoliales) and *P. nigrum* (Piperales), the orthologous regions of the C1
765 and C2 of *A. fimbriata* located in same chromosome (Fig. 3a and Extended Data Figs.
766 4a-c and 6). In the *C. kanehirae* and *L. cubeba* (Laurales) genomes, two of the
767 orthologous regions of the C1 and C2 of *A. fimbriata* located in separated locations of
768 the same chromosome (likely resulted from genomic inversion events), and the other
769 orthologous region of the C1 and C2 located in different chromosomes
770 (Supplementary Fig. 3.11a,c and Extended Data Fig. 6c,d). These results suggested
771 that the translocation of C2 into the D1 and D2 regions was specifically occurred in *A.*
772 *fimbriata* (Fig. 3b and Extended Data Figs. 4 and 6). Moreover, we noticed many
773 Laurales-specific fusion and inversion events that led to the connection patterns of
774 (C1-A1-A2-B1-B2-D1-D2-C2) and (A1-B1-B2-D1-D2, C1, C2) in *C. kanehirae* and
775 *L. cubeba* (Fig. 3b, Supplementary Fig. 3.11a,c and Extended Data Fig. 6c,d), which
776 are different from the pattern in Magnoliales (Fig. 3a,b and Extended Data Figs. 4b,c

777 and 6a). Notably, the two paralogous regions in *L. chinense* (as well as in *M. biondii*)
778 exhibit exactly the same connection pattern (A1-A2-B1-B2, C1-C2, D1-D2) (Fig. 3
779 and Extended Data Figs. 4b,c and 6a). Again, considering the ancient WGD shared by
780 Magnoliales and Laurales, we predict that the identified connection pattern in
781 Magnoliales is the ancestral pattern of Magnoliales and Laurales, while many
782 additional rearrangements occurred in the Laurales following the extra recent WGD.
783 Therefore, the connection patterns of the involved genomic regions in the ancestral
784 genome of magnoliids are predicted as (A1-A2-B1-B2, C1-C2, D1-D2, and E) (Fig.
785 3b). In *A. trichopoda* genome, the connection pattern exhibited as (A1-A2-C1, B1-C2,
786 B2, D1-D2, and E) (Fig. 3a,c and Supplementary Fig. 3.10a). In *N. colorata*, we
787 found the connection patterns exhibiting as (A1-A2-C1, B1-B2-C2, D1-D2, and E)
788 and (A1-A2, C1, B1-B2-C2, D1-D2, and E) (Fig. 3c and Supplementary Fig. 3.10b).

789 Then, we compared the genome structure of *A. fimbriata* with four monocots (*A.*
790 *comosus*, *A. setaceus*, *S. polystachya* and *E. guineensis*), four eudicots (*T. sinense*, *V.*
791 *vinifera*, *A. yangbiense*, and *A. coerulea*), and *Ceratophyllum demersum*, respectively.
792 To exclude the effect of the Piperales- and *A. fimbriata*-specific translocations
793 (including the merge of E[A1] and the translocation of C2 into D1-D2 region), we
794 also performed these intergenomic comparisons using *L. chinense* as another
795 companion reference (Extended Data Figs. 4-6 and Supplementary Figs. 3.11-3.14),
796 because these specific translocations are absent in the *L. chinense* genome. In the four
797 monocot genomes, we found that the connection pattern of the involved genomic
798 regions consistently shows as (A1-A2-B1-B2, C1-C2, D1-D2, and E), therefore such
799 pattern could be referred as the ancestral pattern of monocots (Fig. 3b, Extended Data
800 Fig. 5 and Supplementary Fig. 3.12). Notably, the ancestral connection pattern of
801 monocots seems to be the same as that of magnoliids, but different from those
802 identified in *A. trichopoda* and *N. colorata* (Fig. 3c). In four eudicot genomes, the
803 connection patterns of the involved genomic regions all show as (A1-A2-C1, B1-B2,
804 D1-D2-C2, and E), which could represent the ancestral pattern of eudicots (Fig. 3b
805 and Supplementary Figs. 3.13 and 3.14). The connection of A1-A2-C1 in eudicots is
806 consistent with that in *A. trichopoda* and *N. colorata* genomes, but differ from the
807 pattern in magnoliids and monocots (A1-A2-B1-B2) (Fig. 3c). Further microsynteny
808 examinations revealed that the locations of the breakpoints (Af7:7.85 Mb) is similar
809 when comparing *A. fimbriata* with these genomes of monocots, eudicots, *Amborella*,
810 and *N. colorata* (Supplementary Fig. 3.17). We also compared the genomic structure
811 of two monocot representatives (*A. comosus* and *S. polystachya*) with two eudicots (*T.*
812 *sinense* and *V. vinifera*) to further confirmed these connection patterns
813 (Supplementary Figs. 3.18-3.21).

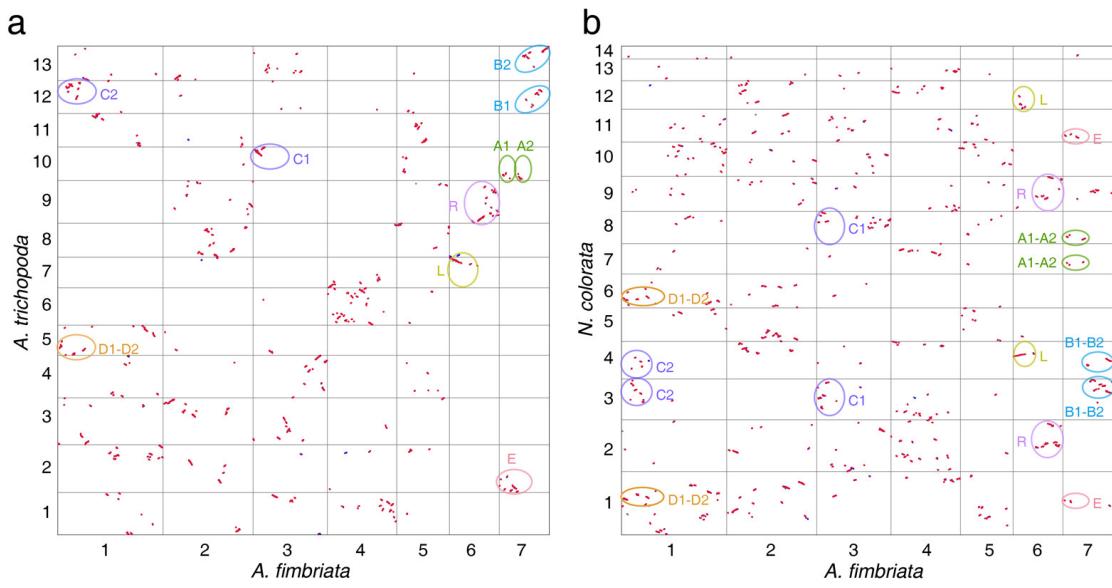
814 In *C. demersum*, we found a few signals supporting the integrated orthologous regions
815 of the *A. fimbriata* A1-A2-B1-B2 on chromosomes 2 and 5, and the integrated
816 orthologous regions of the *A. fimbriata* C2-D1-D2 on chromosomes 1 and 3

818 (Supplementary Fig. 3.22). Given the three lineage-specific WGDs in *Ceratophyllum*
819 and the followed extensive chromosome merges (currently with only 12
820 chromosomes), the identified connection pattern in *Ceratophyllum* could be recently
821 derived rather than inherited from the common ancestor of the major clades of
822 angiosperms. Therefore, we are not able to confidently suggest the chromosomal
823 rearrangement history in *Ceratophyllum* using current genome, and excluded this
824 lineage in the following investigations.

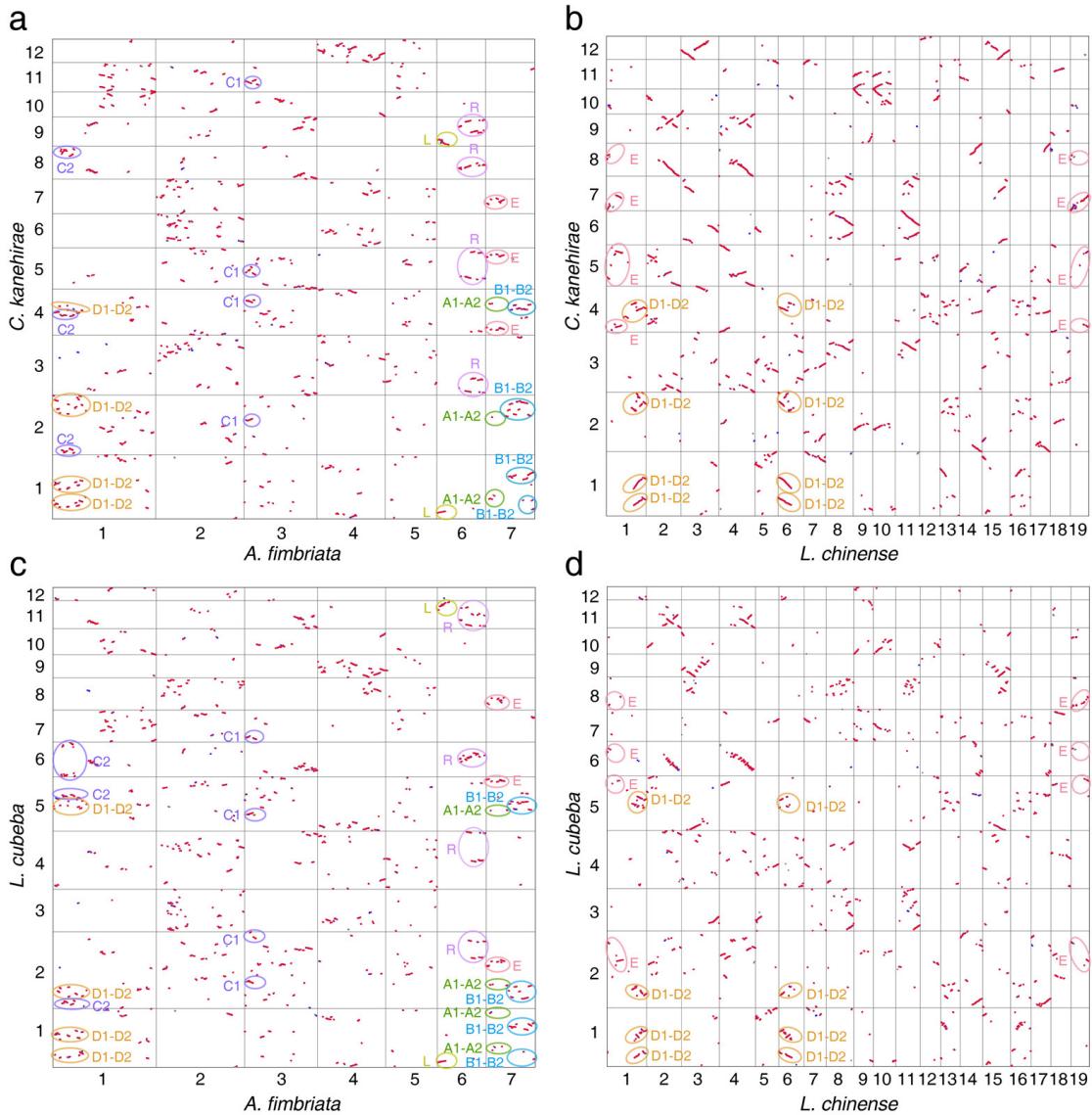
825
826 From the above analyses, the connection patterns of involved genomic regions in
827 most recent common ancestor of extant angiosperms could be inferred as (A1-A2-C1,
828 B1-B2-C2, D1-D2, and E) (Fig. 3c). We reconstructed the history of chromosomal
829 rearrangements following the divergence of extant angiosperms (Fig. 3c). After its
830 split from other angiosperms, the *Amborella* genome experienced a lineage-specific
831 genomic translocation event involving the B2 region that was separated from the
832 ancestral B1-B2-C2 region. In *N. colorata* genome, one set of the paralogous regions
833 (from the recent WGD) remains as the ancestral angiosperm connection pattern, while
834 the other set experienced a translocation event involving the C1 region that was
835 separated from the ancestral A1-A2-C1 region. In the common ancestor of eudicots,
836 the C2 region was split out of the B1-B2-C2 region, and connected with the D1-D2 to
837 form a D1-D2-C2 connection pattern. In the common ancestor of monocots, a
838 reciprocal genomic translocation occurred involving: two fissions resulted the split of
839 the C1 region out of the ancestral A1-A2-C1 region and the C2 region out of the
840 ancestral B1-B2-C2 region; two fusions resulted the fusion of the C1-C2 region and
841 the A1-A2-B1-B2 region. Notably, such scenario was also identified in magnoliids,
842 which may represent an important genomic synapomorphy between magnoliids and
843 monocots.

844 Therefore, our results suggested that the chromosomal rearrangement events we
845 detected here were shared by magnoliids and monocots, and likely not shared with
846 eudicots. It seems hard to completely exclude the possibility that ancient eudicots also
847 shared these rearrangements, but they were subsequently reversed these exchanges
848 soon after the divergence of eudicots from magnoliids and monocots. However, given
849 the similar genomic structure between eudicots and representatives of the ANA grade,
850 we speculate the chance of such reverse scenario should be very small, because both
851 of the fissions and fusions in the common ancestor of eudicots would have had to
852 occur at the previous break points and also in a very limited time frame. Thus, we
853 propose that the reciprocal genomic translocations only shared by magnoliids and
854 monocots may represent an important genomic synapomorphy and that magnoliids
855 might be sister to monocots while the clade of eudicots diverged earlier (Fig. 3c).

856



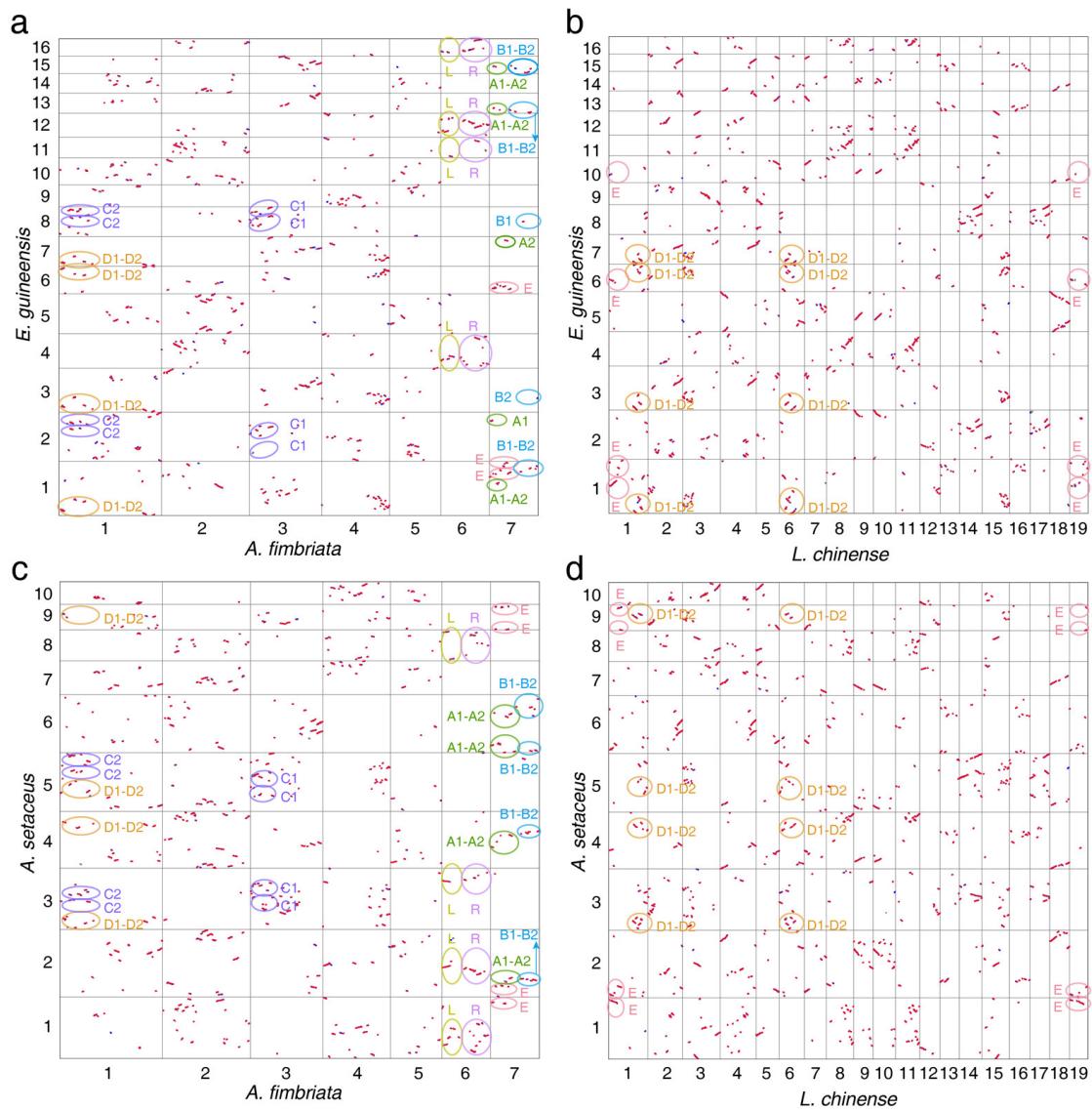
857 **Supplementary Fig. 3.10 | Genomic comparisons of the *A. fimbriata* with the *A. trichopoda* and *N. colorata* genomes.** (a) Syntenic dotplot between the *A. fimbriata* and the *A. trichopoda* genomes. (b) Syntenic dotplot between the *A. fimbriata* and *N. colorata* genomes. Genomic syntentic blocks (≥ 5 anchor gene pairs) were shown in dotplots according to their genomic locations in *A. fimbriata*, *A. trichopoda*, and *N. colorata*. If the anchor gene pairs are the best BLAST hit among the genomes, they are plotted as red dots; otherwise, they are shown in blue dots. The circled blocks indicate the syntentic blocks related to the rearranged genomic regions associated with the Chr6 and Chr7 in *A. fimbriata*, and they are marked with A1, A2, B1, B2, C1, C2, D1, D2, E, L, and R according to the defined regions in *A. fimbriata*.



867

868 **Supplementary Fig. 3.11 | Genomic comparisons of the *A. fimbriata* and *L.*
869 *chinense* with the *C. kanehirae* and *L. cubeba* genomes, respectively.** (a) Syntetic
870 dotplot between the *A. fimbriata* and *C. kanehirae* genomes. (b) Syntetic dotplot
871 between the *L. chinense* and *C. kanehirae* genomes. (c) Syntetic dotplot between the
872 *A. fimbriata* and *L. cubeba* genomes. (d) Syntetic dotplot between the *L. chinense* and
873 *L. cubeba* genomes. The orthologous region of the D1-D2 and E in *C. kanehirae* and
874 *L. cubeba* could be further verified by the syntetic relationship to the corresponding
875 D1-D2 and E regions in *L. chinense*.

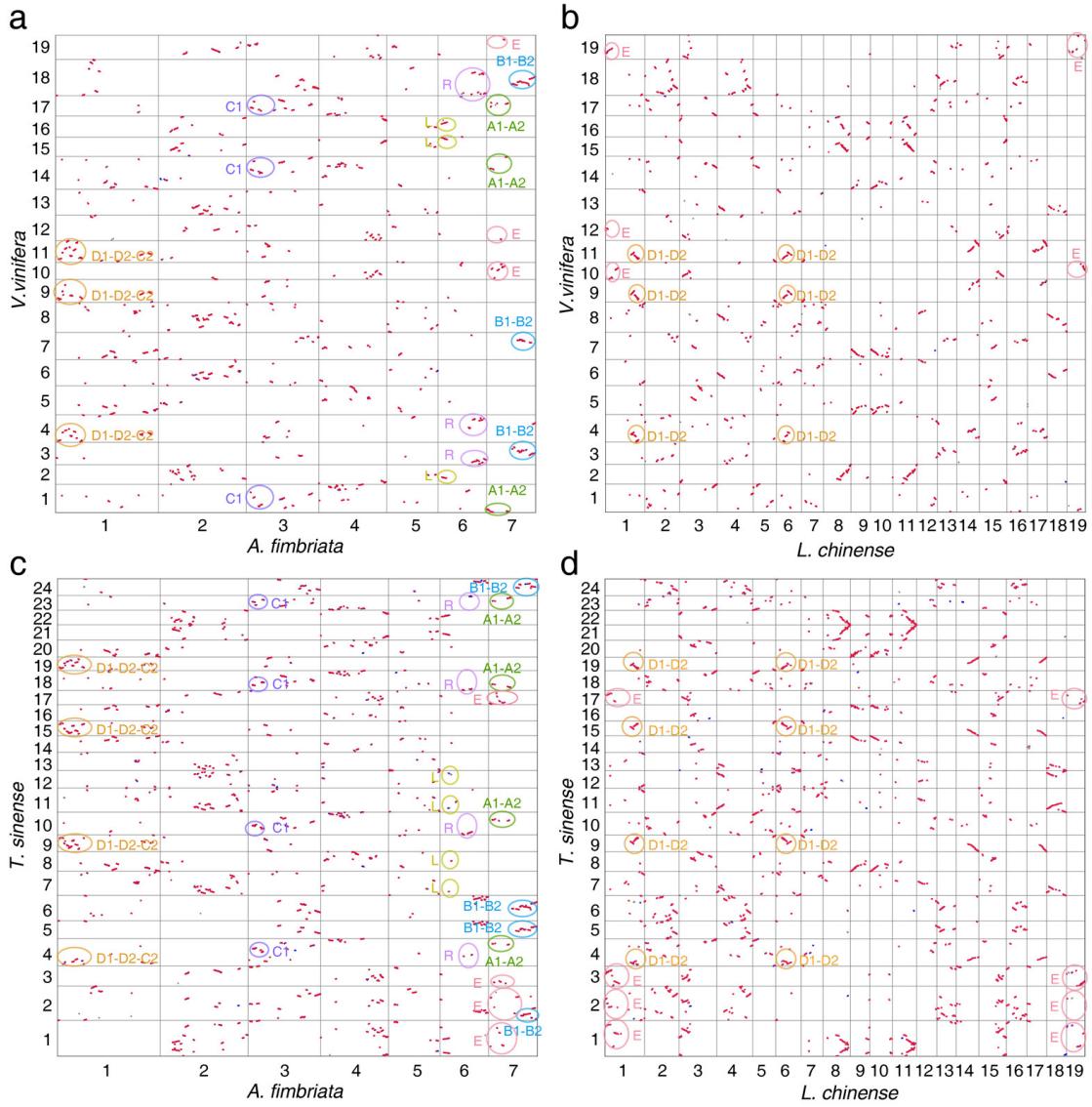
876



885

877 **Supplementary Fig. 3.12 | Genomic comparisons of the *A. fimbriata* and *L.*
878 *chinense* with the *E. guineensis* and *A. setaceus* genomes.** (a) Syntenic dotplot
879 between the *A. fimbriata* and *E. guineensis* genomes. (b) Syntenic dotplot between the
880 *L. chinense* and *E. guineensis* genomes. (c) Syntenic dotplot between the *A. fimbriata*
881 and *A. setaceus* genomes. (d) Syntenic dotplot between the *L. chinense* and *A.*
882 *setaceus* genomes. The orthologous region of the D1-D2 and E in *E. guineensis* and
883 *A. setaceus* could be further verified by the synteny relationship to the corresponding
884 D1-D2 and E regions in *L. chinense*.

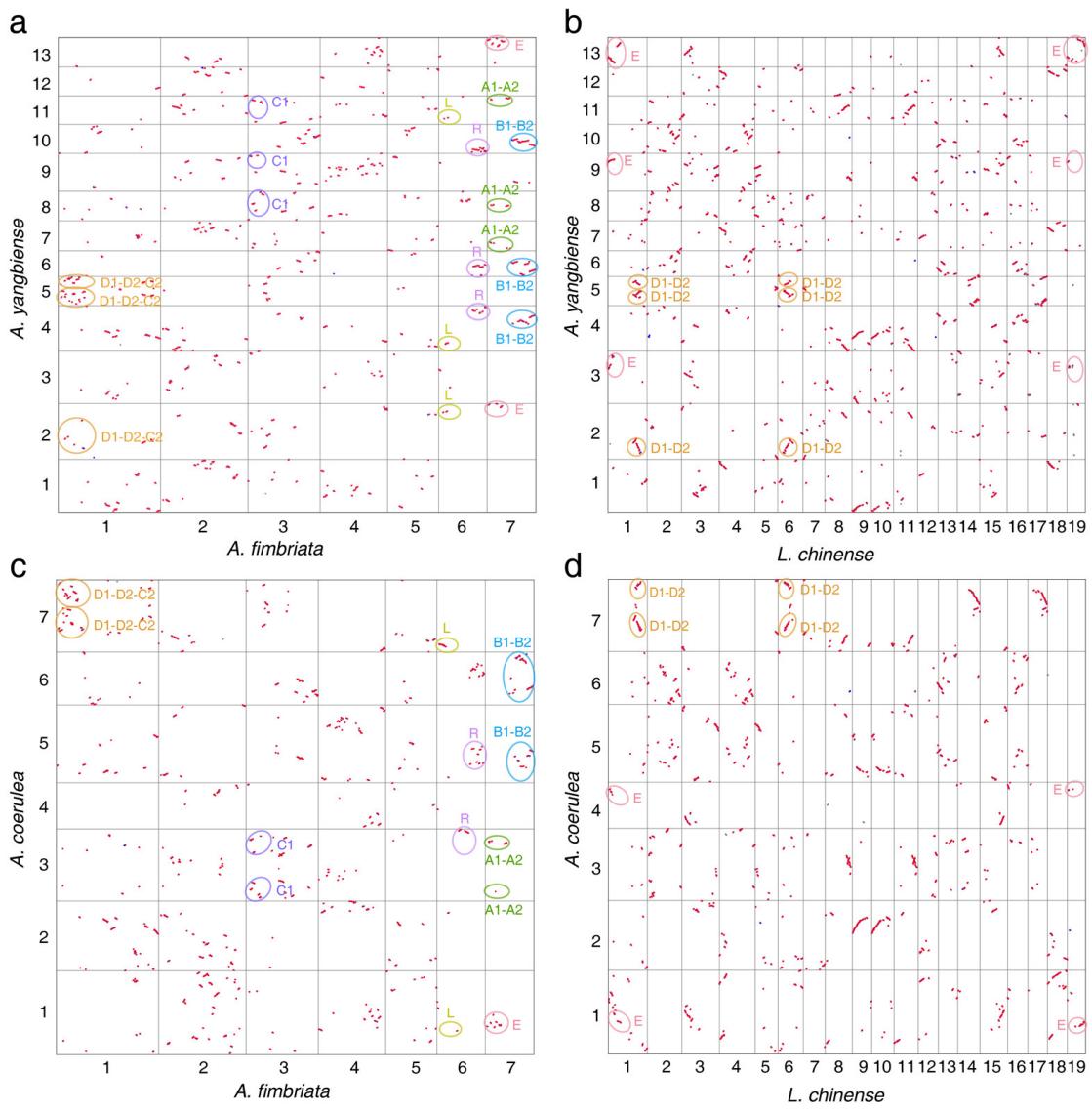
886



887

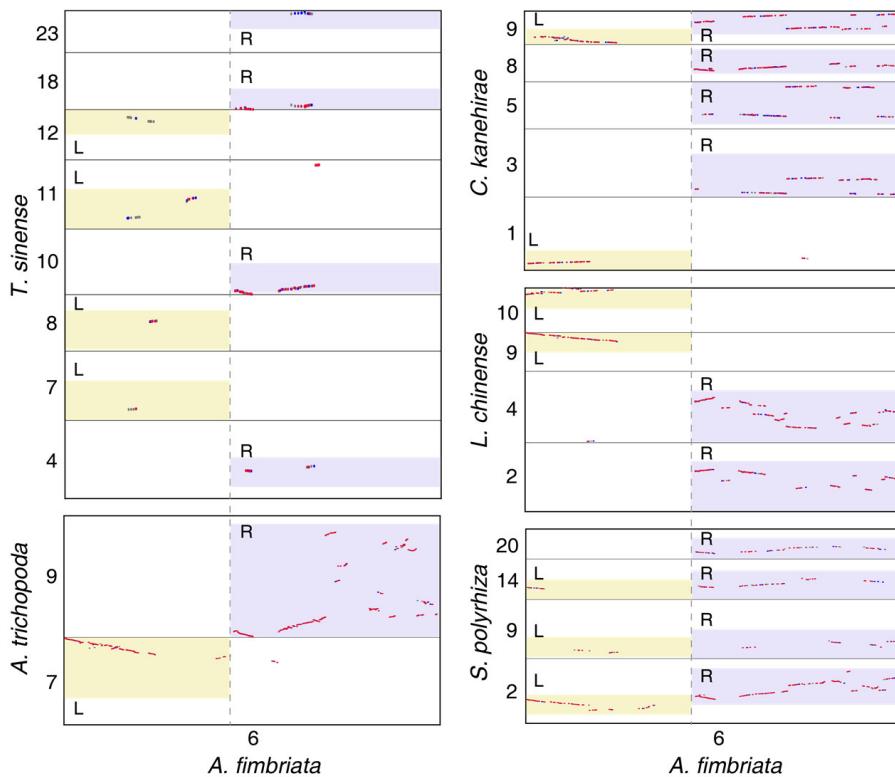
Supplementary Fig. 3.13 | Genomic comparisons of the *A. fimbriata* and *L. chinense* with the *V. vinifera* and *T. sinense* genomes. (a) Syntenic dotplot between the *A. fimbriata* and *V. vinifera* genomes. (b) Syntenic dotplot between the *L. chinense* and *V. vinifera* genomes. (c) Syntenic dotplot between the *A. fimbriata* and *T. sinense* genomes. (d) Syntenic dotplot between the *L. chinense* and *T. sinense* genomes. The orthologous region of the D1-D2 and E in *V. vinifera* and *T. sinense* could be further verified by the syntenic relationship to the corresponding D1-D2 and E regions in *L. chinense*.

895



896 **Supplementary Fig. 3.14 | Genomic comparisons of the *A. fimbriata* and *L.***
 897 ***chinense* with the *A. yangbiense*, and *A. coerulea* genomes.** (a) Syntenic dotplot
 898 between the *A. fimbriata* and *A. yangbiense* genomes. (b) Syntenic dotplot between the
 899 *L. chinense* and *A. yangbiense* genomes. (c) Syntenic dotplot between the *A. fimbriata*
 900 and *A. coerulea* genomes. (d) Syntenic dotplot between the *L. chinense* and *A. coerulea*
 901 genomes. The orthologous region of the D1-D2 and E in *A. yangbiense*, and *A. coerulea*
 902 could be further verified by the syntenic relationship to the corresponding D1-D2 and
 903 E regions in *L. chinense*.

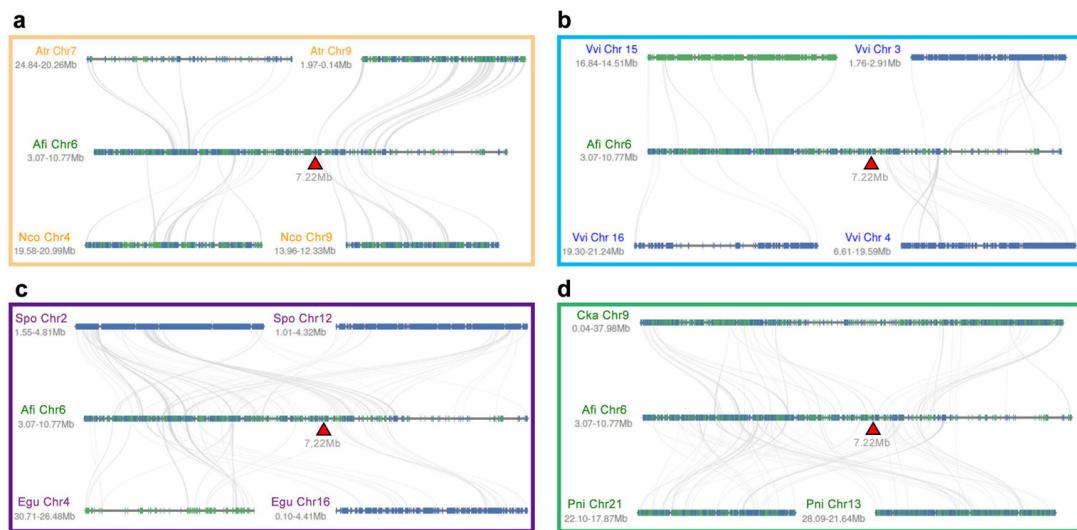
904



905

Supplementary Fig. 3.15 | Local synteny relationships among the selected genomic regions that associated with the structural rearrangements of *A. fimbriata* chromosome 6. The local synteny blocks identified between the *A. fimbriata* genome and the genomes of *A. trichopoda*, *T. sinense*, *S. polyrhiza*, *L. chinense*, and *C. kanehirae*. These synteny blocks were highlighted and named according to the defined regions in *A. fimbriata*, similar to the Supplementary Fig. 3.10.

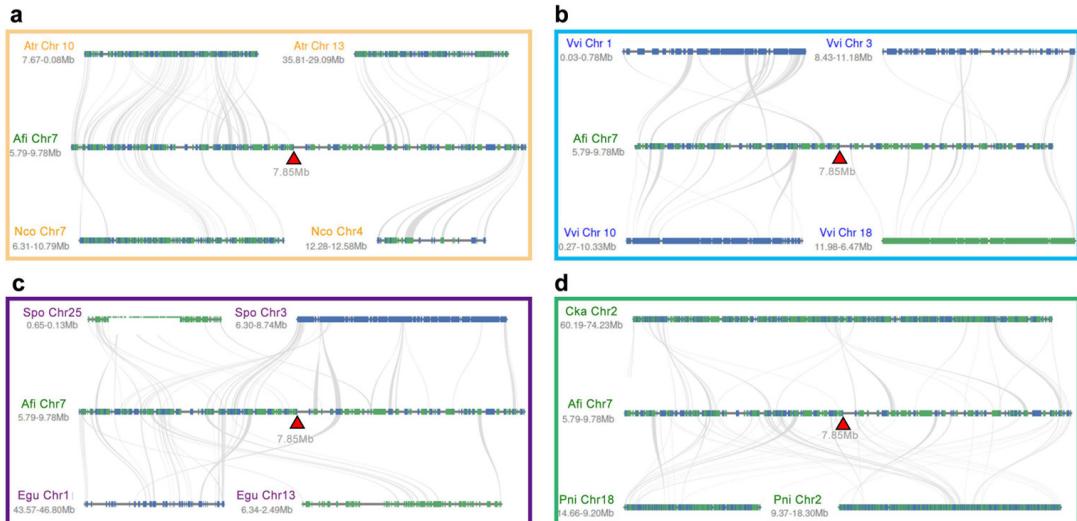
912



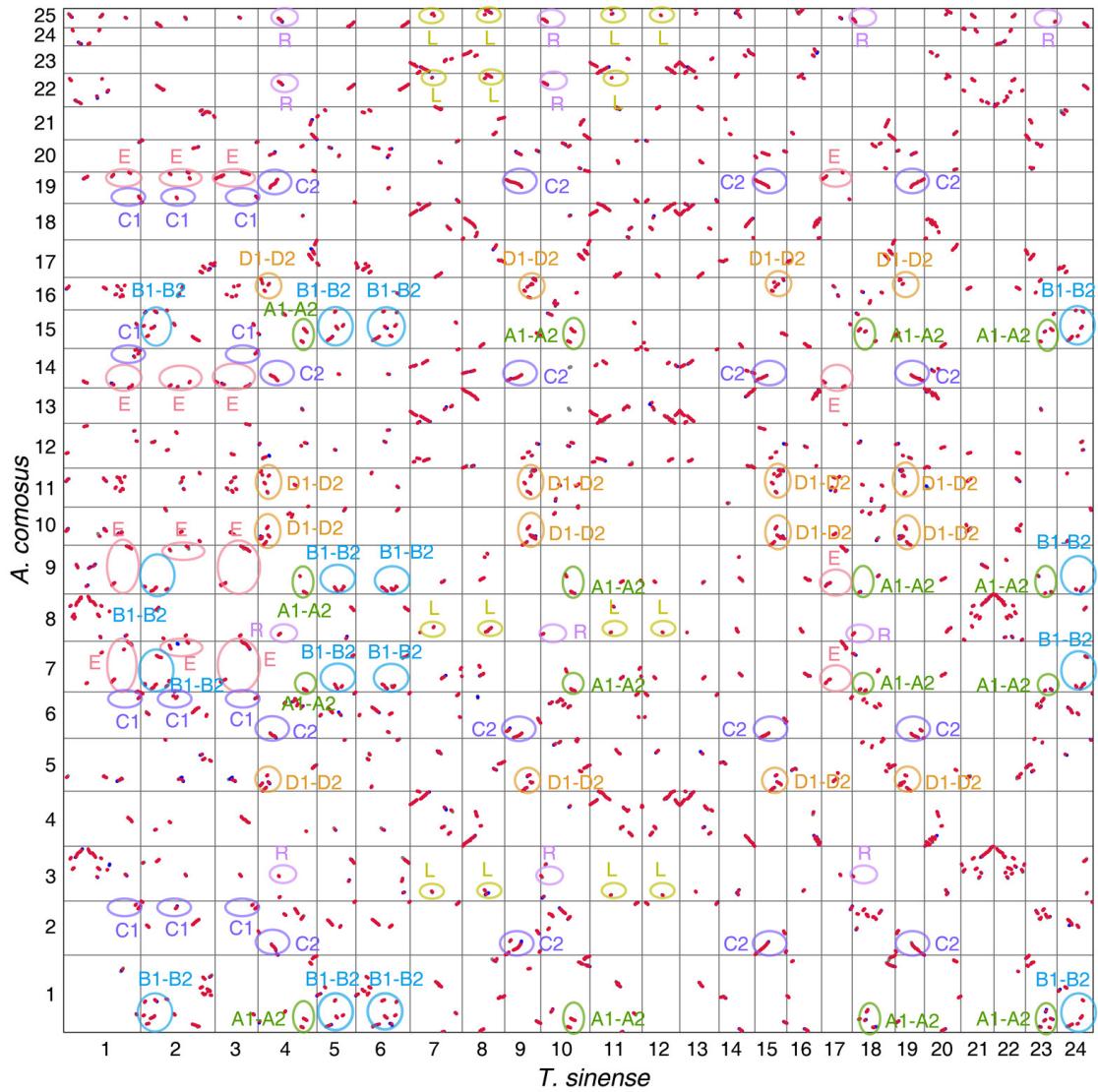
913

Supplementary Fig. 3.16 | Intergenomic microsynteny using the chromosome 6 of *A. fimbriata* as a reference. Syntenic relationships flanking the fusion point of

915 chromosome 6 of *A. fimbriata* (Afi) when comparing to the other genomes including
 916 (a) *A. trichopoda* (Atr) and *N. colorata* (Nco), (b) *V. vinifera* (Vvi), (c) *S. polyrhiza*
 917 (Spo) and *E. guineensis* (Egu), and (d) *P. nigrum* (Pni) and *C. kanehirae* (Cka).
 918 Rectangles represent annotated genes, and the grey lines connect syntenic gene pairs.
 919 The genomic fusion point is likely in the region around the locus indicated by the red
 920 triangle.

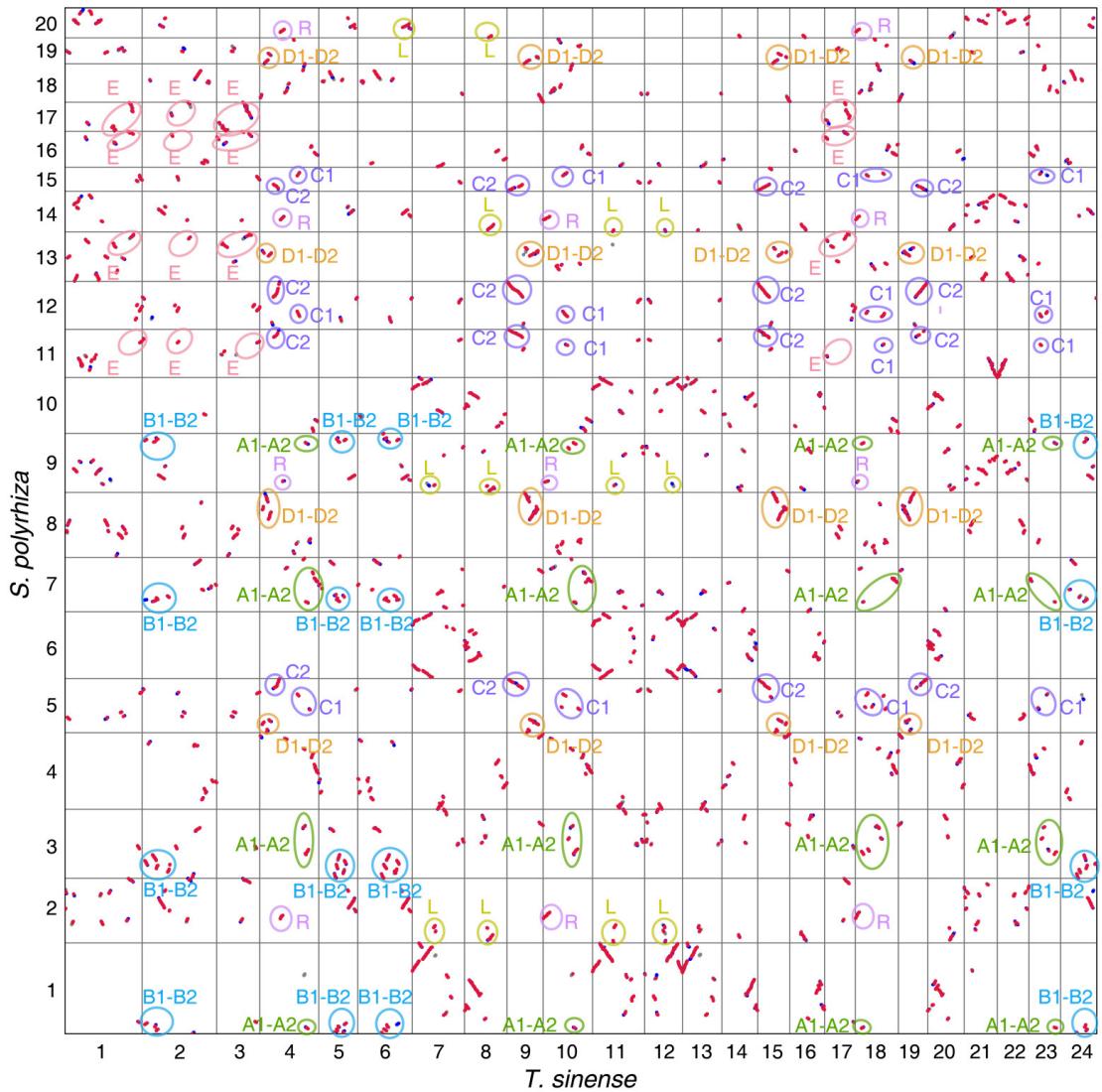


921
 922 **Supplementary Fig. 3.17 | Intergenomic microsynteny using chromosome 7 of *A.***
 923 ***fimbriata* as a reference.** Syntenic relationships flanking the fusion point of
 924 chromosome 7 of *A. fimbriata* (Afi) when compared to the other genomes including
 925 (a) *A. trichopoda* (Atr) and *N. colorata* (Nco), (b) *V. vinifera* (Vvi), (c) *S. polyrhiza*
 926 (Spo) and *E. guineensis* (Egu), and (d) *P. nigrum* (Pni) and *C. kanehirae* (Cka).
 927 Rectangles represent annotated genes, and the grey lines connect syntenic gene pairs.
 928 The genomic fusion point is likely in the region around the locus indicated by the red
 929 triangle.



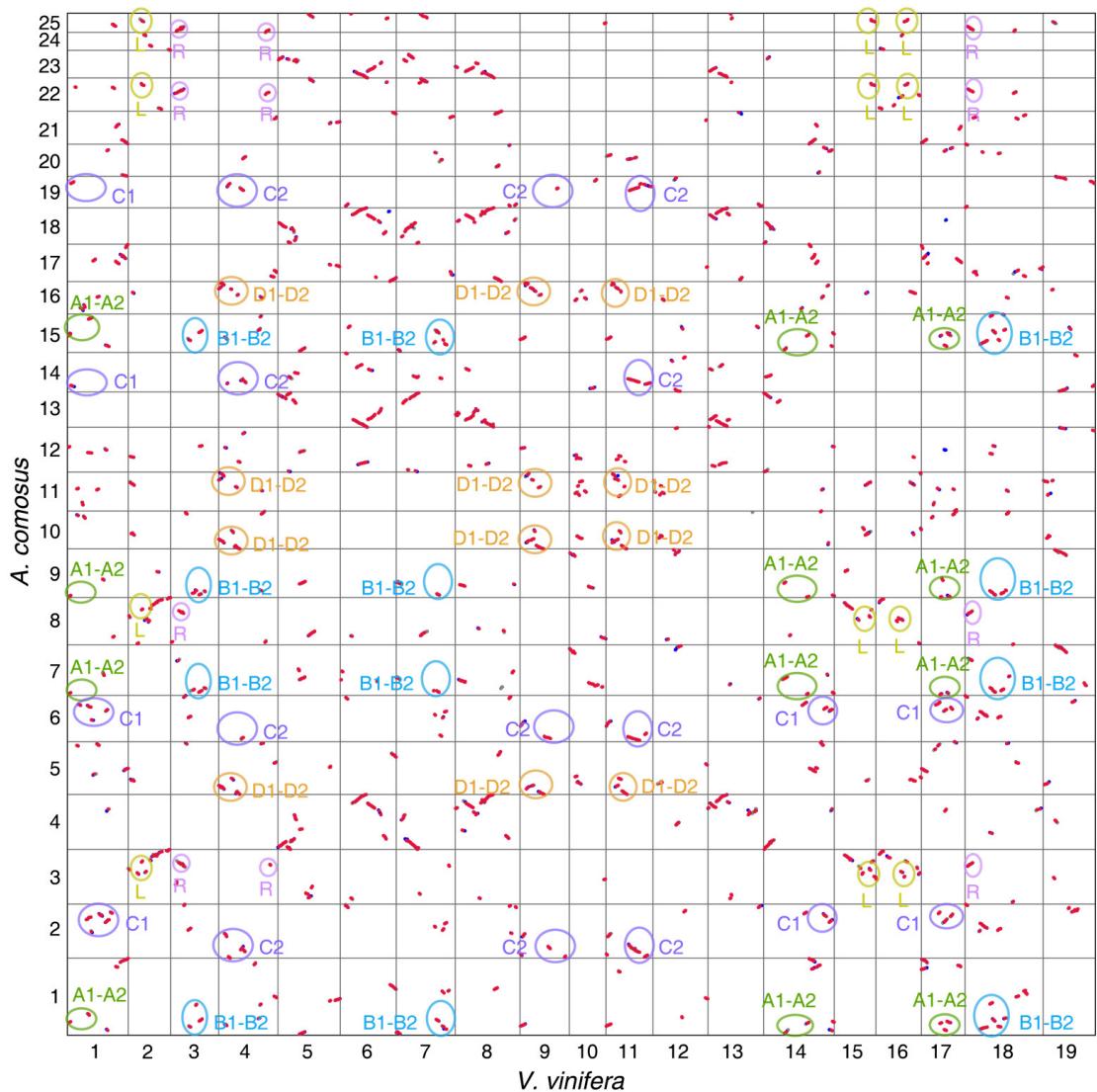
930

931 **Supplementary Fig. 3.18 | Genomic comparison of the *A. comosus* with the *T.***
 932 ***sinense* genome.**



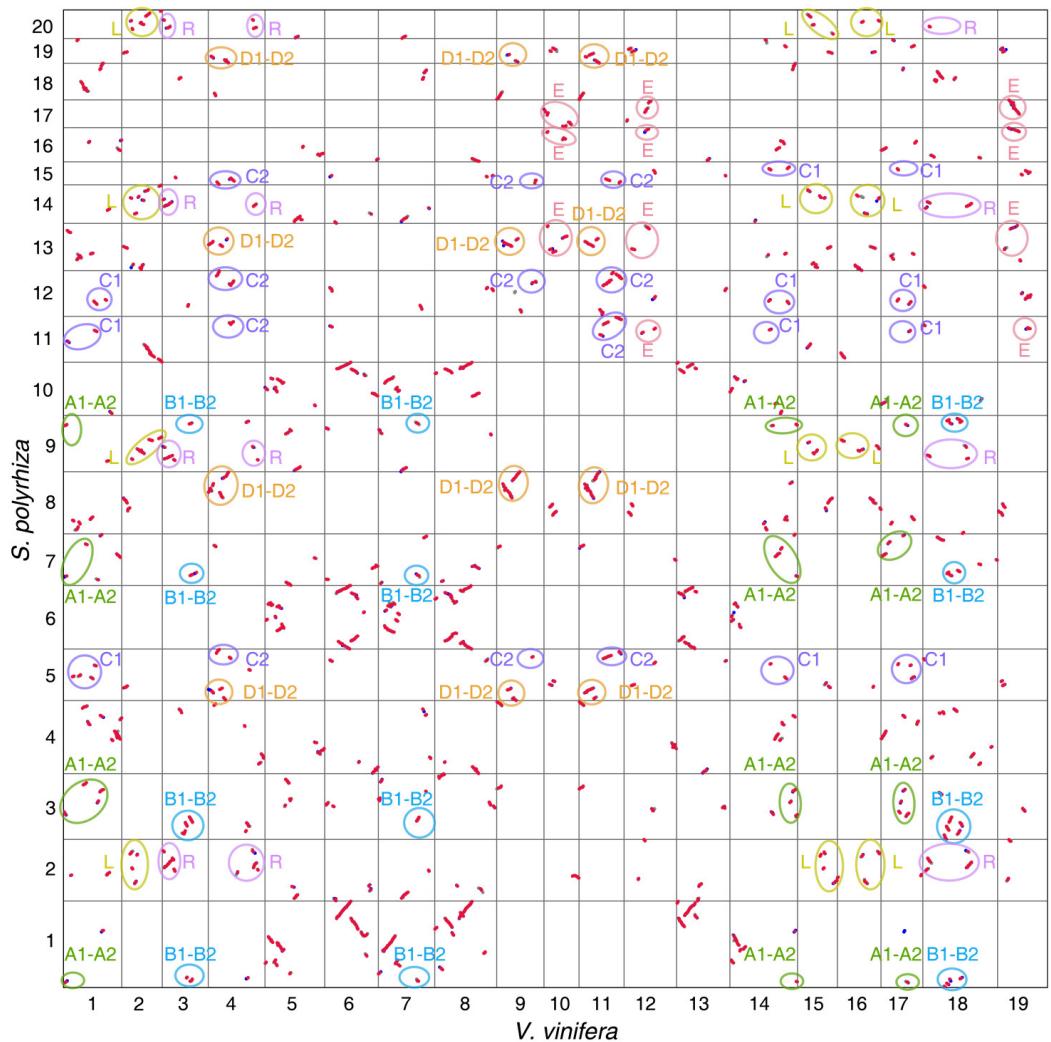
933

934 **Supplementary Fig. 3.19 | Genomic comparison of the *S. polystachya* with the *T. sinense* genome.**



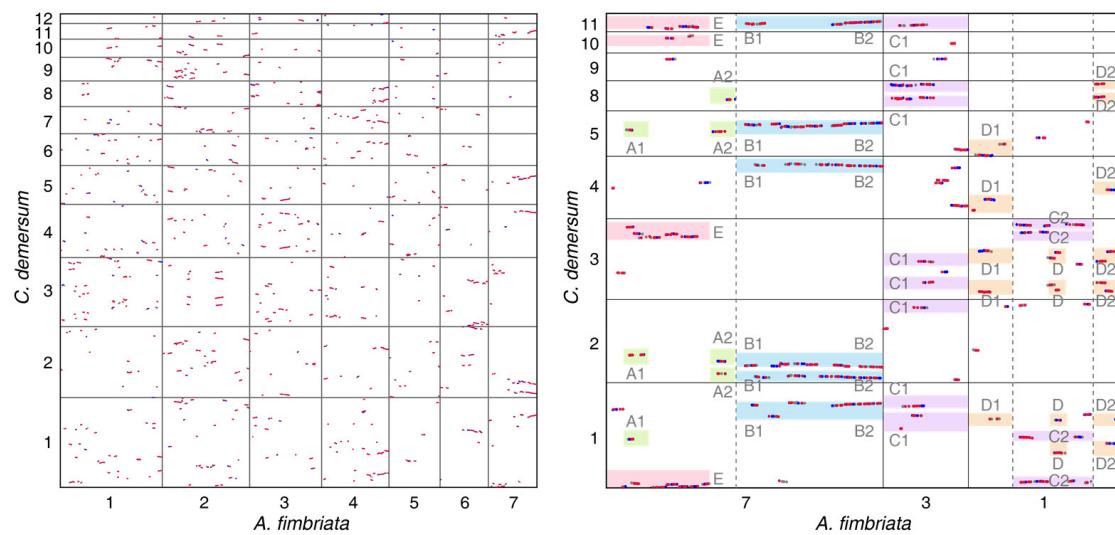
936

937 **Supplementary Fig. 3.20 | Genomic comparison of the *A. comosus* with the *V.***
 938 ***vinifera* genome.**



939

940 **Supplementary Fig. 3.21 | Genomic comparison of the *S. polystachya* with the *V.***
 941 ***vinifera* genome.**



942

943 **Supplementary Fig. 3.22 | Genomic comparison of the *A. fimbriata* with the *C.***
 944 ***demersum* genome. (a) Syntenic dotplot between the *A. fimbriata* and the *C.***

945 *demersum* genome. (b) Local syntenic relationships among the selected genomic
946 regions that associated with the structural rearrangements of *A. fimbriata* chromosome
947 7. Similar to the Fig. 3a, the specific genomic regions associated with the *A. fimbriata*
948 chromosome 7 fusion were named regions of E, A1, A2, B1, B2, C1, C2, D, D1 and
949 D2 as marked on top of the plot.

950 **4 Phylogenomic investigation of the relationship among magnoliids, monocots,
951 and eudicots**

952 **4.1 Phylogenomic analyses of strictly and mostly single-copy gene families**

953 **Materials and Methods**

954 In order to clarify the phylogenetic position of magnoliids relative to eudicots and
955 monocots, comprehensive phylogenomic analyses were performed using different
956 datasets and approaches (Supplementary Table 4.1). Four main topologies were
957 defined here as T1: ((magnoliids, (monocots, eudicots)), outgroup), T2: ((monocots,
958 (eudicots, magnoliids)), outgroup), T3: ((eudicots, (monocots, magnoliids)),
959 outgroup), and T4: ((eudicots, monocots, magnoliids), outgroup). Two strategies were
960 used for screening OGs based on gene copy number: the strictly single-copy (SSC)
961 and mostly single-copy (MSC) gene families.

962 **Strictly and mostly single-copy gene family identification.** Because the genomes of
963 black pepper (*P. nigrum*) and opium poppy (*P. somniferum*) each experienced a very
964 recent WGD event^{66,80}, we allowed them two gene copies at most, and the other 20
965 species strictly a single gene. A custom Python script was used to extract SSC gene
966 families from the OrthoMCL results of 22 species (Supplementary Note 2.4). We also
967 extracted MSC gene families which permitted some species to lack the homologous
968 gene, but we still required the OGs with genes in at least 4 species (4/7) in eudicots, 3
969 species (3/5) in monocots, 3 species (3/5) in magnoliids, and 3 species (3/5) in
970 outgroups (including 2 representatives from ANA grade, 2 gymnosperms, and one
971 lycophyte).

972 **Phylogenetic analysis.** For each gene family, protein sequences were aligned using
973 MUSCLE v3.8.31⁸¹, and nucleotide sequences were then forced to fit the amino acid
974 alignments using PAL2NAL v14⁸². We also forced nucleotide sequences on the
975 amino acid alignments using a custom Python script to obtain codon-preserving
976 alignments of nucleotide sequences. Finally, we retrieved four different alignments
977 for each gene family: (1) amino acid (or peptide, pep) alignments; (2) nucleotide
978 alignment (nucleotides forced to the amino acid alignment; or coding sequence, cds);
979 (3) codon alignments with third-position removed (codon1&2); and (4) codon

980 alignments with first- and second-position removed (codon3). Here we specifically
981 investigated the phylogeny using the alignment from codon3, because previous report
982 showed apparently decreased supporting values for the sister relationship of
983 magnoliids and eudicots when using alignment of codon1&2 comparing to using the
984 cds alignment⁸³.

985 For the concatenation-based phylogeny, gene alignments were concatenated as a
986 single supermatrix, and the tree was inferred under the “PROTGAMMAAUTO” and
987 “GTRGAMMA” model of amino acid and nucleotide substitution using RAxML
988 v8.2.12⁸⁴. We also constructed individual gene trees by 100 rapid bootstrapping
989 replicates and searching for the best-scoring maximum likelihood (ML) tree in one
990 single run (-f a option). We checked the bootstrap support (BS) values for the nodes
991 associated with the phylogenetic relationship among monocots, eudicots, and
992 magnoliids, and summarized the topologies with BS value greater than or equal to
993 0%, 10%, 50%, or 80%, respectively. Coalescent-based phylogeny was inferred from
994 the individual ML gene trees with different BS cutoff values using ASTRAL-II
995 v5.5.11⁸⁵. The reliability of each internal branch in the coalescent-based species tree
996 was further evaluated using the local posterior probability (LPP) measurement. In
997 addition, we also used another coalescent-based phylogenetic method, MP-EST
998 (<https://github.com/liliu1871/mp-est>), to carry out additional phylogenetic analyses.

999 We selected 35,944 OGs with more than 4 members from OrthoMCL results, and
1000 used the above method to construct the gene tree. ASTRAL-Pro
1001 (<https://github.com/chaoszhang/A-pro>) and STAG
1002 (<https://github.com/davidemms/STAG>) method were used for species trees estimated
1003 from multi-copy gene families.

1004 **Assessing conflict at specific nodes.** To investigate the extent of incongruence that is
1005 present in the phylogenomic data matrix, we performed the following two
1006 assessments for ML trees based on amino acid and nucleotide sequences, respectively.
1007 First, we used phyparts v0.0.1⁸⁶ to count the number of genes supporting certain
1008 topologies. Secondly, we used built-in LPPs of ASTRAL to estimate branch support
1009 and to test for polytomies^{87,88}.

1010 **Results and Discussion**

1011 In total, we identified 98 SSC and 535 MSC gene families from the orthogroups of the
1012 selected 22 species (Supplementary Note 2.4). For the analyses using coding sequence
1013 (cds) and amino acid sequence (pep) of the 98 SSC gene families, no individual trees
1014 from cds/pep alignments supported magnoliids sister to a clade consisting of eudicots
1015 and monocots (T1); 8% of cds trees and 4% of pep trees supported magnoliids as the
1016 sister lineage to eudicots (T2); 3% of cds trees and 2% of pep trees supported

1017 magnoliids as sister lineage to monocots (T3); and the other trees (~90%) show no
1018 resolution (T4) (Fig. 4b and Supplementary Table 4.2). For the dataset 535 MSC gene
1019 families, 1% of the cds and pep trees supported that magnoliids are at the basal
1020 position of the angiosperms (T1); 5% of cds trees and 2% of pep trees supported
1021 magnoliids as sister lineage to eudicots (T2); 6% of cds trees and 2% of pep trees
1022 supported that magnoliids as sister lineage to monocots (T3); and the other trees
1023 (~90%) also show no resolution (T4) (Fig. 4b and Supplementary Table 4.2). In
1024 conclusion, most of the individual gene trees show weak support for relationships or
1025 no resolution regarding the phylogenetic relationship of magnoliids, monocots, and
1026 eudicots, resulting in the T4 topology (Fig. 4b and Supplementary Table 4.2).

1027 The concatenation-based phylogenetic inferences were performed for the 98 SSC and
1028 535 MSC gene families using nucleotide and amino acid sequences, respectively. The
1029 resulting phylogenetic trees all supported magnoliids and eudicots as sister lineages
1030 (T2; Supplementary Fig. 4.1), which is consistent with previous reports⁸⁹⁻⁹¹.

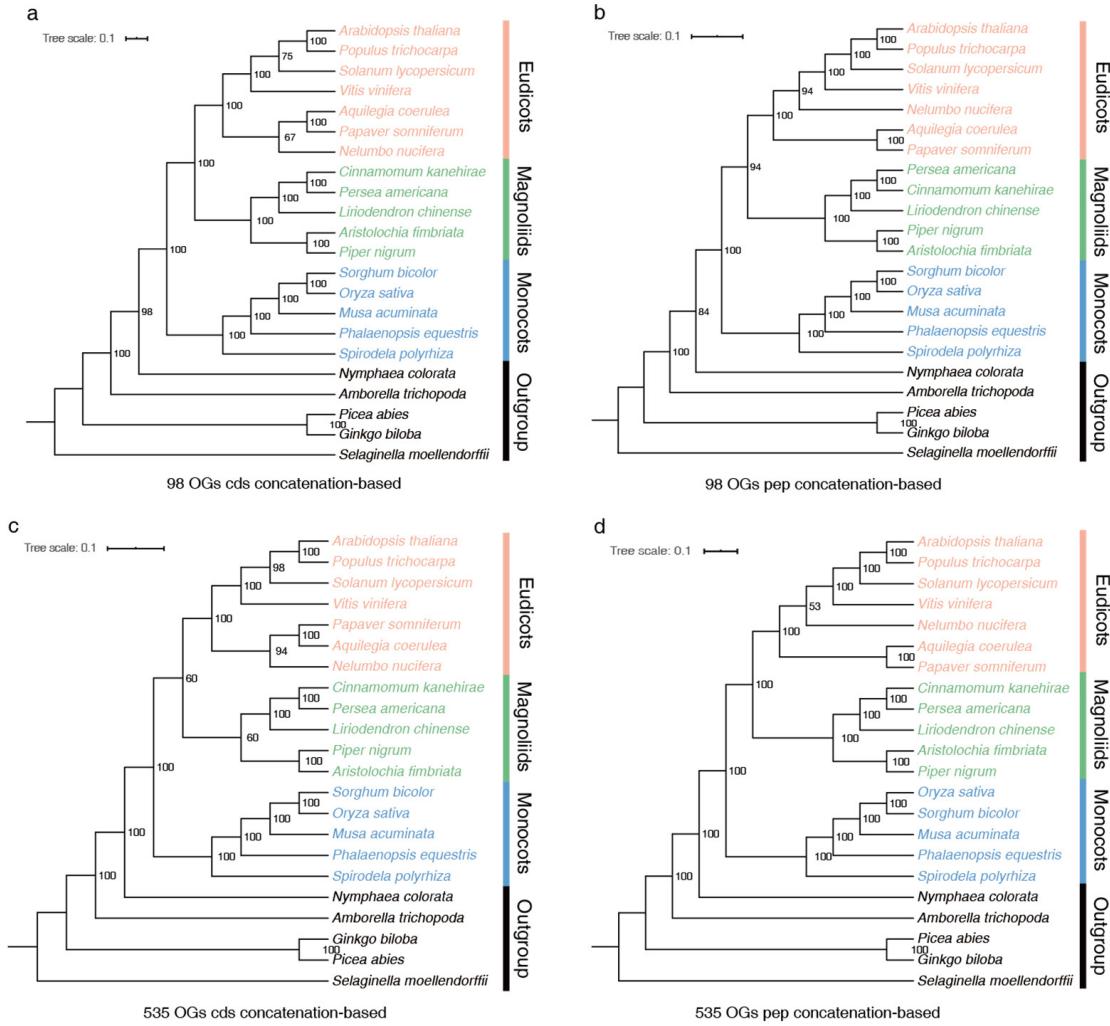
1031 Then, we performed ASTRAL phylogenetic analyses using different datasets of input
1032 gene trees which have different BS cutoff values for the nodes of the magnoliids,
1033 monocots, and eudicots lineages (BS \geq 0%, BS \geq 10%, BS \geq 50%, and BS \geq 80%).
1034 Coalescent trees using all 98 SSC seem to support T2 with magnoliids and eudicots as
1035 sister groups with low LPP (T2; details in <https://itol.embl.de/shared/Berlin>). In
1036 addition, we found that including the low BS input trees of the 535 MSC gene
1037 families resulted in T2 topology, but with low LPP support (Supplementary Fig.
1038 4.2a,b). Notably, if only inputting good support trees (BS \geq 50% or BS \geq 80%), we
1039 obtained phylogenies supporting T3 with high LPP values (Supplementary Fig.
1040 4.2c,d).

1041 We further performed the coalescent-based phylogenetic analyses inputting good
1042 support trees (the relevant node with BS \geq 50%) of the 98 SSC and 535 MSC datasets
1043 using nucleotide and amino acid sequences respectively. The 98 SSC dataset
1044 moderately supported the T2 topology, with magnoliids sister to eudicots
1045 (Supplementary Fig. 4.3a,b). However, the 535 MSC dataset strongly supported T3,
1046 with magnoliids sister to monocots (Supplementary Fig. 4.3c,d). We used the MP-
1047 EST method to analyze the amino acid and nucleotide sequence data of 535 MSC.
1048 Since the input file of MP-EST does not support gene trees for non-binary analysis,
1049 the bootstrap of gene trees is not collapsed. These results show that the amino acid
1050 sequence support magnoliids and monocots are sister groups (T3), but the nucleotide
1051 sequence support magnoliids and eudicots are sister groups (T2) (Supplementary Fig.
1052 4.4a,b). In addition, we use the bootstrap of RAxML to conduct bootstrap analysis on
1053 the MP-EST, and the results are consistent with the above results (Supplementary Fig.
1054 4.4c,d).

1055 In addition, we reconstructed phylogenies by partitioned codons employing both
1056 concatenation- and ASTRAL-based approaches. The resulted concatenation-based
1057 trees using the codon1&2 of the 98 SSC and 535 MSC gene families highly supported
1058 the magnoliids sister to eudicots (T2; Supplementary Fig. 4.5a,c), while using the
1059 codon3 resulted in low resolution and inconsistent topologies (Supplementary Fig.
1060 4.5b,d). Coalescent analyses from codon1&2 of the 98 SSC gene families supported
1061 T2 but with moderate resolution (Supplementary Fig. 4.6a). Notably, the coalescent
1062 analyses using codon1&2 from the 535 MSC gene families suggested the T3 with full
1063 support (Supplementary Fig. 4.6c). When using codon3, both 98 SSC and 535 MSC
1064 gene families moderately and highly supported T3, respectively, while the placement
1065 of *S. moellendorffii* within the monocots seems to be incorrect (Supplementary Fig.
1066 4.6b,d).

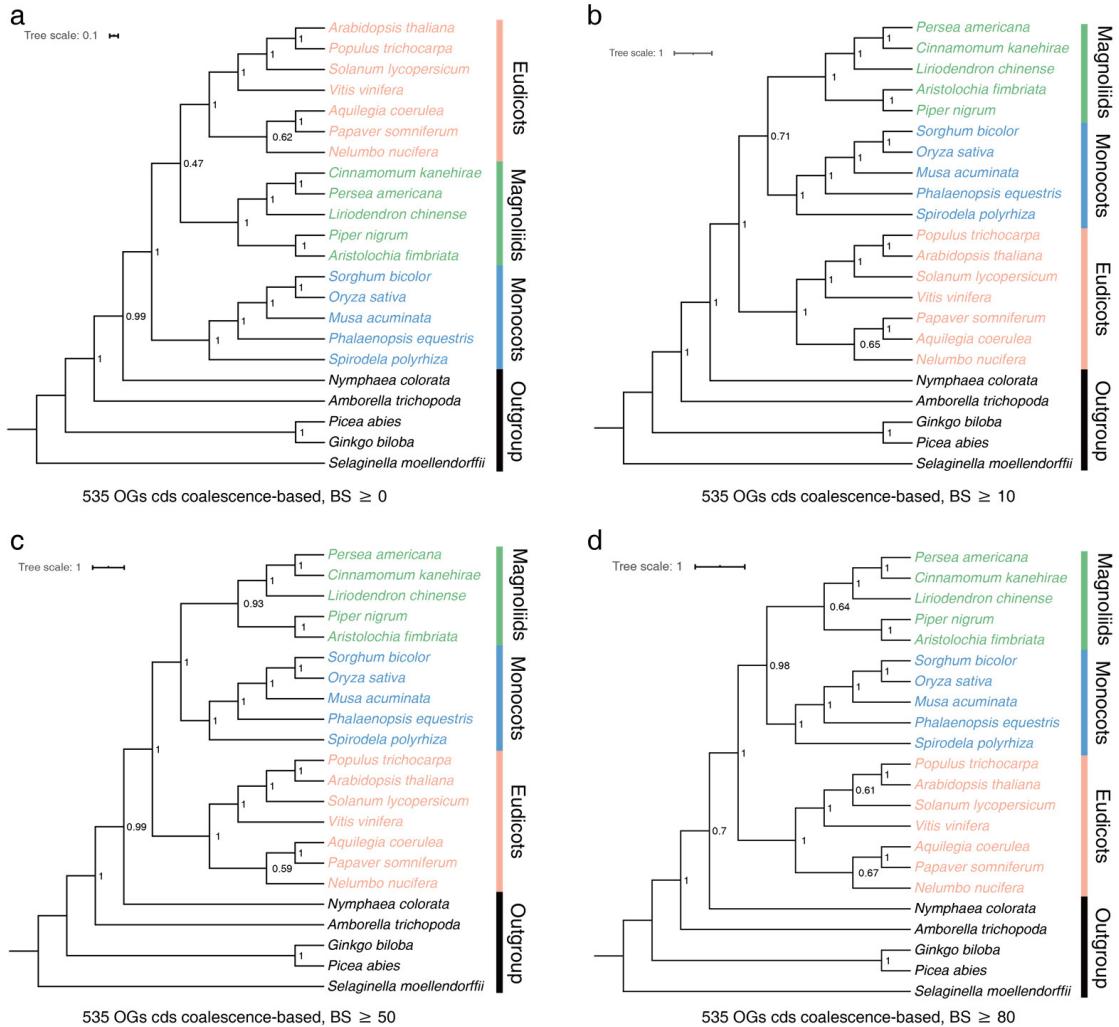
1067 We also explored the impact of incomplete lineage sorting (ILS) on the phylogenomic
1068 incongruence. The quartet score values were investigated for the variations among the
1069 gene trees and species trees. If we input gene trees directly, the resulted quartet
1070 frequencies are similar for the three topologies (Supplementary Figs. 4.7b-e and
1071 Extend Data Fig. 7a-d), which seems consistent to those of previous studies^{73,83}.
1072 However, if we collapse the tree nodes with BS value lower than 50%, the gene tree
1073 quartet frequencies of the 98 SSC gene families provide support for T2 with
1074 magnoliids as sister lineage to eudicots (Supplementary Fig. 4.7f-i and Supplementary
1075 Table 4.3). However, the quartet frequency analyses of the 535 MSC gene families
1076 with collapsed tree nodes support the T3 topology with magnoliids sister to monocots
1077 (Extended Data Fig. 7e-h and Supplementary Table 4.3). In addition, we used the SSC
1078 and MSC OGs for polytomy testing, which could not reject the null hypothesis that
1079 the node of magnoliids, eudicots, and monocots is a polytomy (T4) (Fig. 4 and
1080 Supplementary Table 4.3).

1081 SSC/MSC gene families greatly limit the availability of genomic data. In order to use
1082 more genomic data, we used multicopy gene tree summary methods ASTRAL-Pro
1083 and STAG to analyze the 22,563 OGs of 22 species (gene family members greater
1084 than 4). These results show that ASTRAL-Pro and STAG both support magnoliids
1085 and eudicots as sister groups (T2), but the support value of STAG is low (0.24)
1086 (Supplementary Fig. 4.8).



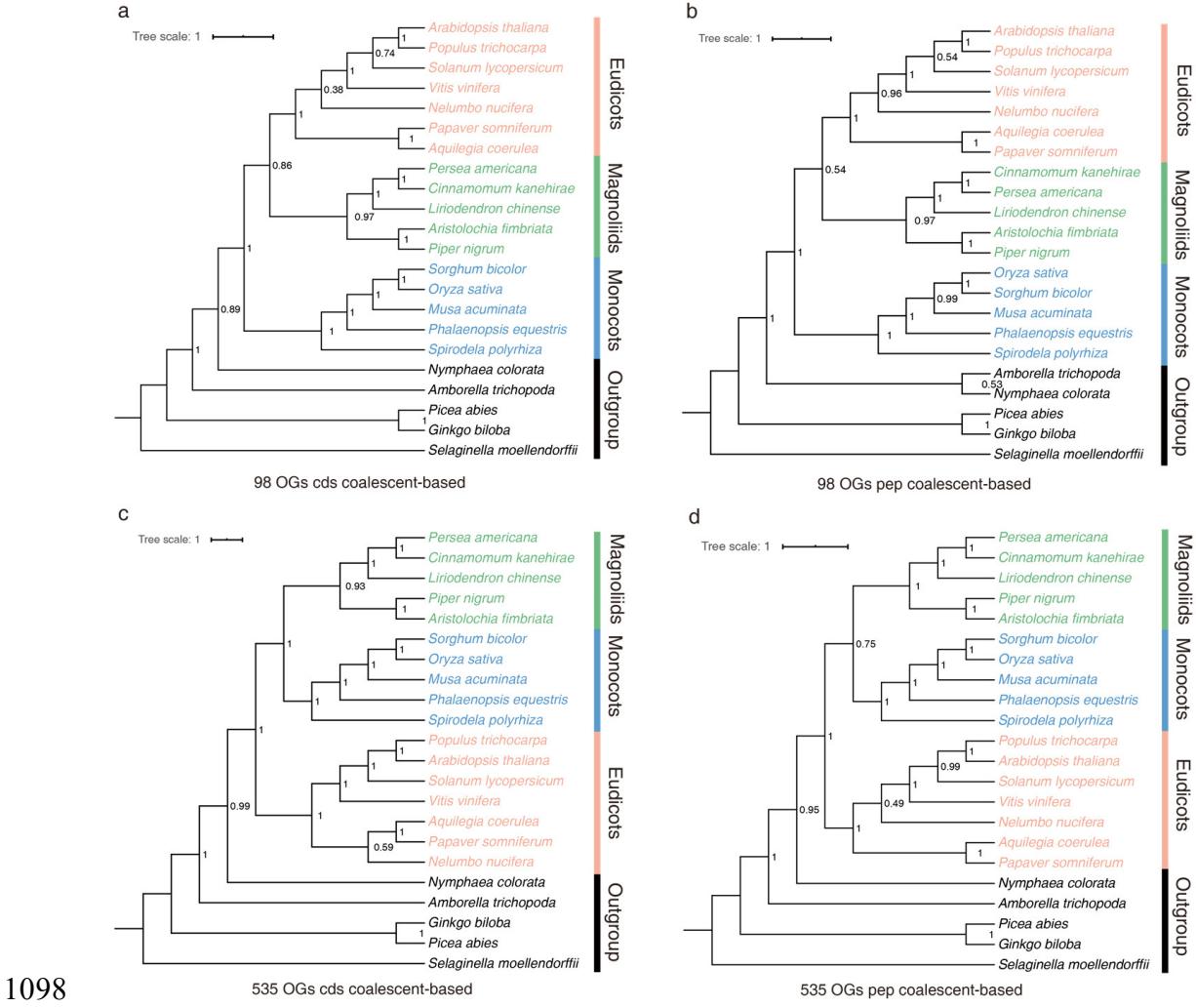
1087

1088 **Supplementary Fig. 4.1 | Phylogenetic relationships inferred from the 98 SSC**
1089 **and 535 MSC gene families using concatenation-based ML approaches.** Bootstrap
1090 support (BS) values for the concatenated-based analyses are shown on the branch to
1091 each node.



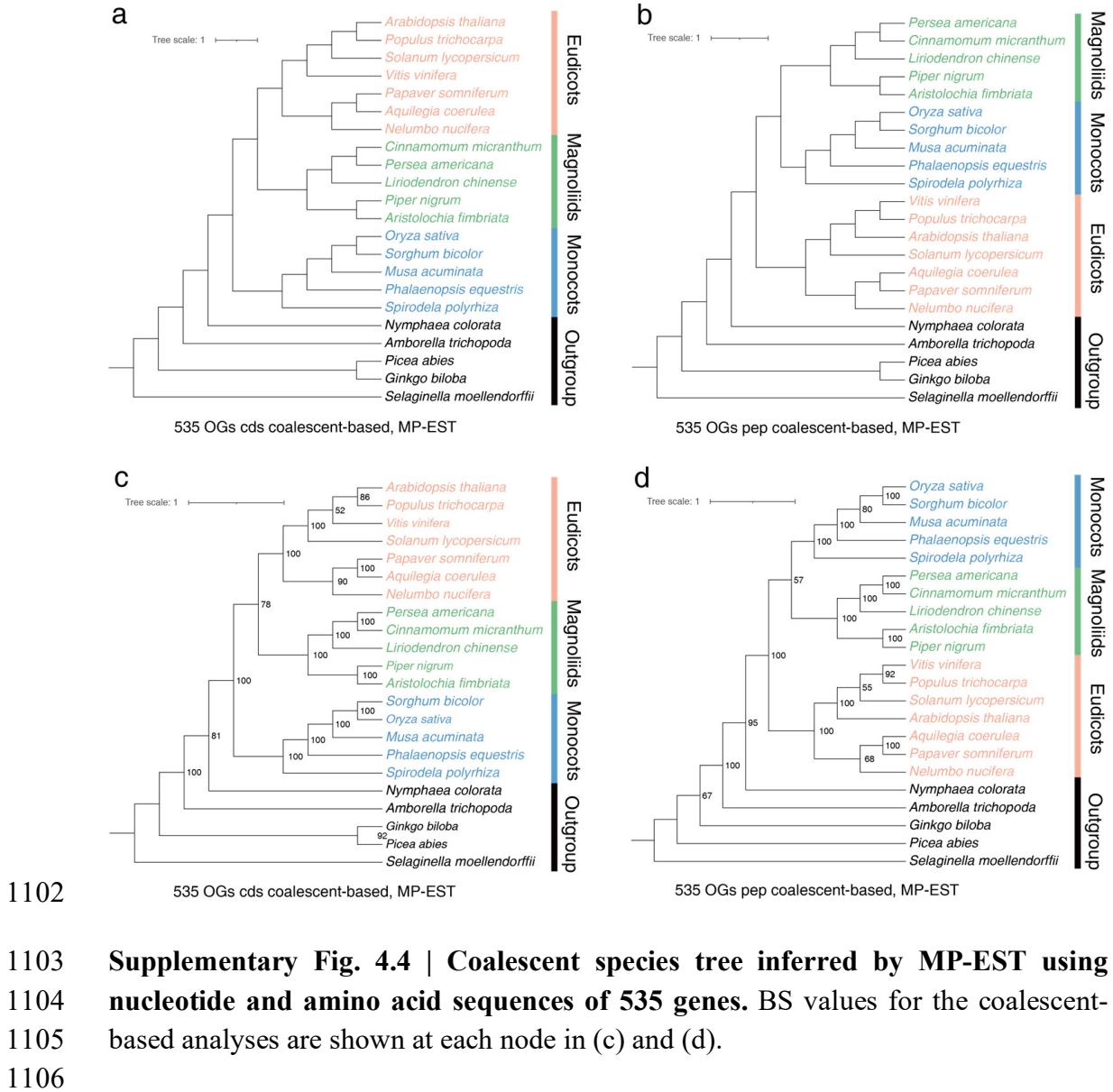
1092

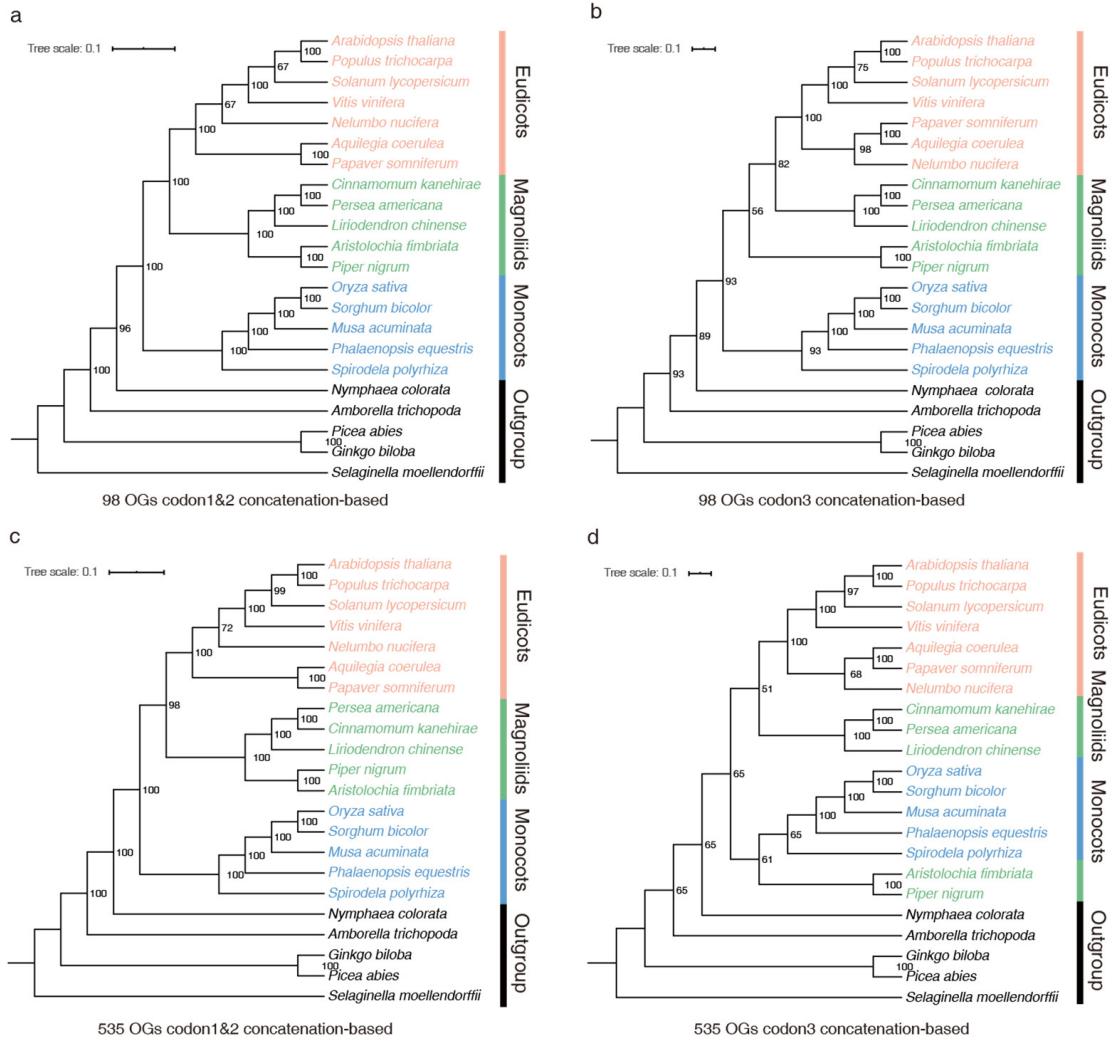
1093 **Supplementary Fig. 4.2 | Effect of BS values of the input individual gene trees for**
 1094 **coalescent-based phylogenetic analyses.** Numbers denote LPP values as
 1095 percentages. When including weakly supported input trees, the resulting phylogenies
 1096 supported T2 but with low support as in (a) and (b). When filtering out the low-BS
 1097 trees, the resulting phylogenies supported T3 with high LPP values as in (c) and (d).



1098

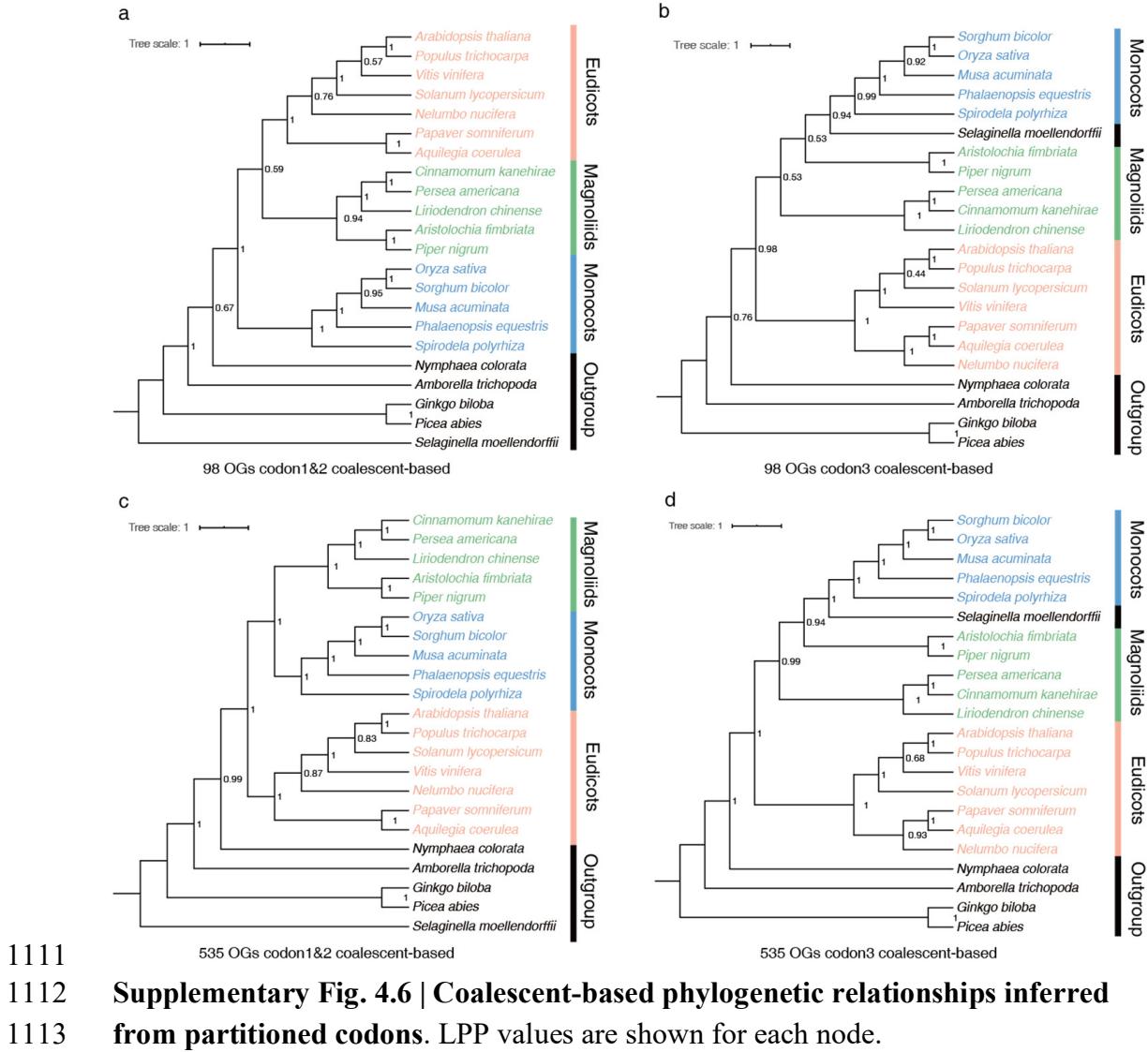
1099 **Supplementary Fig. 4.3 | Coalescent-based inferences using well-supported (the**
 1100 **relevant node with BS $\geq 50\%$) ML trees of the 98 SSC and 535 MSC gene families.**
 1101 **LPP values for the coalescent-based analyses are shown at each node.**

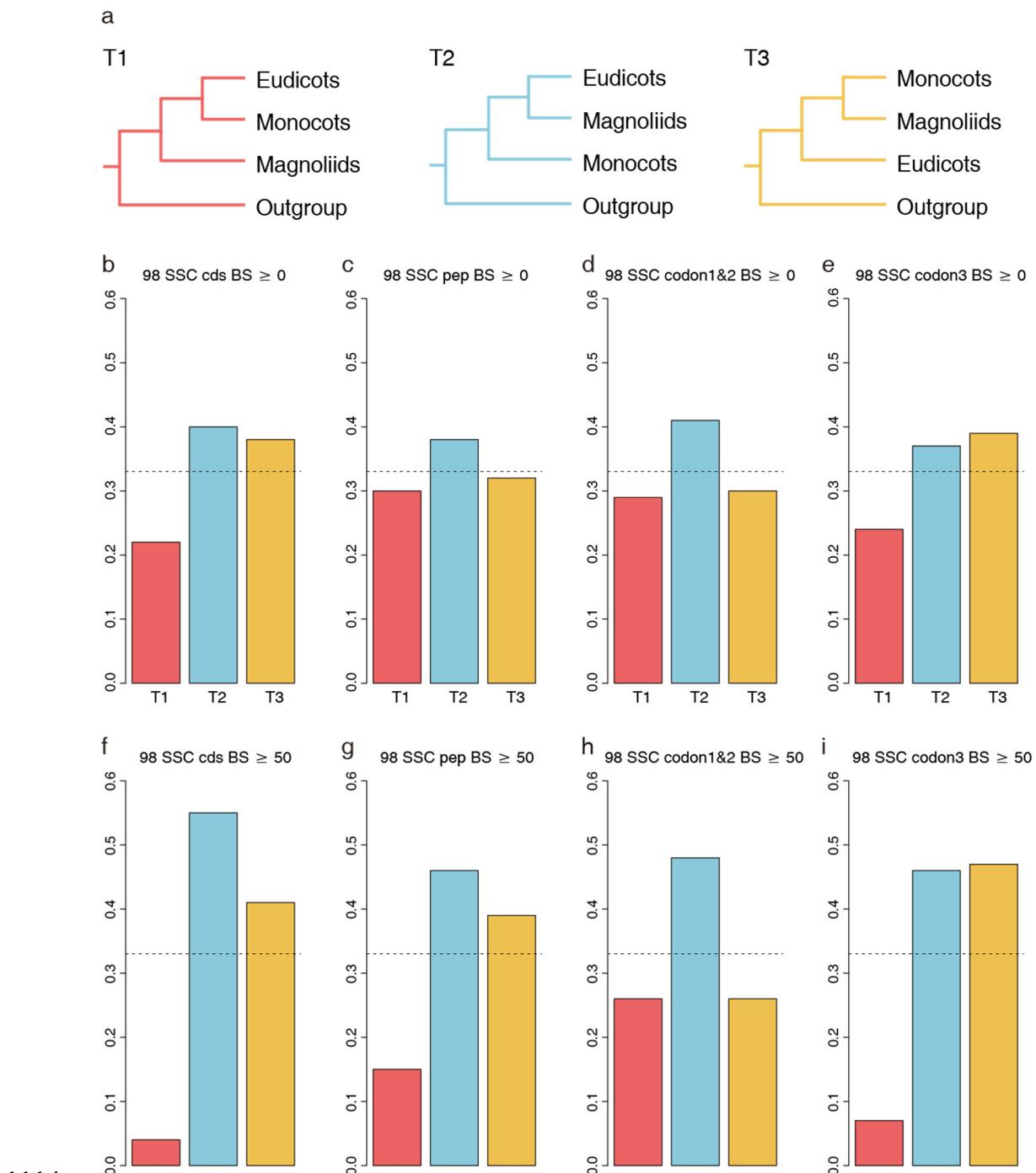


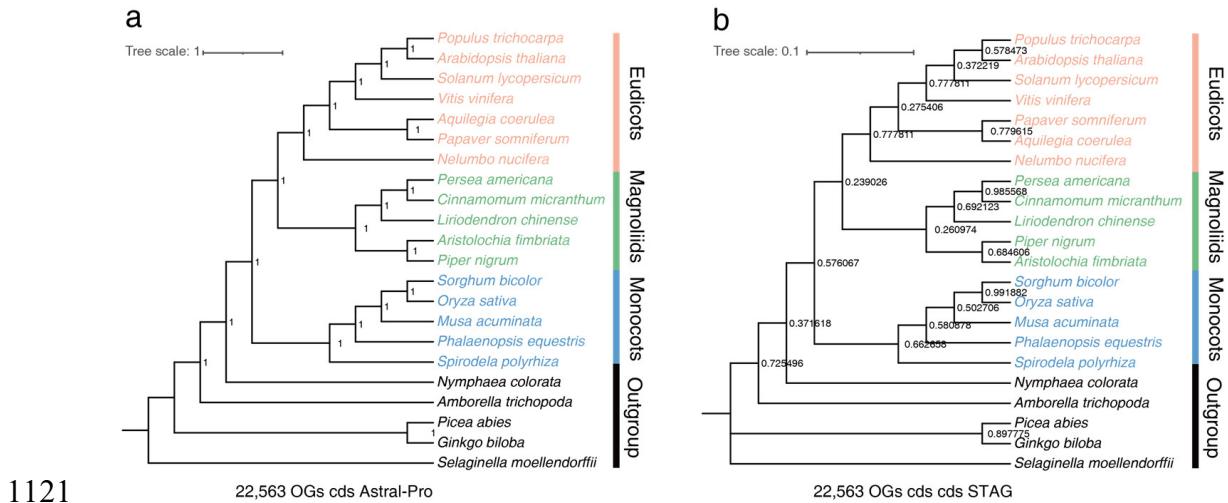


1107

1108 **Supplementary Fig. 4.5 | Concatenation-based phylogenetic relationships**
 1109 **inferred from the partitioned codons.** BS values for the concatenation-based
 1110 analysis are shown at each node.







1122 **Supplementary Fig. 4.8 | Species tree inferred by ASTRAL-Pro and STAG method using all gene family trees.** LPP and support values for the ASTRAL-Pro and STAG analysis are shown at each node.

1125 4.2 The impact of taxon sampling on the resulting phylogenetic topology

1126 Materials and Methods

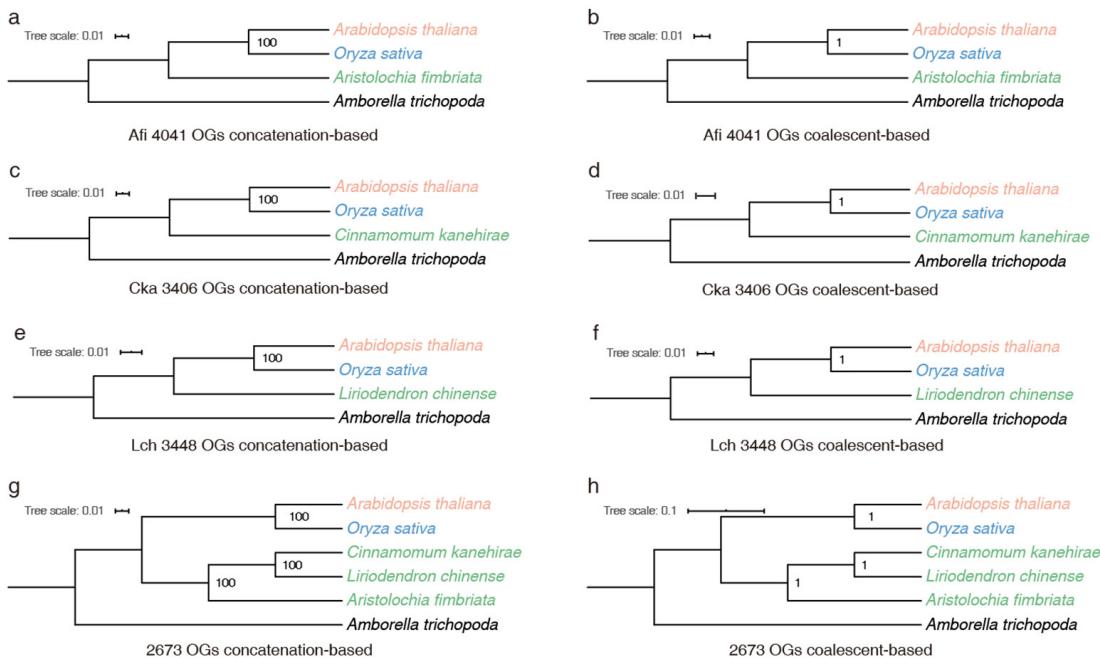
1127 To investigate the impact of taxon sampling on phylogenomic analyses, we
 1128 constructed datasets of (i) four species including one eudicot (*A. thaliana*), one
 1129 monocot (*O. sativa*), one angiosperm from the ANA grade (*A. trichopoda*), and one of
 1130 three magnoliid taxa (*L. chinense*, *C. kanehirae*, *A. fimbriata*), (ii) six species
 1131 including *A. thaliana*, *O. sativa*, *A. trichopoda*, and three magnoliid plants (*L.*
 1132 *chinense*, *C. kanehirae*, *A. fimbriata*), and (iii) 16 species including one angiosperm
 1133 from the ANA grade (*A. trichopoda*), seven eudicots (*A. thaliana*, *P. trichocarpa*, *S.*
 1134 *lycopersicum*, *V. vinifera*, *N. nucifera*, *A. coerulea*, and *P. somniferum*), five
 1135 monocots (*O. sativa*, *S. bicolor*, *M. acuminata*, *P. equestris*, *S. polyrhiza*), and three
 1136 magnoliids (*L. chinense*, *C. kanehirae*, *A. fimbriata*). Associated single-copy gene
 1137 families were extracted from OrthoMCL results by custom Python scripts, and the
 1138 concatenation- and coalescent-based phylogenetic analyses were performed. All of
 1139 these analyses were rooted with *Amborella*, which has been well-supported as the
 1140 single living sister lineage to all other extant angiosperms⁵⁵⁻⁵⁹.

1141 Results and Discussion

1142 We identified 4,041 SSC gene families for phylogenetic analysis of the four-species
 1143 dataset and found that approximately 40% (1,614 out of 4,041) of the individual trees
 1144 support the T1 topology with BS values greater than 50%, and 22%, 21%, and 18% of
 1145 the 4,041 individual gene trees supported T2, T3, and T4, respectively (Fig. 4b and

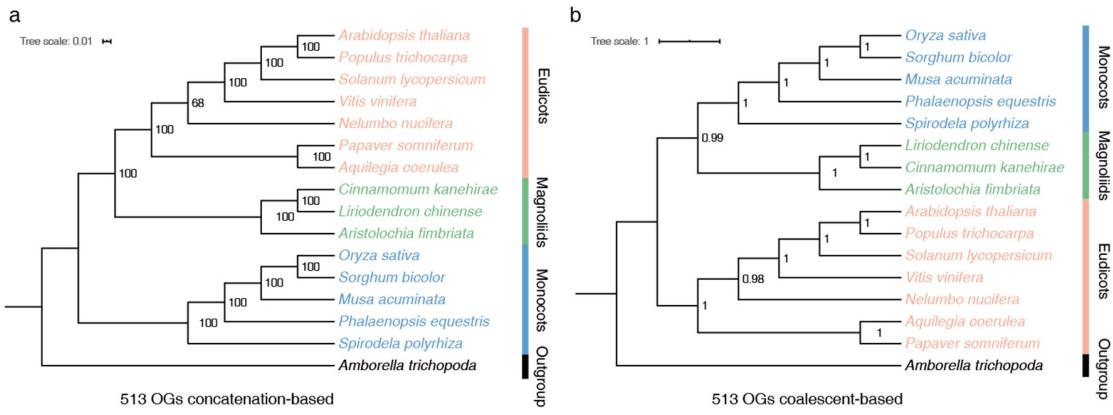
1146 Supplementary Table 4.2). The concatenation- and coalescent-based (BS \geqslant 50)
 1147 phylogenetic analyses all supported magnoliids as the sister lineage to the clade of
 1148 eudicots and monocots (Supplementary Fig. 4.9a,b). When replacing *A. fimbriata*
 1149 with the other two magnoliids, *C. kanehirae* and *L. chinense*, we obtained similar
 1150 results (Fig. 4b, Supplementary Fig. 4.9c-f and Supplementary Table 4.2). For the six-
 1151 species dataset, we obtained 2,673 SSC gene families, and found that both
 1152 concatenation- and coalescent-based phylogenies strongly support the T1 topology as
 1153 well (Supplementary Fig. 4.9g,h). When we selected 16 species, we only found 513
 1154 SSC gene families. The concatenation-based analyses fully supported T2
 1155 (Supplementary Fig. 4.10a), while the coalescent-based phylogeny supported T3 with
 1156 high LPP (Supplementary Fig. 4.10b).

1157 It is generally a tradeoff between the number of species and the number of SSC gene
 1158 families for phylogenomic investigation of the species tree. A large number of the
 1159 SSC gene families might better represent the genome-scale divergence, while limited
 1160 taxon sampling may yield incorrect relationships, due to long-branch attraction and
 1161 other artifacts. In addition, alternative phylogenetic approaches may produce different
 1162 topologies. Thus, our study suggests that the discordant topologies among eudicots,
 1163 monocots, and magnoliids proposed by previous studies result, at least in part, from
 1164 different taxon sampling strategies and phylogenetic approaches (Supplementary
 1165 Table 4.4).



1166
 1167 **Supplementary Fig. 4.9 | Alternative topologies resulting from the four- and six-**
 1168 **species datasets.** The number of corresponding orthogroups (OGs) and the
 1169 phylogenetic approaches are shown under the trees. Concatenation- and coalescent-
 1170 based approaches were employed for phylogenomic analysis. The resulting topologies

1171 all supported the magnoliids as sister to a clade of eudicots and monocots. *A.*
1172 *fimbriata* (Afi), *C. kanehirae* (Cka), *L. chinense* (Lch).



1173
1174 **Supplementary Fig. 4.10 | Discordant topologies from concatenation- and**
1175 **coalescent-based approaches for the selected 16-species dataset.** The number of
1176 corresponding OGs and phylogenetic approaches are shown under the trees.

1177 4.3 Phylogenetic analysis of chloroplast genes

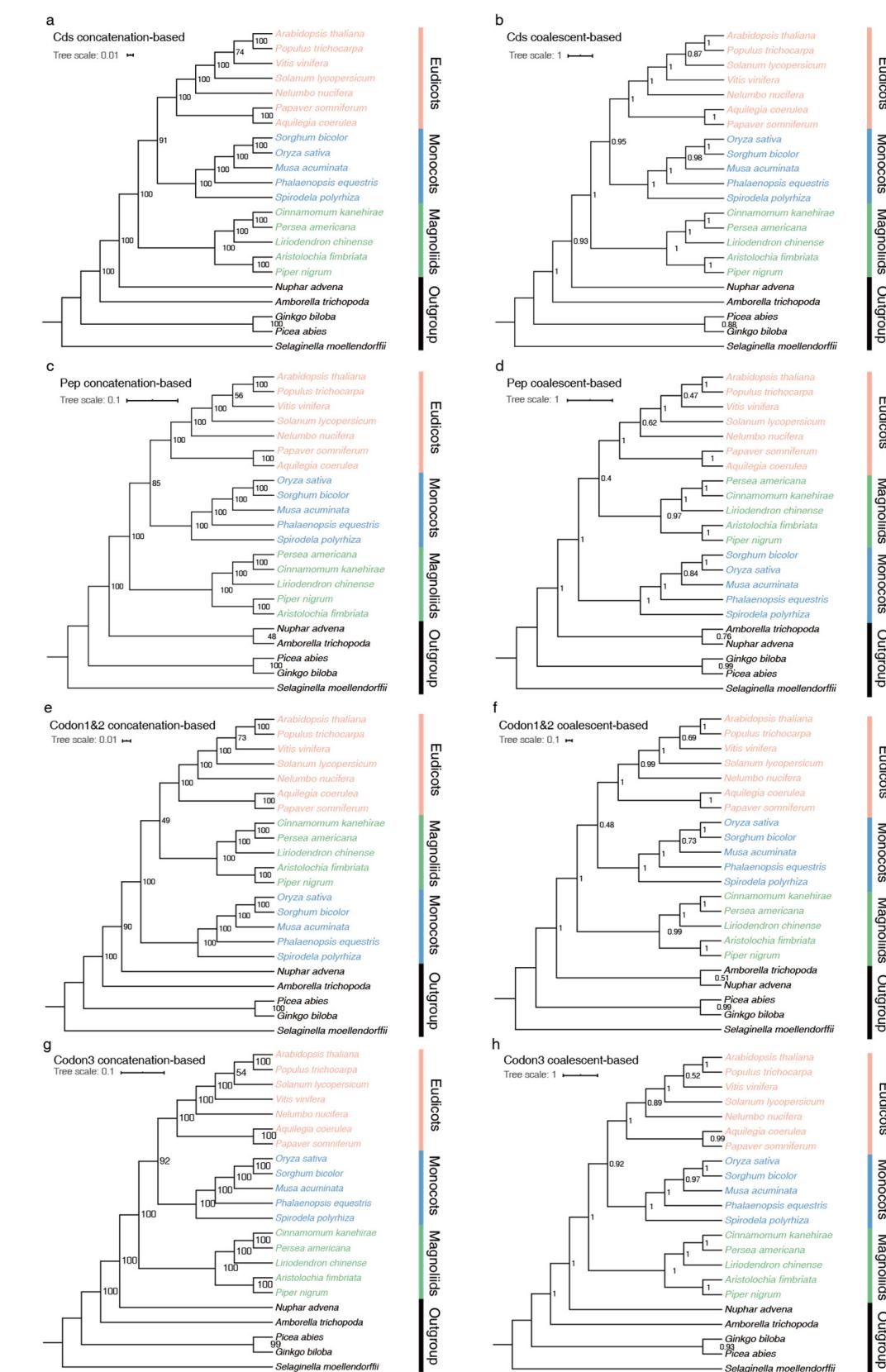
1178 Materials and Methods

1179 We obtained the well-annotated chloroplast genome of *A. fimbriata*, and also
1180 downloaded the chloroplast genomes of the other 21 land plants used for the gene
1181 family analyses in Supplementary Note 2.4 from NCBI. Together, we used the same
1182 set of 22 species in the above nuclear genome phylogenomic analyses as in
1183 Supplementary Note 4.1. Here, we used the chloroplast genome of *Nuphar advena* to
1184 represent Nymphaeales instead of *N. colorata*, because the chloroplast genome of *N.*
1185 *colorata* has not been fully annotated⁵⁹. We manually checked the chloroplast
1186 genomes and extracted 79 protein-coding genes from the selected genomes. After
1187 concatenating the 79 gene alignments into a supermatrix, the ML tree was constructed
1188 using RAxML v8.2.12 under the “PROTGAMMAAUTO” and “GTRGAMMA”
1189 model for amino acid and nucleotide substitution, respectively, with 1000 bootstrap
1190 replicates⁸⁴. Single gene trees were also constructed. In addition, we also used the first
1191 and second codon (codon1&2) and the third codon (codon3) for the above analyses.
1192 Even though coalescent methods are not really appropriate for plastid genomes
1193 because they are uniparentally inherited, as discussed in Stull et al. 2020⁹². We still
1194 performed the coalescent-based phylogeny inferred from the individual ML gene trees
1195 with BS $\geq 50\%$ using ASTRAL-II v5.5.11⁸⁵.

1196 Results and Discussion

1197 Among the 79 chloroplast gene families, 7, 2, 2, and 68 gene trees supported T1, T2,
1198 T3, and T4, respectively (Fig. 4b and Supplementary Tables 4.2 and 4.3), therefore,
1199 the majority of the individual trees also show weak or no resolution about the
1200 phylogenetic relationship of monocots, eudicots, and magnoliids. Concatenation- and
1201 coalescent-based analyses of nucleotide sequences of the 79 protein-coding genes
1202 mostly supported magnoliids as the sister clade to the highly supported clade of
1203 eudicots and monocots (T1; Supplementary Fig. 4.11a-c), which is consistent with
1204 previous reports (Supplementary Table 4.4)^{93,94}. The coalescent-based tree using
1205 amino acid sequences placed magnoliids as sister to eudicots with 40% LPP (T2;
1206 Supplementary Fig. 4.11d). We also performed concatenation- and coalescent-based
1207 analyses on 79 chloroplast genes for codon1&2 and codon3 sequences, respectively.
1208 Surprisingly, we found poor support for alternative relationships among monocots,
1209 eudicots, and magnoliids based on concatenation- and coalescent-based analyses for
1210 codon1&2 (Supplementary Fig. 4.11e,f). However, both codon3-based analyses
1211 highly supported magnoliids as sister to the clade of monocots and eudicots (T1;
1212 Supplementary Fig. 4.11g,h). These results seem to be consistent with the previous
1213 study using nuclear genomes that removed codon3 and decreased the support value of
1214 magnoliids + eudicots⁸³. We speculate that this could be due to the biased codon
1215 usage among the different groups of angiosperms (Discussed more below).

1216



1217 **Supplementary Fig. 4.11 | Phylogenetic relationships inferred from 79**
1218 **chloroplast protein-coding genes using concatenated- and coalescent-based ML**
1219 **approaches.** BS and LPP values for the concatenated- and coalescent-based analysis,
1220 respectively, are shown on the branch to each node.

1221 **4.4 Codon usage bias and the most plausible phylogenetic topology**

1222 **Materials and Methods**

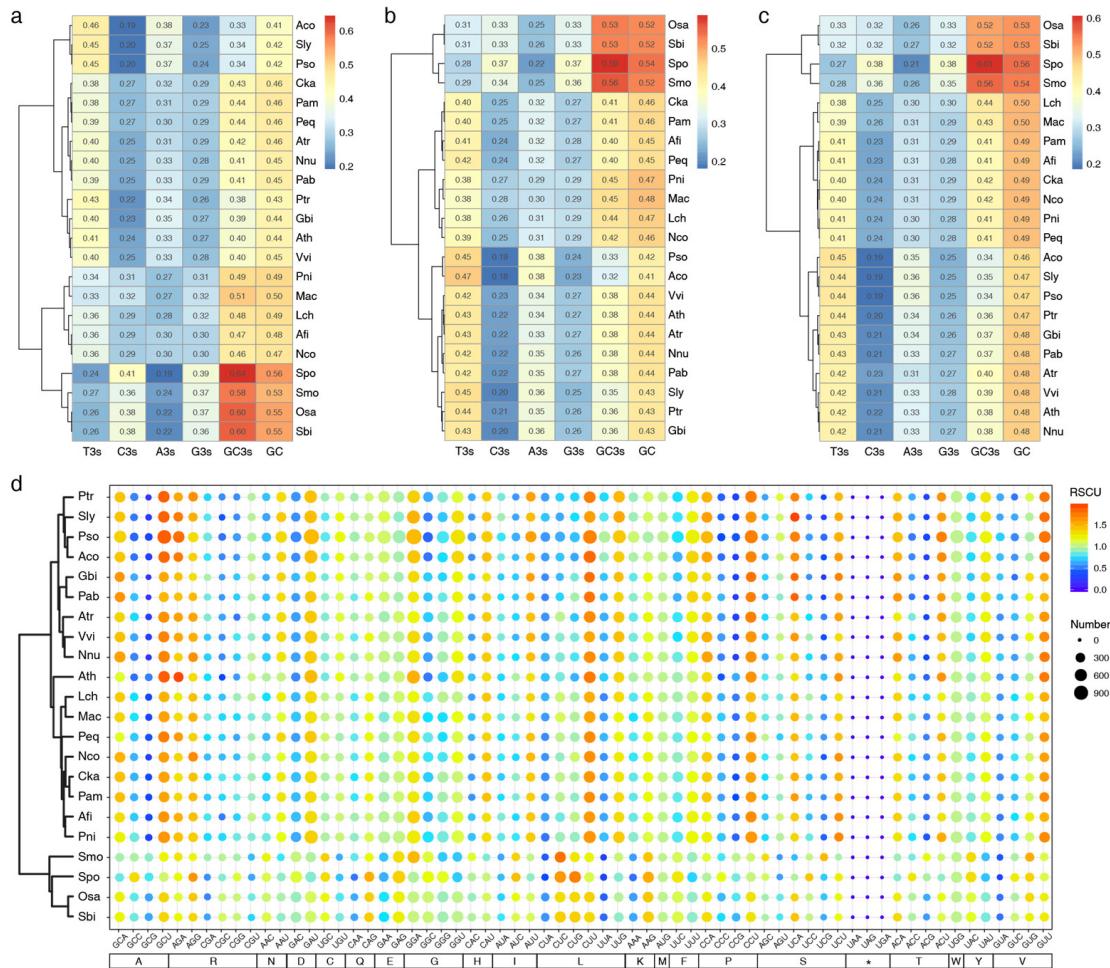
1223 The phenomenon that synonymous codons are not used at equal frequencies in coding
1224 sequences in many organisms is referred as codon usage bias. To investigate the
1225 codon usage patterns in the 22 species included in our phylogenetic analyses, we
1226 performed codon usage analyses for all of nuclear and chloroplast genes in the
1227 genome, and the MSC nuclear genes and chloroplast genes used for phylogenetic
1228 analyses, respectively. First, gene sequences with an internal stop codon in the
1229 reading frame were removed. Then, using CodonW v1.4.2
1230 (<http://codonw.sourceforge.net>) with parameters of “-totals -all_indices”, three indices
1231 of codon usage bias – i.e., relative synonymous codon usage (RSCU), effective
1232 number of codons (ENC), and the frequency of the nucleotides G+C at the third
1233 position (GC3s) – were calculated. Two codon composition indices, GC content of the
1234 entire gene (GC) and the frequency of each individual base at the third position of
1235 codons (A3s, T3s, G3s, C3s), were also tabulated. Differences in codon usage patterns
1236 among these species were shown using pheatmap package in R (www.R-project.org).

1237 **Results and Discussion**

1238 The nuclear genomes of four species (*O. sativa*, *S. bicolor*, *S. polystachya* and *S.*
1239 *moellendorffii*) exhibited different codon usage patterns from the remaining 18
1240 genomes included here. These four species tend to use purine-rich codons more
1241 frequently than pyrimidine-rich codons, and they also have relatively higher GC and
1242 GC3s content compared to other species (Supplementary Fig. 4.12a-c). Consistently,
1243 the RSCU-based cluster analysis also revealed that the preference of synonymous
1244 codon usage in these four species is different from that in the other analyzed species
1245 (Supplementary Fig. 4.12d). This might explain the unexpected placement of *S.*
1246 *moellendorffii* in some of the phylogenetic analyses (Supplementary Fig. 4.6b,d). We
1247 found less variation among species in codon usage in chloroplast genes than in
1248 nuclear genes. Only *S. moellendorffii* shows different codon usage than the other 21
1249 species in chloroplast genes (Supplementary Fig. 4.13), consistent with at least some
1250 previous analyses⁹⁵.

1251 We therefore removed *S. moellendorffii*, *O. sativa*, *S. bicolor*, and *S. polystachya* and
1252 performed the phylogenetic analyses of the 535 MSC nuclear gene families using
1253 concatenation- and coalescent-based approaches. The concatenation-based
1254 phylogenies still supported T2 with magnoliids sister to eudicots (Supplementary Fig.
1255 4.14a,c). However, the coalescent analyses using cds, codon1&2, and codon3
1256 sequences highly supported T3, with magnoliids as sister lineage to monocots
1257 (Supplementary Fig. 4.14b,d,f).

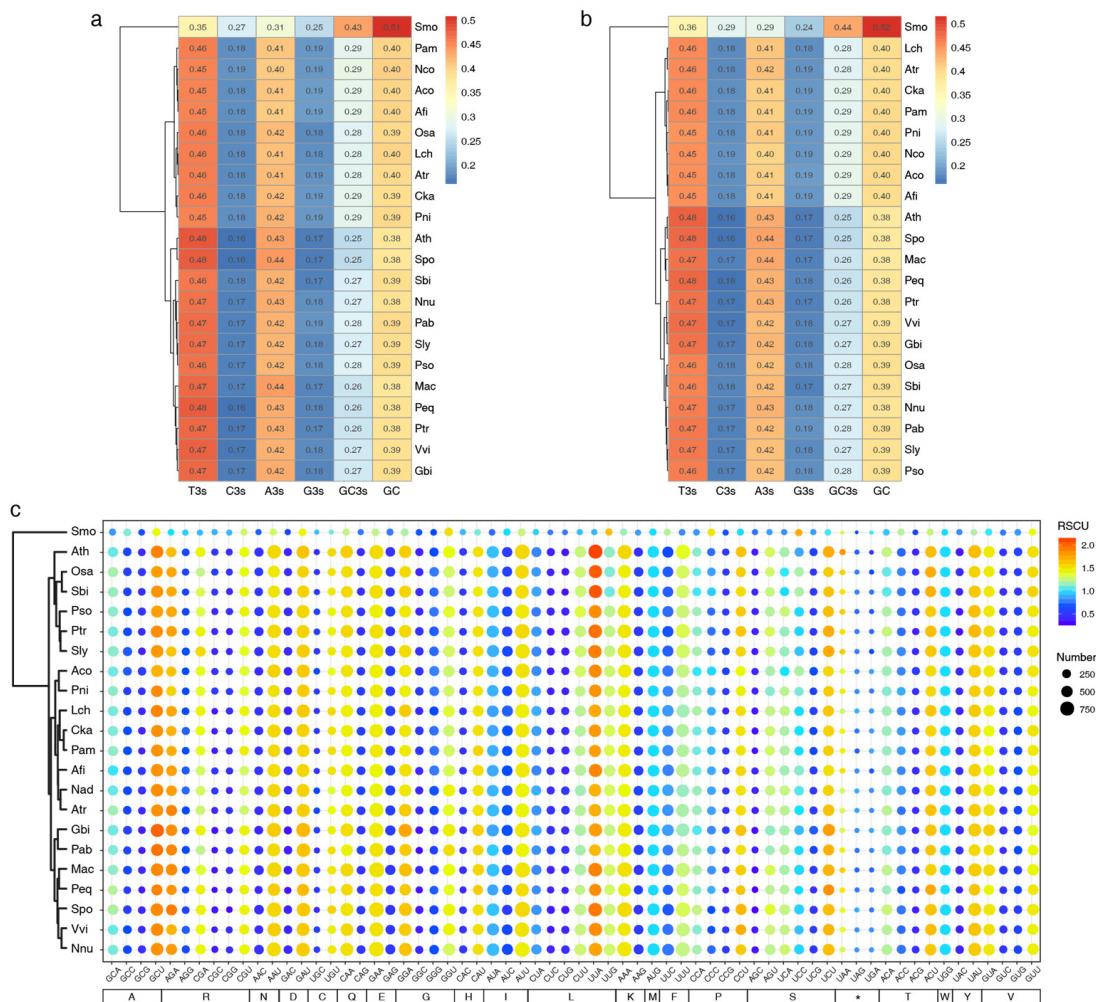
1258 We also excluded the four codon-biased species (*S. moellendorffii*, *O. sativa*, *S.*
 1259 *bicolor* and *S. polystachya*) in the analyses of chloroplast genes. Both the concatenation-
 1260 and coalescent-based analyses using cds or codon3 seem to high/good supported T1
 1261 (Supplementary Fig. 4.15a,b,e,f). However, the concatenation- and coalescent-based
 1262 analyses of the codon1&2 data produced different topologies, but with weak BS/LPP
 1263 support (Supplementary Fig. 4.15c,d).



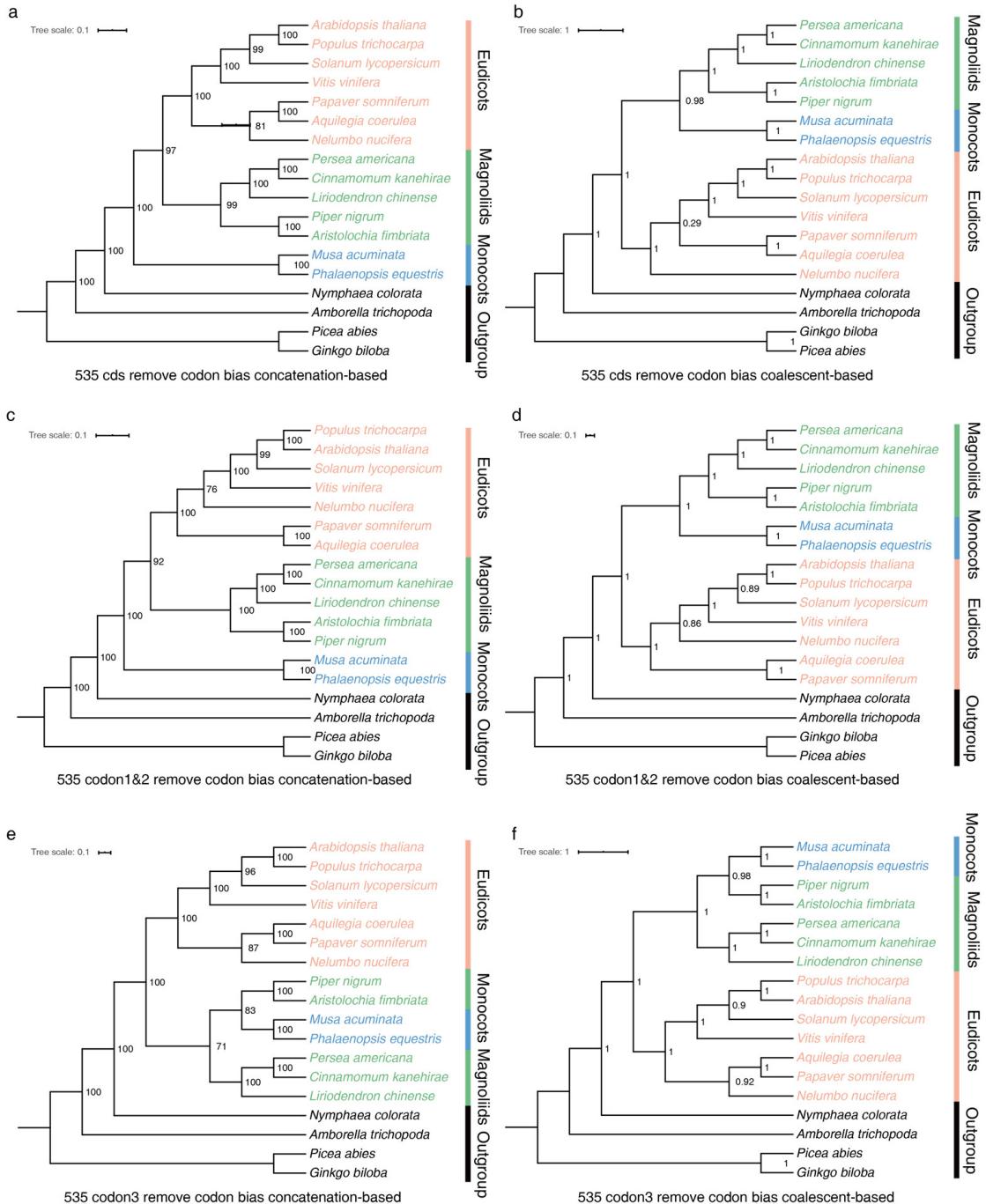
1264

1265 **Supplementary Fig. 4.12 | Codon composition and usage bias of nuclear genes in**
 1266 **the 22 analyzed species.** (a) Comparisons of the third codon composition, GC3s, and
 1267 GC content of genome-level nuclear genes. (b) Comparisons of the third codon
 1268 composition, GC3s, and GC content of the mostly single-copy (MSC) gene families
 1269 used for phylogenetic analyses. (c) Comparisons of the third codon composition,
 1270 GC3s, and GC content of the codons of MSC genes which encode the same amino
 1271 acids in the 22 species. (d) Cluster analysis of the analyzed 22 species based on the
 1272 RSCU for the codons of MSC genes which encode the same amino acids in the 22
 1273 species. Numbers refer to the number of codons used for each species. *A. thaliana*
 1274 (*Ath*), *P. trichocarpa* (*Ptr*), *S. lycopersicum* (*Sly*), *V. vinifera* (*Vvi*), *P. somniferum*
 1275 (*Pso*), *A. coerulescens* (*Aco*), *N. nucifera* (*Nnu*), *O. sativa* (*Osa*), *P. equestris* (*Peq*), *S.*
 1276 *polystachya* (*Spo*), *S. bicolor* (*Sbi*), *M. acuminata* (*Mac*), *A. fimbriata* (*Afi*), *P. nigrum*

1277 (Pni), *P. americana* (Pam), *C. kanehirae* (Cka), *L. chinense* (Lch), *N. colorata* (Nco),
 1278 *A. trichopoda* (Atr), *G. biloba* (Gbi) and *S. moellendorffii* (Smo).

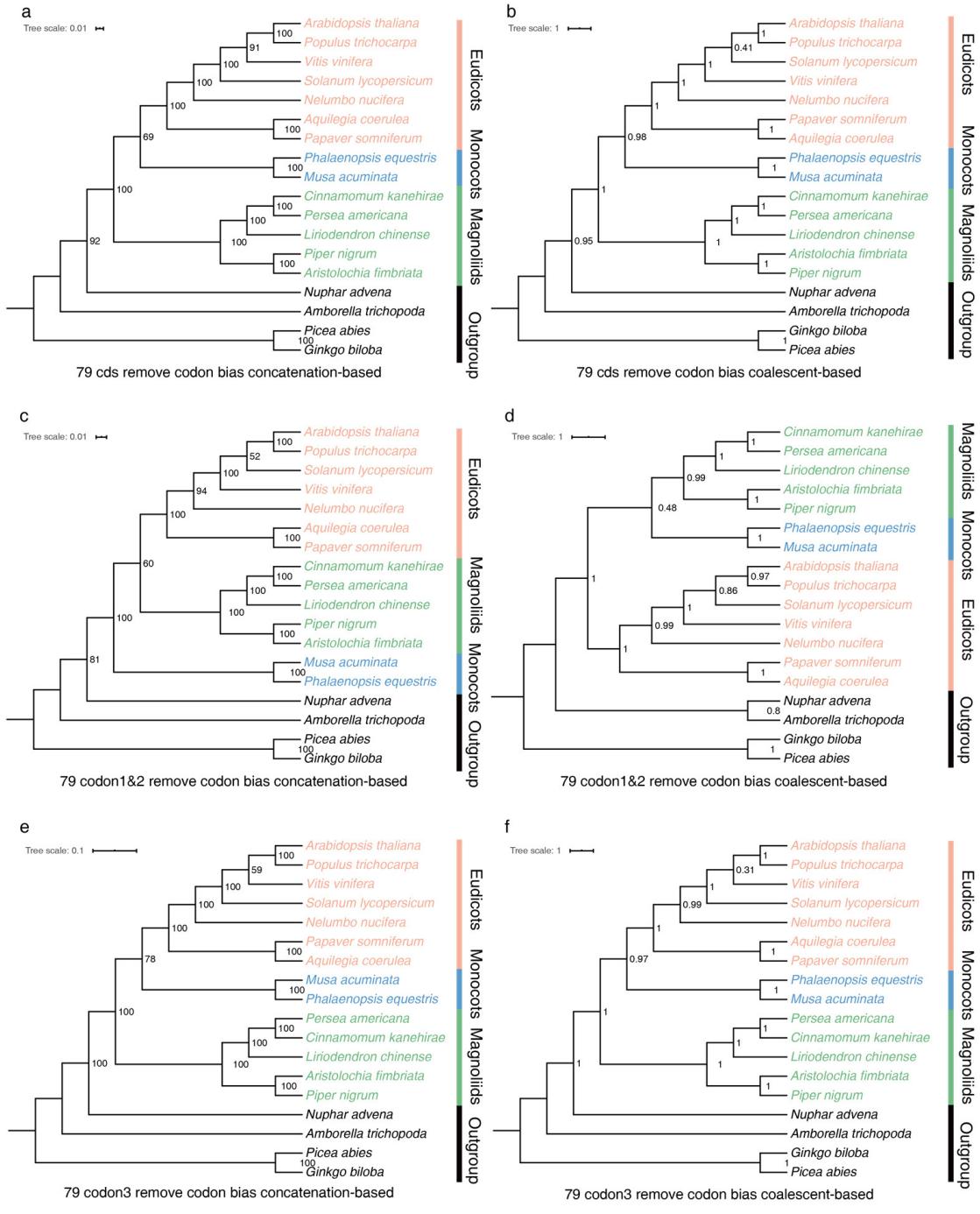


1279
 1280 **Supplementary Fig. 4.13 | Codon composition and usage bias of the chloroplast genes in the 22 analyzed species.** (a) Comparisons of the third codon composition, GC3s, and GC content of all of the chloroplast genes. (b). Comparisons of the third codon composition, GC3s, and GC content of the 79 chloroplast gene families used for phylogenetic analyses. (c). Cluster analysis of the 22 analyzed species based on RSCU for the 79 chloroplast gene families. Numbers refer to the number of used codons in each species. *A. thaliana* (Ath), *P. trichocarpa* (Ptr), *S. lycopersicum* (Sly), *V. vinifera* (Vvi), *P. somniferum* (Pso), *A. coerulea* (Aco), *N. nucifera* (Nnu), *O. sativa* (Osa), *P. equestris* (Peq), *S. polystachya* (Spo), *S. bicolor* (Sbi), *M. acuminata* (Mac), *A. fimbriata* (Afi), *P. nigrum* (Pni), *P. americana* (Pam), *C. kanehirae* (Cka), *L. chinense* (Lch), *N. colorata* (Nco), *A. trichopoda* (Atr), *G. biloba* (Gbi) and *S. moellendorffii* (Smo).



1292

1293 **Supplementary Fig. 4.14 | Phylogenetic relationships inferred from the 535 MSC**
 1294 **gene families after removing the species with codon biases (*S. moellendorffii*, *O.***
 1295 ***sativa*, *S. bicolor* and *S. polystachya*), using concatenated- and coalescent-based**
 1296 **approaches. BS and LPP values for the concatenated- and coalescent-based analyses,**
 1297 **respectively, are shown on branches leading to nodes.**



1298

1299 **Supplementary Fig. 4.15 | Phylogenetic relationships inferred from the 79**
1300 **protein-coding chloroplast genes after removing the species with codon biases (*S.***
1301 ***moellendorffii*, *O. sativa*, *S. bicolor* and *S. polystachya*), using concatenated- and**
1302 **coalescent-based approaches.** BS and LPP values for the concatenated- and
1303 **coalescent-based analyses, respectively, are shown on branches leading to nodes.**

1304

1305 **4.5 Molecular dating and gene family evolution**

1306 **Materials and Methods**

1307 The Bayesian method MCMCTree in the PAML v4.9e package⁹⁶ was employed to
1308 estimate the divergence times using the T3 topology that was constructed with the 98
1309 SSC gene families from 22 species and rooted with *S. moellendorffii* as the input tree.
1310 Following fossil dates were used for the calibration procedure: maximum age of 400
1311 Ma for the divergence of *S. moellendorffii*⁹⁷, a minimum age of 309 Ma for the
1312 crown-group seed plants⁹⁸, a minimum age of 125 Ma for the eudicots⁹⁹, a maximum
1313 age of 113 Ma for the monocots¹⁰⁰⁻¹⁰², and a maximum age of 113 Ma for the
1314 magnoliids¹⁰³. Branch lengths were estimated using BASEML from the PAML
1315 package under the GTR + G model (model = 7)⁹⁶. The overall substitution rate (rgene
1316 gamma) and rate-drift parameter (sigma2 gamma) were set as G (1, 5.6) and G (1,
1317 4.0) respectively. We ran all analyses twice to check for consistency and to ensure the
1318 effective sample size was above 200 in Tracer v1.7 ([http://tree.](http://tree.bio.ed.ac.uk/software/tracer/)
1319 [bio.ed.ac.uk/software/tracer/](http://tree.bio.ed.ac.uk/software/tracer/)).

1320 Orthogroups constructed above (Supplementary Note 2.4) were used to investigate the
1321 expansion and contraction of gene families during evolutionary history using CAFÉ
1322 v4.1¹⁰⁴. As instructed in the CAFÉ manual, we removed the 375 gene families that
1323 have more than 100 genes in each orthogroup, and the other 51,388 gene families
1324 were used for the gene family evolution analysis (automatic λ and μ estimation,
1325 significance level for fast-evolving families 5%).

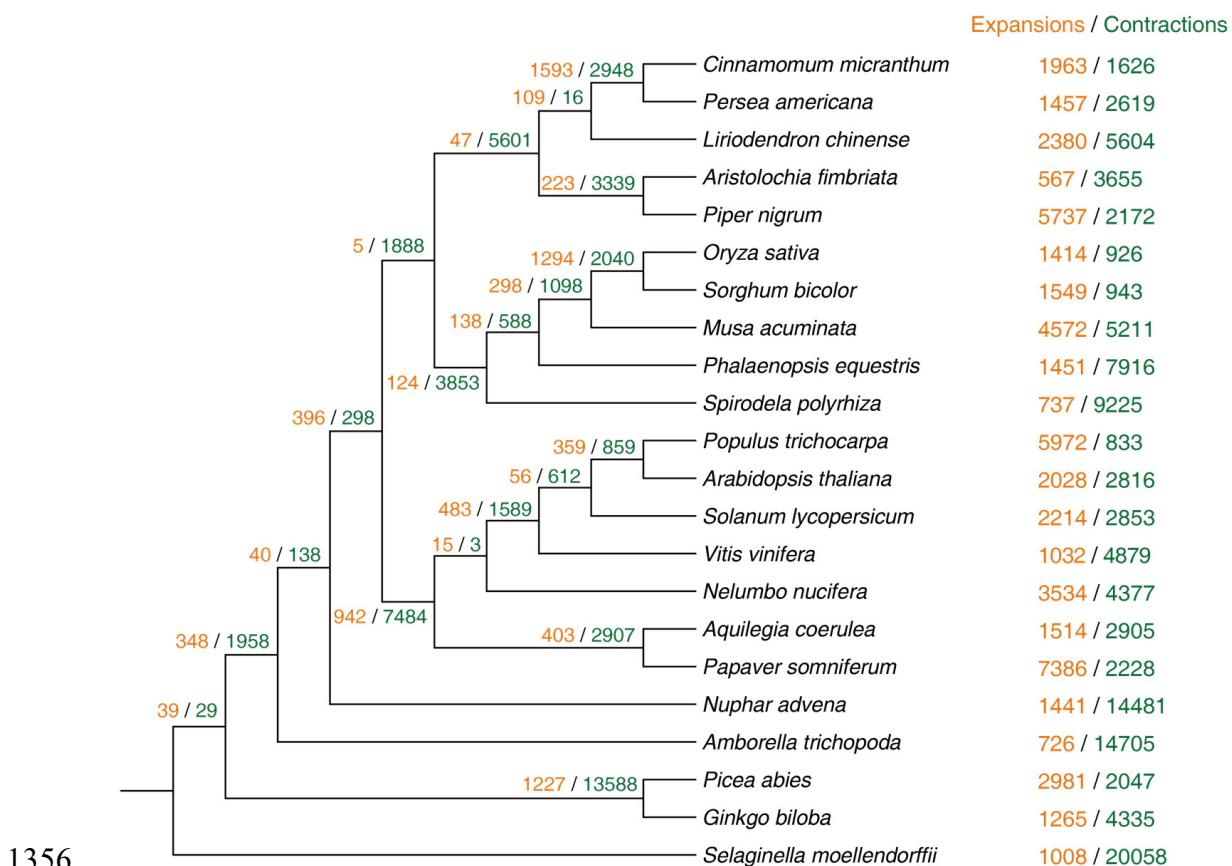
1326 Due to the crucial roles of transcription factor (TF) for biological processes, we also
1327 specifically identified their family members in *A. fimbriata* and compared with other
1328 representative angiosperms. For each TF family, the *A. thaliana* genes were obtained
1329 from TAIR (<https://www.arabidopsis.org/>), and the corresponding OG containing
1330 these genes were retrieved. If multiple OGs were identified for the same TF family,
1331 we further combined them together as one family.

1332 **Results and Discussion**

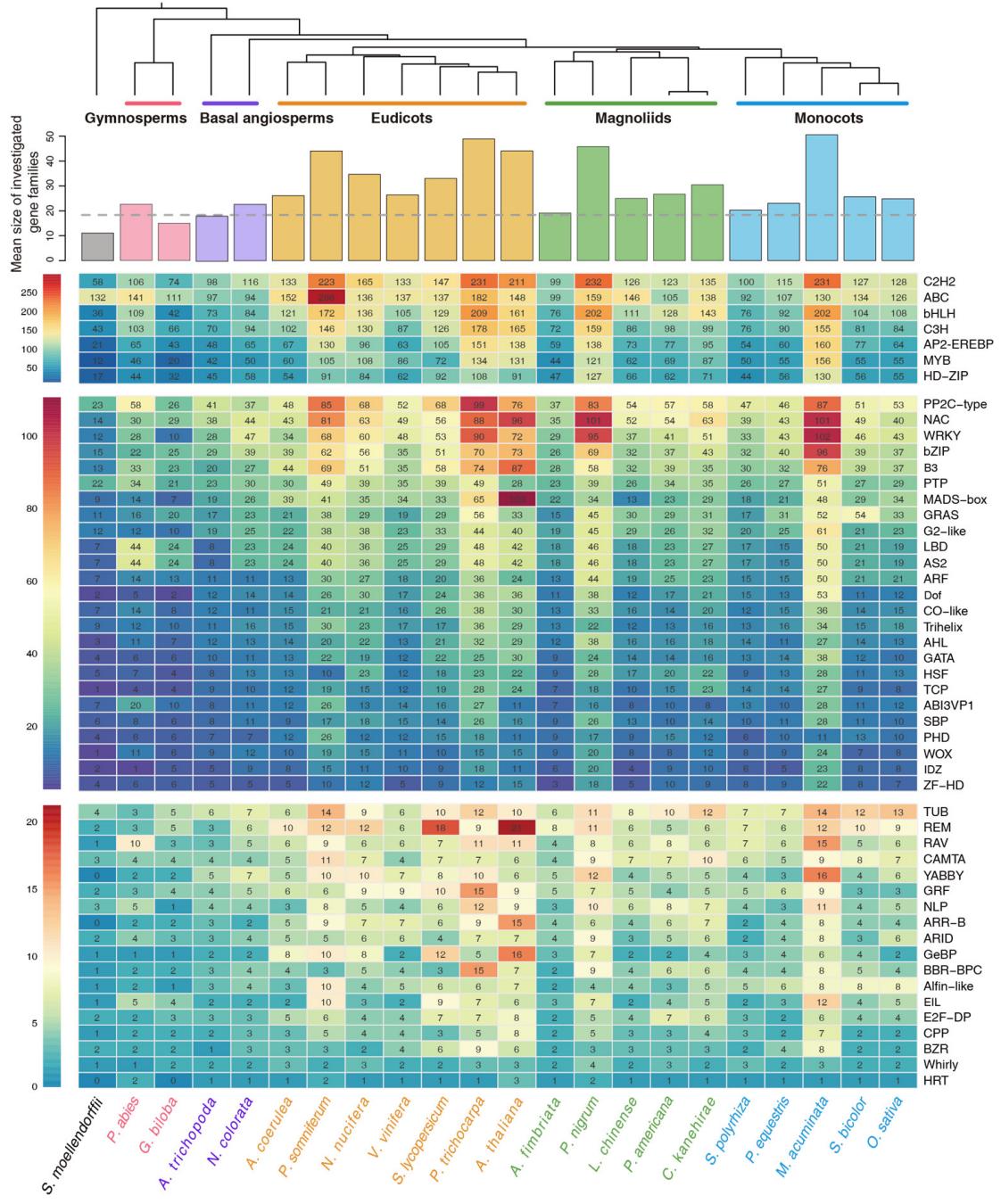
1333 Molecular dating of angiosperm lineages, using the stringent set of 98 SSC genes and
1334 with age calibrations based on 5 fossil dates, inferred the crown age of angiosperms at
1335 ~250 Ma (Fig. 4d). The split between monocots and magnoliids was estimated at
1336 ~186 Ma, and the divergence time between magnoliid + monocot clade and eudicots
1337 was at ~193 Ma. The rapid divergence among the magnoliids, monocots, and eudicots
1338 is likely responsible for the great difficulty in reconstructing relationships among
1339 these clades⁵⁶.

1340 In addition, we found that 348 gene families expanded in the common ancestor of
 1341 angiosperms, whereas 1,958 gene families contracted (Supplementary Fig. 4.16). In
 1342 the common ancestor of eudicots, 942 gene families expanded, and 7,484 gene
 1343 families contracted (Supplementary Fig. 4.16). In magnoliids, the black pepper
 1344 genome has the largest number of expanded orthogroups (5,737), while only 567
 1345 orthogroups expanded in the *A. fimbriata* lineage. We also found that 2,380, 1,457,
 1346 and 1,963 orthogroups expanded in the *L. chinense*, *P. americana*, and *C. kanehirae*
 1347 lineages, respectively (Supplementary Fig. 4.16). *A. fimbriata* shows the least gene
 1348 family expansion among magnoliids, which may be due to its lack of lineage-specific
 1349 WGDs.

1350 We found increasing numbers of TFs during land plant evolution, and the *P.*
 1351 *somniferum*, *P. trichocarpa*, *A. thaliana*, *P. nigrum*, and *M. acuminata* genomes have
 1352 larger mean family size of TF than the other species. Notably, the *A. fimbriata*
 1353 genome, as well as *Amborella*, exhibit the lowest numbers of TFs among the
 1354 angiosperms (Supplementary Fig. 4.17). The family sizes of TFs are also well
 1355 correlated with the number and timing of WGDs.



1356
 1357 **Supplementary Fig. 4.16 | Estimation of gene family expansions and**
 1358 **contractions.** Orange/green numbers on each branch indicate the numbers of gene
 1359 family expansions and contractions, respectively.



1360

1361

Supplementary Fig. 4.17 | Variation in gene copy numbers of transcription

factors during land plant evolution. The *A. fimbriata* genome exhibits low mean size of gene families, similar to that of *A. trichopoda*. Transcription factors were classified into three main groups based on their total number of homologs in the 22 selected species.

1366 5 High-specialized flower morphology and floral development in *A. fimbriata*

As described above, the *Aristolochia* flowers exhibits unique morphology and highly specialized modifications (Extended Data Fig. 1). In the absence of lineage-specific

1369 WGDs, the newly sequenced genome of *A. fimbriata* offers an opportunity to study
1370 the genetic basis underlying the flower development and modifications in contrast to
1371 other angiosperms. It might shed new insights on the evolution of floral genes and
1372 regulatory networks that ultimately help us to understand the developmental
1373 mechanisms that contribute to the highly modified floral features.

1374 **5.1 Phylogeny of the MADS-box gene family**

1375 MADS-box genes encode a large family of transcription factors that are involved in
1376 various developmental processes in plants, animals, and fungi¹⁰⁵⁻¹⁰⁷. In green plants,
1377 the MADS-box gene family has been classified into type I and type II. Type I contains
1378 Ma, M β , and My subfamilies, while type II has been further divided into MIKC^C and
1379 MIKC* groups^{105,108,109}. The MIKC^C genes play essential roles in the flower
1380 development and floral regulatory networks, thus they are the most widely studied
1381 MADS-box genes¹⁰⁸⁻¹¹⁰.

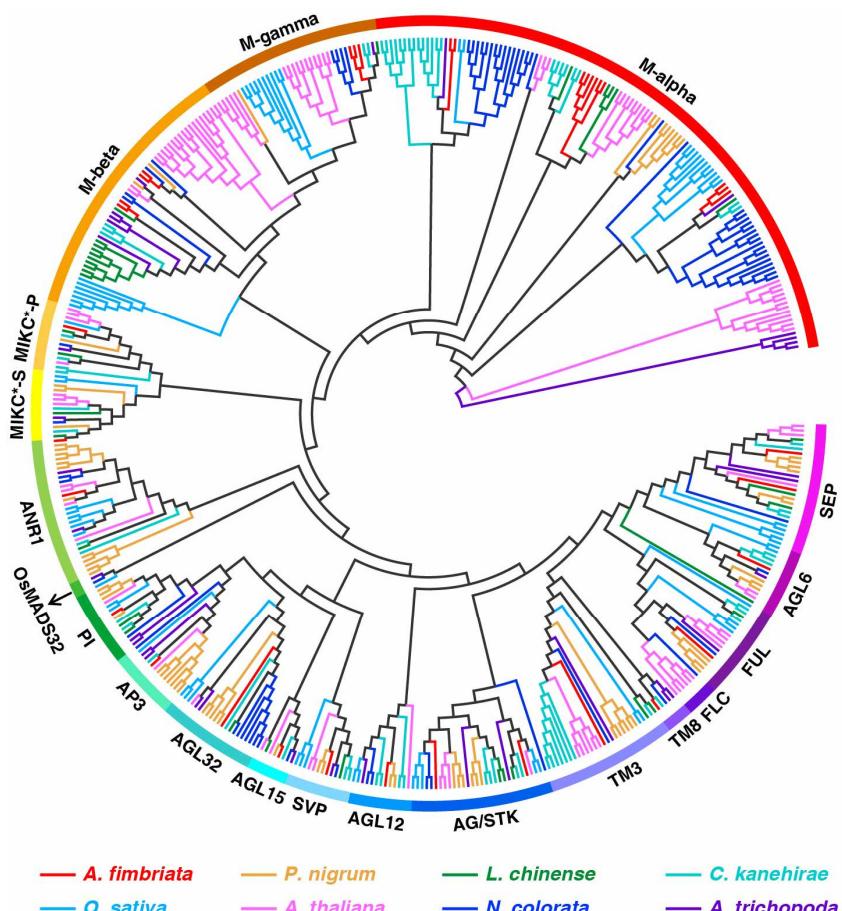
1382 **Materials and Methods**

1383 To identify MADS-box genes in *A. fimbriata*, we first retrieved the well-studied and
1384 annotated subtypes of MADS-box genes in *A. thaliana*¹¹¹, *O. sativa*¹¹², *Amborella*⁵⁵
1385 and *N. colorata*⁵⁹. The above constructed orthogroups (Supplementary Note 2.4) that
1386 contain these well-annotated MADS-box genes were obtained. To avoid any missing
1387 putative homologs, we further performed the BLASTP analysis using these well-
1388 annotated MADS-box genes as queries, as well as the profile hidden Markov model
1389 (HMM) searches using SPF-TF (PF00319) from the Pfam database¹¹³ as seed, against
1390 the genome-wide amino acid sequences in *A. fimbriata*, employing the BLASTP⁵² (*E*-
1391 value < 10⁻³) and the hmmsearch in HMMER v3.3¹¹⁴ (-domE 0.001) respectively.
1392 The identified putative MADS-box genes were further manually inspected and
1393 screened using InterProScan¹¹⁵. We also identified the MADS-box genes in three
1394 other magnoliids (*C. kanehirae*, *P. nigrum* and *L. chinense*) using the same process.
1395 Due to the large number of MADS-box gene family members, we mainly focused on
1396 these eight species representing major clades of angiosperms for phylogenetic
1397 analysis.

1398 To construct the phylogeny of these MADS-box genes, amino acid sequences were
1399 aligned first using MAFFT v7.312¹¹⁶, and the nucleotide sequences were forced to fit
1400 the amino acid alignment using PAL2NAL v14⁸². The poorly aligned regions were
1401 removed using trimAL v3¹¹⁷ with parameter of “-gt 0.3”. Phylogenetic tree was
1402 constructed using maximum likelihood in RAxML v8.2.12⁸⁴ under the
1403 “GTRGAMMA” model with 100 bootstrap replicates.

1404 **Results and Discussion**

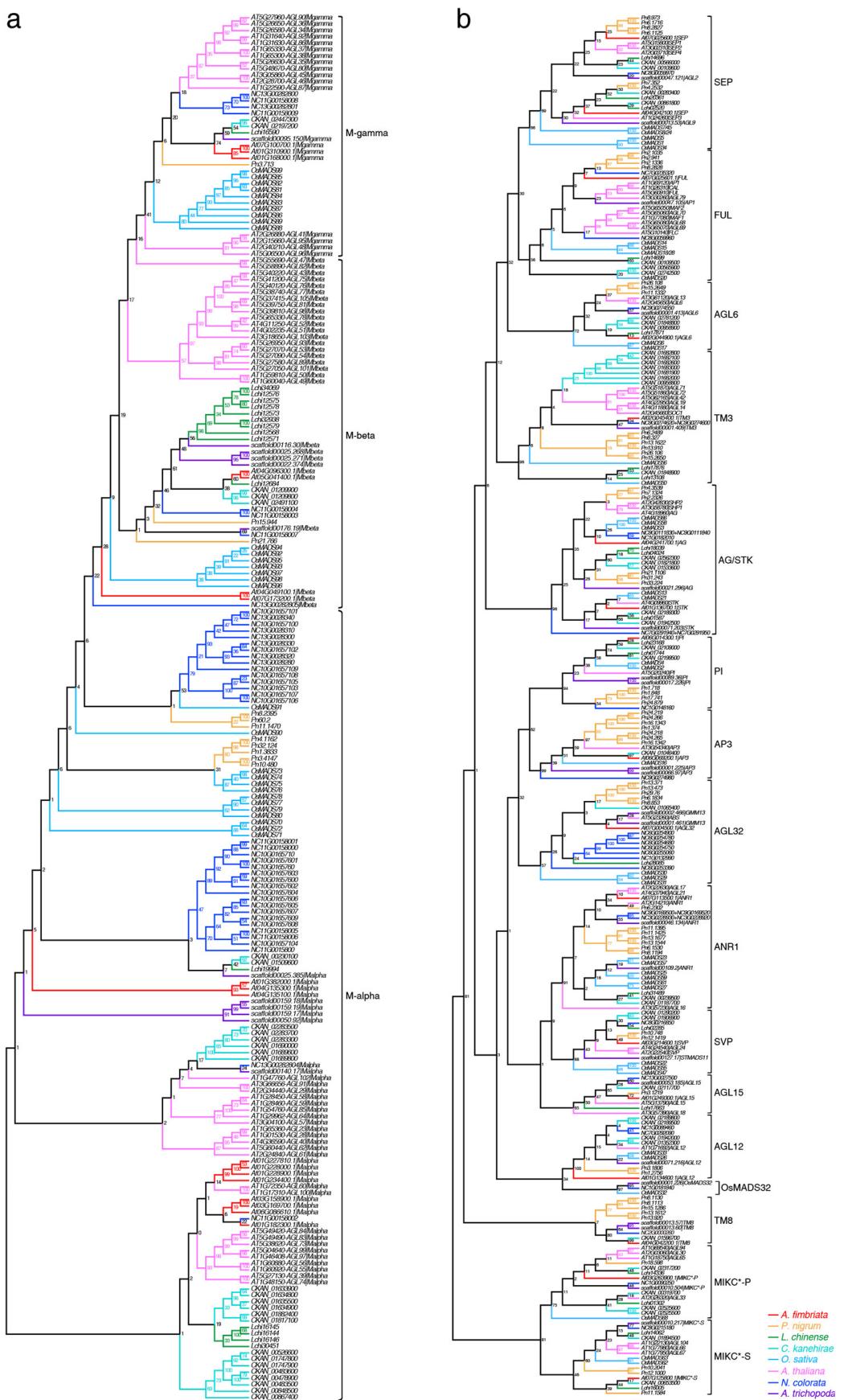
1405 Based on the phylogeny and the well-annotated genes in the above mentioned four
 1406 species, 18 type I and 17 type II MADS-box genes were identified in *A. fimbriata*
 1407 (Supplementary Tables 5.1 and 5.2 and Supplementary Figs. 5.1 and 5.2). We further
 1408 specifically examined the floral organ identity determination program. Surprisingly,
 1409 we found that the *A. fimbriata* genome contains only one homologue for each of the
 1410 eight classes of floral organ identity genes including: *APETALA1/FRUITFULL* and
 1411 *AGAMOUS-LIKE6* (*API/FUL* and *AGL6*, A function for sepals and petals),
 1412 *APETALA3* and *PISTILLATA* (*AP3* and *PI*, B function for petals and stamens),
 1413 *AGAMOUS* (*AG*, C function for stamens and carpels), *SEEDSTICK* (*STK*, D function
 1414 for ovule), and *SEPALLATA* (*SEP1* and *SEP2*, E function for interacting with ABCD-
 1415 function proteins) (Supplementary Figs. 5.1 and 5.2 and Fig. 5b). This is likely due to
 1416 the lack of any additional WGD in *A. fimbriata* since the origin of extant
 1417 angiosperms. In addition, we found high conservation of gene structure and exon
 1418 sequences of floral organ identity genes between *A. fimbriata* and *Amborella*, while
 1419 the intron length of *Amborella* genes are generally longer than that of *A. fimbriata*
 1420 orthologues (Extended Data Fig. 8).



1421
 1422 **Supplementary Fig. 5.1 | Phylogeny of MADS-box genes from the selected eight**
 1423 **species.** The phylogenetic tree was constructed using maximum likelihood in
 1424 RAxML, and the MADS-box genes are classified into 20 clades. Branches are colored
 1425 according to the species color scheme under the tree.

1426

1427 **Supplementary Fig. 5.2 | Phylogenetic trees of type I (a) and type II (b) MADS-**
 1428 **box genes from eight species with their gene names.**



1429 **5.2 Alternative splicing of MADS-box genes in *A. fimbriata***

1430 Alternative splicing (AS) is a key post-transcriptional processing step of pre-mRNA,
1431 which is prevalent in eukaryotic organisms and sometimes can widely lead to great
1432 complexity of the functional transcripts from a single gene locus^{118,119}. Previous
1433 studies have illustrated that AS plays crucial roles in the processes of floral
1434 development¹²⁰, vernalization, and biotic and abiotic stress responses¹²¹. Given the
1435 relatively small number of MADS-box genes and other floral regulatory TFs in *A.*
1436 *fimbriata* (Fig. 5a), we particularly investigated AS events of MADS-box genes in *A.*
1437 *fimbriata*, to explore whether AS contributes to the development of the highly
1438 modified flower structure of *A. fimbriata*.

1439 **Materials and Methods**

1440 In addition to the full-length transcript sequencing in the annotation section
1441 (Supplementary Note 2), we further collected and pooled the flower buds at different
1442 developmental stages (from stage 5 to anthesis)¹²² together in relatively equal amount
1443 to perform much deeper transcriptome sequencing to get the potential AS for floral
1444 genes. The extracted RNA from the mixed sample was used for isoform sequencing
1445 (Iso-Seq) on the PacBio Sequel II platform. The raw sequence data were processed by
1446 SMRT Link v8.0 software (<https://www.pacb.com/support/software-downloads/>).
1447 Circular consensus sequences (CCSs) were generated from the raw subreads BAM
1448 file to identify full-length (FL) reads using ccs with parameters of “--min-passes 1 --
1449 min-length 100”. Then, full-length non-chimeric (FLNC) reads were identified if they
1450 have the 5'-primer, 3'-primer, and poly(A) tail. Lastly, FLNC reads from the same
1451 isoform were clustered and further polished using subreads.

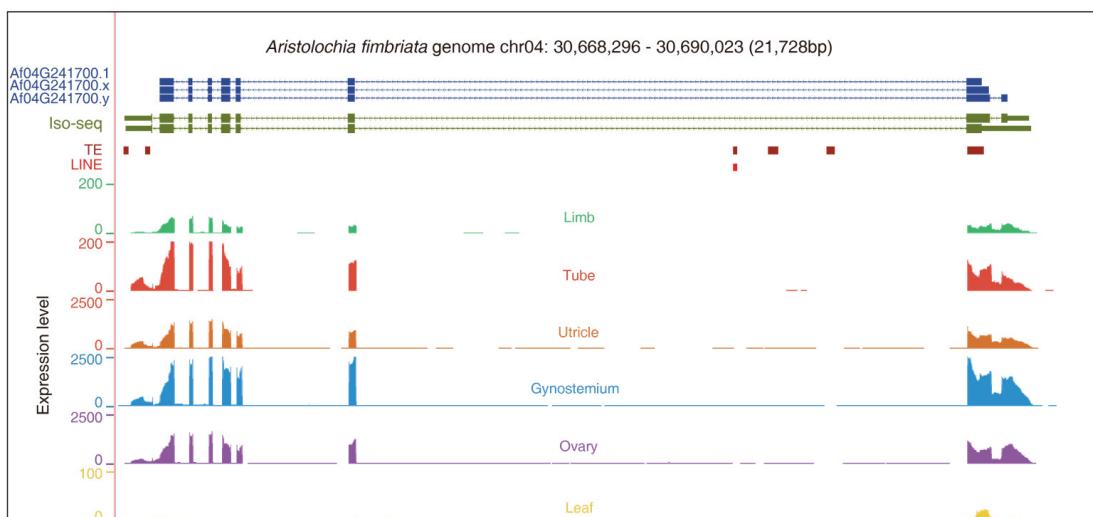
1452 The high-quality and full-length consensus sequences were aligned to the *A. fimbriata*
1453 genome using the Genome Mapping and Alignment Program (GMAP) software⁴⁸
1454 with parameters of “-S -A -f sampe --max-intronlength-middle 5000 -t 15”.
1455 SpliceGrapher v0.2.7 software¹²³ was used to identify AS events in *A. fimbriata*, in
1456 which the models of splice sites were designed as donors of GT or GC and acceptor of
1457 AG, respectively.

1458 **Results and Discussion**

1459 The quality control assessments indicated the extracted RNA was of high quality,
1460 with the RNA integrity number (RIN) = 9.7 and the 28S/18S ribosomal ratio = 1.6.
1461 The generated raw subreads data were approximately 133.09 Gb in size, and the
1462 average length of subreads was approximately 2,126 bp. After processing with SMRT
1463 Link, we obtained 132,090 high-quality and polished full-length consensus sequences,
1464 which were further mapped to the *A. fimbriata* reference genome. We found 97.8% of

1465 all annotated genes (21,282/21,751) with at least one transcript, and all of the MADS-
1466 box genes were detected. These results indicate that the Iso-Seq data set has very high
1467 sequencing depth and a wide range of coverage.

1468 In total, we identified 3,539 AS events from 2,678 genes in *A. fimbriata* genome,
1469 including intron retention (IR; 2,940, 83.1%), exon skipping (ES; 126, 3.6%),
1470 alternative 3' splice sites (A3SS; 280, 7.9%), and alternative 5' splice sites (A5SS;
1471 193, 5.5%). IR was the most abundant type of AS event. Notably, we found no AS
1472 event for 33 of the 35 annotated MADS-box genes, and only two (*AG* and *AGL15*)
1473 were detected to have potential AS events (both with IR type) (Supplementary Fig.
1474 5.3). *A. fimbriata*, with a relatively small number of MADS-box genes, might employ
1475 mechanisms other than multiple alternatively spliced isoforms to generate its complex
1476 floral morphology.



1477
1478 **Supplementary Fig. 5.3 | Genomic locations, overlapped TE/LINE elements, and**
1479 **expression levels of the *AG* gene in *A. fimbriata*.**

1480 **5.3 Expression patterns of floral organ identity genes**

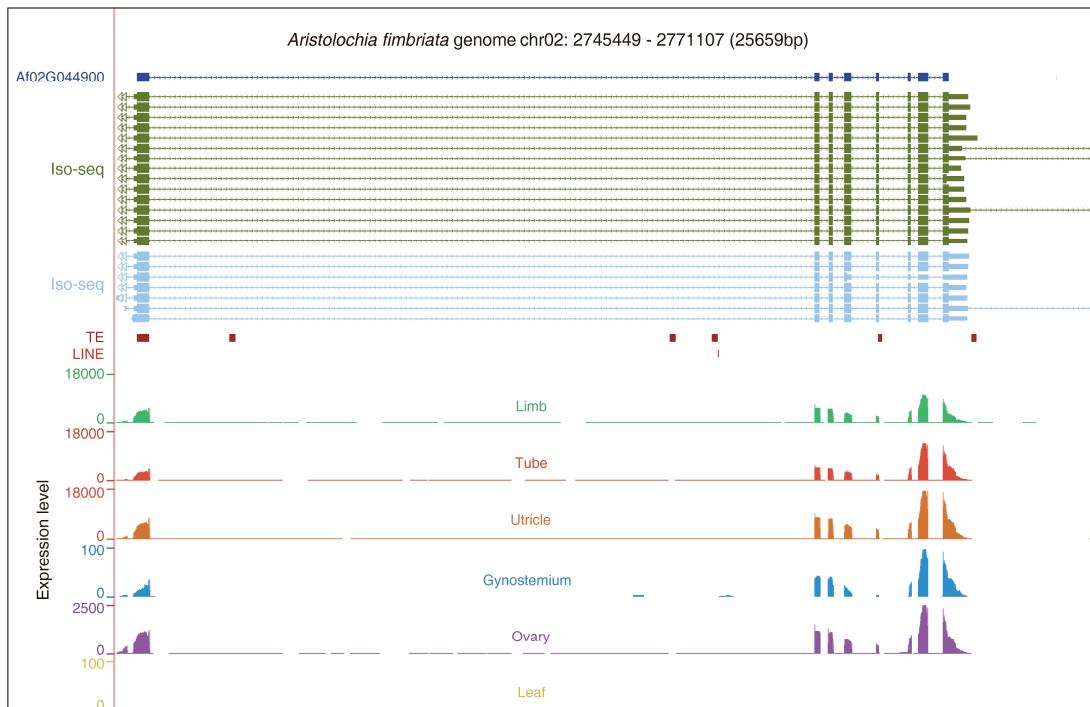
1481 **Materials and Methods**

1482 In this study, we wanted to obtain a general idea about the expression patterns of these
1483 floral organ identity genes, with the understanding that the early stage flower buds are
1484 usually used for studying the role of these genes in specifying floral organs. Here, the
1485 flowers at anthesis and late pre-anthesis (developmental stages 7 and 8) from living
1486 plants were collected and further dissected into 5 different tissue types (limb, tube,
1487 utricle, gynostemium, and ovary). Young leaves were also collected. Each sample had
1488 three biological replicates. Total RNA was extracted and sequenced on an Illumina

1489 HiSeq4000 instrument. RNA-Seq raw reads from three replicates of the 11 samples
1490 were preprocessed using Trimmomatic²⁶ to remove adaptor sequences and low-
1491 quality reads with the parameters of “PE -phred33 ILLUMINACLIP: Trimmomatic-
1492 0.36/adapters/TruSeq3-PE.fa: 2:30:10 LEADING: 20 TRAILING: 20
1493 SLIDINGWINDOW: 4:20 MINLEN: 50”. The clean reads were then mapped to the
1494 reference genome using HISAT2 with default parameters. The expression abundance
1495 values were calculated using Stringtie¹²⁴ with options “-e -A”, and we averaged the
1496 abundance values from the three biological replicates to obtain levels of gene
1497 expression.

1498 **Results and Discussion**

1499 We found that all of the floral organ identity genes exhibited relatively stable
1500 expression patterns during the pre-anthesis and anthesis stages (Fig. 5c), which were
1501 largely consistent with those of other well-studied species, such as *A. thaliana*¹²⁵. For
1502 the A-class genes, we found that *AfFUL* has relatively low expression levels, while
1503 *AfAGL6* is highly expressed in the perianth (limb, tube, and utricle) as well as in the
1504 ovary (Fig. 5c and Supplementary Fig. 5.4). For B-class genes, *AfAP3* and *AfPI* are
1505 widely expressed in all studied floral organs, and their overlapping expression pattern
1506 in the sepal-derived perianth might be associated with the petaloidy of the perianth
1507 (Fig. 5c). The *AfAG* gene is mainly expressed in the gynostemium, ovary, and utricle
1508 (Fig. 5c and Supplementary Fig. 5.3), and *AfSTK* is highly expressed in the ovary
1509 (Fig. 5c). Despite their wide ranges of expression in the pre-anthetic and anthetic
1510 flowers compared to the more restricted patterns in *A. thaliana*^{126,127}, we found that
1511 the relatively high expression in specified floral organs largely agrees with their
1512 putative ascribed roles in floral organ patterning (Fig. 5c). Further expression
1513 profiling and functional studies using earlier developmental stage flowers are needed
1514 to confirm the function of these floral organ identity genes.



1515

1516 **Supplementary Fig. 5.4 | Genomic locations, overlapped TE/LINE elements, and**
1517 **expression levels of the *AGL6* gene in *A. fimbriata*.**

1518 **5.4 Genetic basis of the floral organ fusion and the bilaterally symmetry of**
1519 **flowers**

1520 Floral organ fusion is a peculiar feature of the *A. fimbriata* flower. The stamens and
1521 stigmatic lobes are fused to form the gynostemium, and the sepals are fused to form a
1522 calyx consisting of an inflated utricle, a narrow tube, and a limb^{2,5}. It has been
1523 suggested that the boundary-specification genes, *CUP-SHAPED COTYLEDON*
1524 (*CUC*) genes belonging to NAC transcription factors, are the key determination factor
1525 for the fusion of floral organs¹²⁸⁻¹³⁰.

1526 In addition, *Aristolochia* flowers exhibit monosymmetric perianth, which is rare in
1527 magnoliids and also rare in angiosperms in general^{131,132}. It has been suggested that
1528 floral symmetry may impact specialized interactions between plants and their insect
1529 pollinators¹³³. Previous studies have proposed the *CYCLOIDEA* (*CYC*)/*TEOSINTE*
1530 *BRANCHED1* (*TB1*) class and *CINCINNATA* (*CIN*) class of Teosinte
1531 branched1/Cycloidea/Proliferating cell factor (TCP) transcription factors as key
1532 regulators of flower symmetry establishment and morphogenesis of leaf-like organs in
1533 model species *Antirrhinum majus* and some other species¹³⁴⁻¹³⁹.

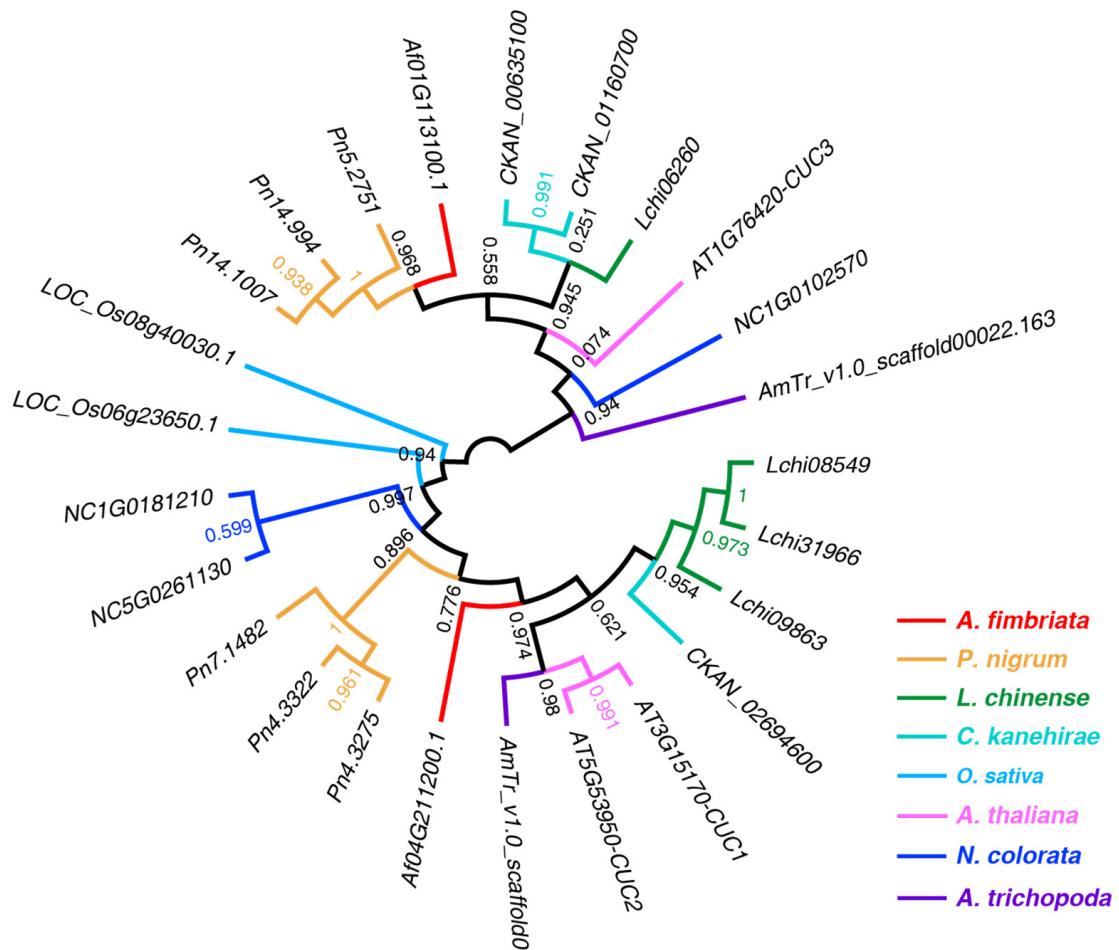
1534 **Materials and Methods**

1535 The associated homologous genes for specific flower features were examined in the
1536 *A. fimbriata* genome. To identify the *CUC* genes in *A. fimbriata*, the well-annotated
1537 *CUC* genes in *A. thaliana* were firstly retrieved from previous studies^{129,130} and TAIR
1538 (<https://www.arabidopsis.org/>). Then, similar to the identification of MADS-box
1539 genes in the above section (Supplementary Note 5.1), we integrated and screened the
1540 results from the associated OG(s), BLASTP and HMM searches to identify the *CUC*
1541 homologs in *A. fimbriata*, *A. trichopoda*, *N. colorata*, *O. sativa*, *C. kanehirae*, *P.*
1542 *nigrum* and *L. chinense*. Similarly, using the well-studied *CYC* and *CIN* genes in *A.*
1543 *majus* and *A. thaliana* as references¹³⁶⁻¹³⁸, the *CYC* and *CIN* homologous genes were
1544 also comprehensively identified. Phylogenetic trees were then constructed using the
1545 methods described above (Supplementary Note 5.1). The expression patterns of these
1546 identified *A. fimbriata* homologous genes were also examined from the
1547 aforementioned transcriptomes in Supplementary Note 5.3.

1548 Results and Discussion

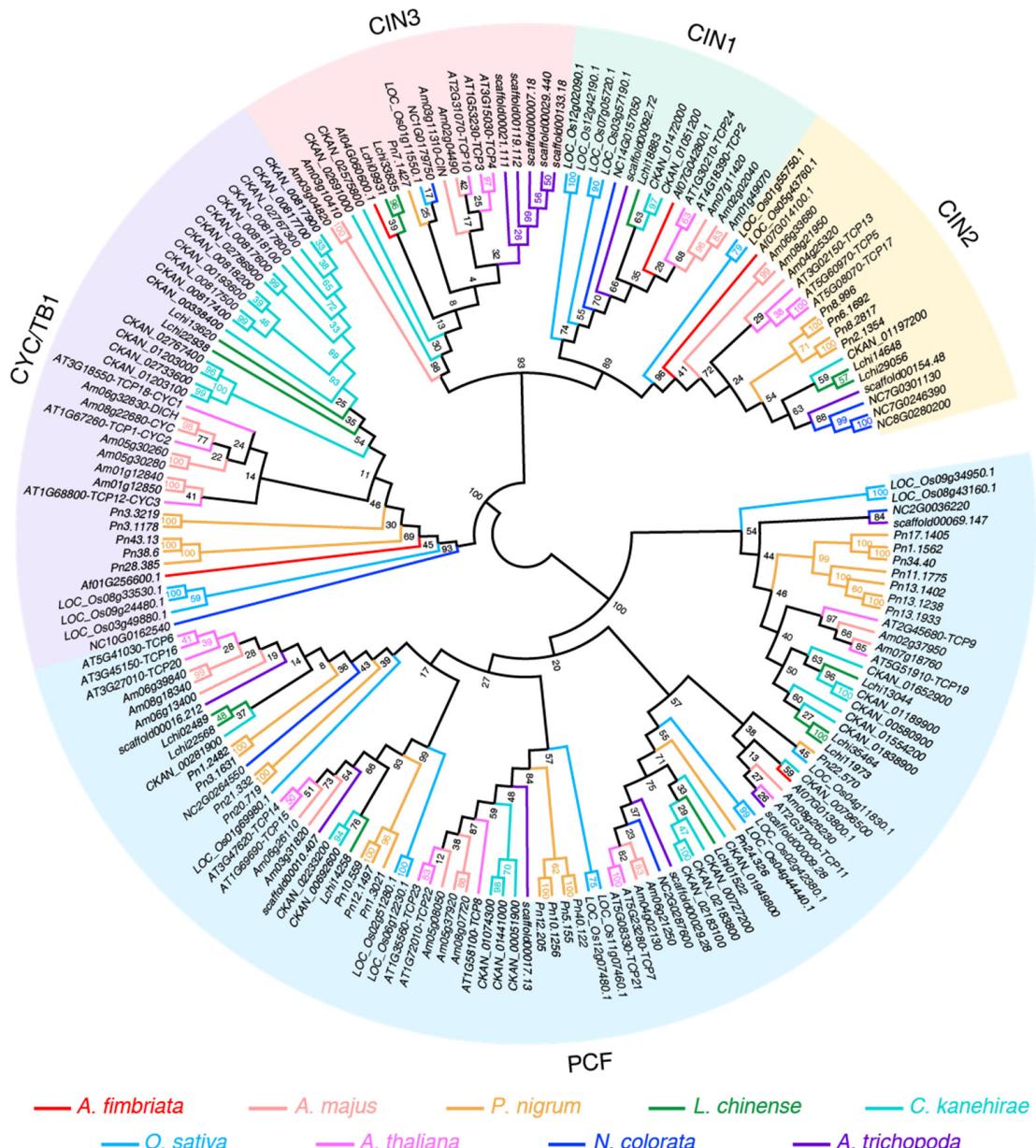
1549 Two *CUC* genes (named *AfCUC2* and *AfCUC3*) were identified in *A. fimbriata*,
1550 which are orthologous to the boundary-specification genes in *A. thaliana* (*CUC1/2*
1551 and *CUC3*, respectively) (Supplementary Fig. 5.5). Our RNA-Seq data showed that
1552 both of them exhibit nearly no expression in various floral organs and leaf tissue
1553 during both pre-anthesis and anthesis stages, except for the weak expression of *CUC2*
1554 in ovary (Fig. 5d). Given the previous knowledge of *CUC* genes in controlling the
1555 specification of organ boundaries¹²⁹, their lack of expression is consistent with the
1556 formation of tubular perianth and the fusion of stamens and the stigmatic region of
1557 carpels.

1558 We identified five TCP genes in *A. fimbriata*. Phylogenetic analysis showed that the
1559 *CYC/TB1* clade includes one *A. fimbriata* gene (named *AfCYC*), all three *CYC* genes
1560 (*CYC1*, *CYC2*, and *CYC3*) in *A. thaliana* and all functional *CYC* genes in *A. majus*,
1561 and that three *A. fimbriata* *CIN* genes (named *AfCIN1*, *AfCIN2*, and *AfCIN3*) were
1562 clustered in three groups, respectively (Supplementary Fig. 5.6). The RNA-Seq data
1563 showed that the *AfCYC* gene is only expressed in the limb of the anthetic flower and
1564 in leaf (Fig. 5e). In contrast, all three *CIN* genes are widely expressed in all floral
1565 organs and the leaves, and exhibit differential expression in perianth basipetally with
1566 the highest expression in the limb region (Fig. 5e). These results, together with their
1567 previously known expression patterns at early developmental stages in *Aristolochia*
1568 *arborea* and *A. fimbriata*^{132,135}, further suggest that both *CYC* and *CIN* genes
1569 contribute to the heterogeneous growth and morphological deformation of the
1570 monosymmetric perianth in *Aristolochia*. Again, further functional genomic studies in
1571 the early flower developmental stages are needed to confirm the roles of these genes
1572 in the establishment of the bilateral symmetry of the perianth in *Aristolochia*.



1573

1574 **Supplementary Fig. 5.5 | Phylogenetic tree of CUC genes from eight species.**
1575 Branches were colored based on the species color scheme on the right of the tree.



1576
1577 **Supplementary Fig. 5.6 | Maximum likelihood tree of TCP genes from nine**
1578 **species.** The phylogenetic tree was constructed using RAxML. Previously recognized
1579 clades were shaded by different colors. Branches were colored based on the species
1580 color scheme below the tree.

1581 **5.5 Floral color and trichome formation genes in *A. fimbriata***

1582 *Aristolochia* flowers often exhibit dull purple-brown color and elaborated inner
1583 epidermal modification^{3,6,8}. To explore the underlying genetic control of dull purple-
1584 brown perianth and a plethora of trichomes in inner perianth epidermis, we explored
1585 the putative homologs of the previously acknowledged anthocyanin biosynthetic

1586 genes and trichome developmental genes, and investigated their expression patterns
1587 and co-expression networks.

1588 **Materials and Methods**

1589 To identify candidate genes responsible for the development of trichomes in the *A.*
1590 *fimbriata* flower, we first cataloged homologs of the genes involved in trichome
1591 formation that were previously identified in other species, such as *A. thaliana*, *S.*
1592 *lycopersicum* and *Cucumis sativus*¹⁴⁰⁻¹⁴³. The related protein sequences of the genes in
1593 these species were used as queries to search against the annotated *A. fimbriata*
1594 proteins using BLASTP with *E*-value cut-off of 10⁻³. The best BLASTP hit in *A.*
1595 *fimbriata* was designated as a homologous candidate gene. Then, phylogenetic
1596 analysis was performed based on orthogroup classification (Supplementary Note 2.4)
1597 to identify the closely related *A. fimbriata* gene with known *A. thaliana* genes.
1598 Finally, the putative genes associated with trichome development in *A. fimbriata* were
1599 inferred based on the integrated evidence from the sequence similarity and
1600 phylogenetic classification.

1601 Anthocyanins are the main floral pigments that give rise to colors ranging from red to
1602 blue¹⁴⁴. To understand the molecular basis of the unique flower color of *A. fimbriata*,
1603 the well-studied anthocyanin biosynthetic genes in *A. thaliana*, *S. lycopersicum* and
1604 *Petunia hybrida*¹⁴⁴⁻¹⁴⁸ were retrieved. Then, candidate genes involved in anthocyanin
1605 biosynthesis in *A. fimbriata* were also identified using the same processes as above
1606 (processes used for the identification of trichome development associated genes in *A.*
1607 *fimbriata*).

1608 For the construction of co-expression networks, we used all RNA-Seq data from 14
1609 samples described above (tissues of flowers at anthesis and pre-anthesis, leaves, and
1610 seedlings with different treatment) and required genes with TPM ≥ 1 in at least one of
1611 the samples to be included in the analysis. Pearson correlation coefficients (PCCs) for
1612 each bidirectional gene pair were calculated to quantify the correlations. Then, we
1613 ranked the PCC values by mutual rank (MR) algorithm to identify the highly
1614 correlated gene pairs. Finally, gene pairs with MR ≤ 300 were referred to as co-
1615 expressed genes¹⁴⁹.

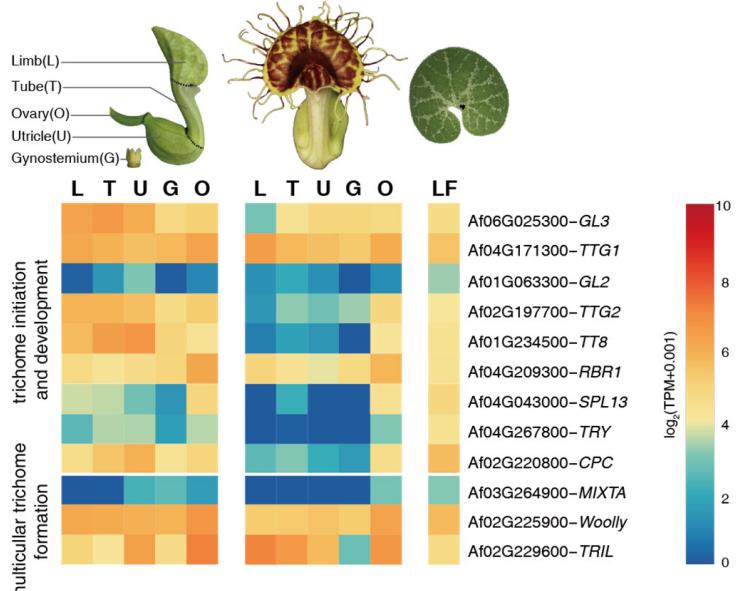
1616 **Results and Discussion**

1617 In *A. fimbriata*, we identified 12 putative homologs of the previously acknowledged
1618 trichome developmental genes^{140,141}, including the trichome initiation and
1619 developmental genes, such as *GLABRA3* (*GL3*), *TRANSPARENT TESTA GLABRA1*
1620 (*TTG1*), *TRANSPARENT TESTA GLABRA2* (*TTG2*), *TRANSPARENT TESTA8* (*TT8*)
1621 and *CAPRICE* (*CPC*), and multicellular trichome formation genes, such as *WIN1*,

1622 *WOOLLY*, and *TRICHOME-LESS (TRIL)* (Supplementary Fig. 5.7). Nearly all of
1623 these genes are single-copy in *A. fimbriata* (Supplementary Table 5.3), suggesting
1624 gene duplication and sub-functionalization are not responsible for the trichome
1625 development in the *A. fimbriata* flowers. The transcriptome data show that most of
1626 these genes show higher expression at pre-anthesis compared to anthesis
1627 (Supplementary Fig. 5.7).

1628 Similarly, fourteen putative anthocyanin biosynthetic genes were identified in *A.*
1629 *fimbriata*, including several critical regulators such as *CHALCONE SYNTHASE*
1630 (*CHS*), *FLAVANONE 3-HYDROXYLASE (F3H)*, *CHALCONE ISOMERASE (CHI)*,
1631 *DIHYDROFLAVONOL 4-REDUCTASE (DFR)*, *ANTHOCYANIDIN SYNTHASE*
1632 (*ANS*), *PRODUCTION OF ANTHOCYANIN PIGMENT 1 (PAPI)*, *TT8*, and *GL3*. We
1633 found that most of these genes are also single-copy in *A. fimbriata* (Supplementary
1634 Table 5.4). However, no homolog was identified to encode the flavonoid 3'5'-
1635 hydroxylase (F3'5'H), a key enzyme for the synthesis of delphinidin-based lilac to
1636 blue anthocyanins, indicating the absence of delphinidin-based anthocyanins in *A.*
1637 *fimbriata* flowers. The genes encoding key enzymes (e.g., *CHS*, *F3H*, *DFR*, and *ANS*)
1638 and several positive regulators show higher expression in perianth (limb, tube, and
1639 utricle) of the flower at the pre-anthesis (developmental stage 8) compared to anthesis
1640 (Fig. 5f). The stage with relatively higher gene expression is same as the previously
1641 known pigmentation stage (flower developmental stage 8) in *A. fimbriata*¹²². These
1642 results suggest these candidate genes are likely involved in the biosynthesis of
1643 cyanidin and/or pelargonidin-based anthocyanins, and therefore may be responsible
1644 for the dull purple-brown color in the perianth of *A. fimbriata* flowers.

1645 By investigating co-expression networks, we found that the MADS-box B-function
1646 genes in *A. fimbriata*, *AP3* and *PI*, are positively co-expressed with three key
1647 anthocyanin biosynthesis candidate genes (*F3H*, *DFR*, *ANS*)¹⁴⁶ and a bHLH
1648 transcription factor (*TT8*) that has been reported to function in both anthocyanin
1649 biosynthesis and trichome formation (Fig. 5g and Supplementary Fig. 5.8)¹⁵⁰. We also
1650 identified CArG-box motifs (MADS-box binding sites) in the promoter regions of
1651 *F3H*, *DFR*, *ANS*, and *TT8* genes (Supplementary Table 5.5). In addition, it also has
1652 been proposed that the expression of *PI* correlates with the occurrence of conical
1653 cellular differentiation and pigment accumulation in *Aristolochia* and other
1654 species^{151,152}. Thus, the results suggest that novel regulatory relationships between the
1655 MADS-box B-function genes and the anthocyanin biosynthesis genes (as well as
1656 several trichome formation genes) might have contributed to the development of the
1657 showy color and elaborated epidermis of the *A. fimbriata* flowers.



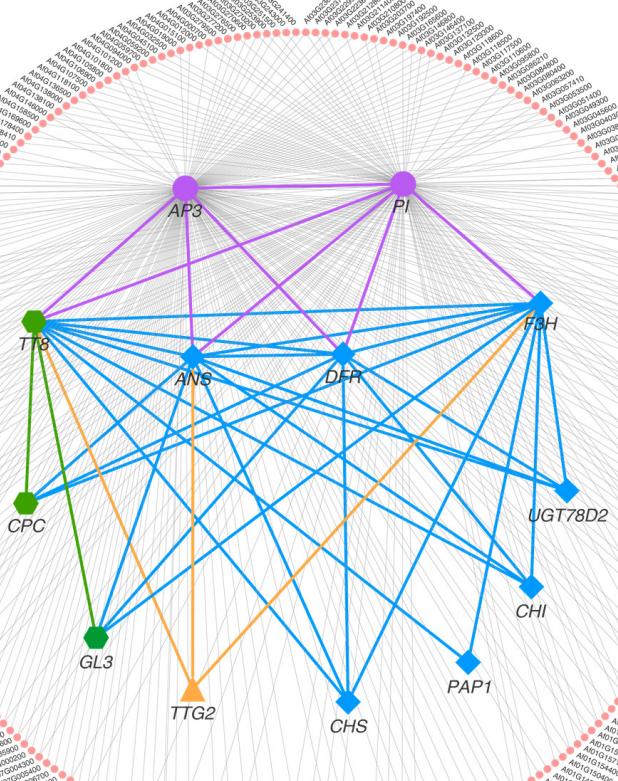
1658

1659

1660

1661 leaf.

Supplementary Fig. 5.7 | Expression patterns of the putative candidate genes involved in trichome formation in the late pre-anthetic and anthetic flower and leaf.



1662

1663

1664

1664 MADS-box B-class genes, genes in
1665 several trichome formation genes

Supplementary Fig. 5.8 | The details of the co-expression network containing MADS-box B-class genes, genes involved in anthocyanin biosynthesis, and several trichome formation genes.

1666 **6 Metabolic gene clusters in *A. fimbriata* and the biosynthesis of terpenoid and**
1667 **aristolochic acids**

1668 *Aristolochia* is well-known for its production of a complex mixture of secondary
1669 metabolites and has been extensively used for traditional medicines⁹. However, a
1670 class of nitrophenanthrene carboxylic acids are naturally synthesized in *Aristolochia*
1671 called aristolochic acids (AAs), which have been demonstrated to be highly
1672 nephrotoxic and carcinogenic to humans¹⁰⁻¹³. Thusly, many herbal medicines are
1673 banned or restricted due to their AAs content. It has been postulated that aporphine
1674 alkaloids may be biogenetical precursors to AAs¹⁵³. Here, we explored the utility of
1675 our genome assembly as a functional genomics resource to investigate the molecular
1676 basis of the secondary metabolites, especially the terpenoids and aristolochic acid.

1677 **6.1 Identification of metabolic gene clusters in the *A. fimbriata* genome**

1678 **Materials and Methods**

1679 First, we annotated genes that were predicted to encode enzymes involving in the
1680 biosynthesis of common secondary metabolites using the KEGG automatic annotation
1681 server (https://www.genome.jp/kaas-bin/kaas_main) with the assignment method of
1682 BBH (bi-directional best hit). In addition, we also employed a metabolic database to
1683 annotate putatively related enzyme genes by the E2P2 package v3.1 using default
1684 parameter settings (<https://gitlab.com/rhee-lab/E2P2/tree/master>). Third, we searched
1685 for potential biosynthetic gene clusters in the *A. fimbriata* genome that are associated
1686 with secondary metabolite biosynthesis using plantiSMASH¹⁵⁴ with default
1687 parameters. Lastly, for the functionally important genes, we further identified the
1688 other gene family members by investigating the orthogroups we constructed in the
1689 above gene family analyses (Supplementary Note 2.4) and requiring these genes with
1690 corresponding Pfam models. Phylogenetic analyses were performed using the
1691 methods described in Supplementary Note 5.1. Amino acid sequence alignments for
1692 the orthologous genes between *A. fimbriata* and the functionally well-studied species
1693 were generated, and the conservation level of previously characterized functional
1694 active residues were carefully examined to better predict the candidates in *A.*
1695 *fimbriata*.

1696 **Results and Discussion**

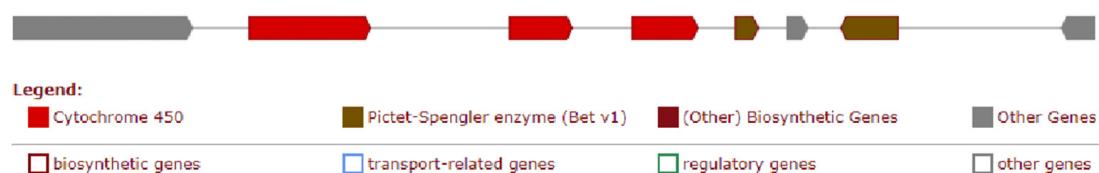
1697 We annotated 1,803 genes across the *A. fimbriata* genome into 20 secondary
1698 metabolism pathways (Supplementary Table 6.1). Four alkaloid biosynthesis
1699 pathways were annotated: ‘Isoquinoline alkaloid biosynthesis’, ‘Indole alkaloid
1700 biosynthesis’, ‘Indole diterpene alkaloid biosynthesis’ and ‘Tropane, piperidine and

1701 pyridine alkaloid biosynthesis'. Two tyrosine anabolism pathways were also
1702 predicted: 'Tyrosine metabolism' and 'Phenylalanine, tyrosine and tryptophan
1703 biosynthesis'. It has been proposed that the potential precursor compounds and
1704 reaction processes related to AAs were in the isoquinoline alkaloid biosynthesis
1705 (ko00950) and tyrosine metabolism (ko00350) pathways¹⁵⁵. Here, we annotated 164
1706 genes in the *A. fimbriata* genome belonging to these two pathways (Supplementary
1707 Table 6.1).

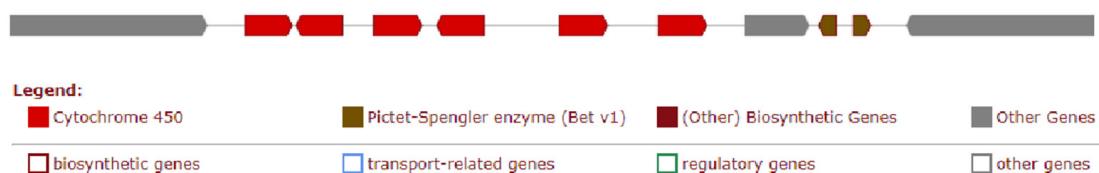
1708 It also has been shown that genes involved in secondary metabolic pathways
1709 sometimes locate together as biosynthetic gene clusters (BGCs) in plant
1710 genomes^{156,157}. By using the computational toolkit plantiSMASH¹⁵⁴, we detected 33
1711 BGCs related to various plant secondary metabolic pathways (Supplementary Fig. 6.1
1712 and Supplementary Table 6.2), which included five alkaloid-, two polyketide-, three
1713 saccharide-, two saccharide-alkaloid-, four saccharide-terpene-, seven terpene-, two
1714 terpene-alkaloid-related, and eight putative gene clusters. The genomic regions of
1715 these BGCs spanned from 19.47 to 358.74 Kb and contained multiple core protein
1716 domains related to various secondary metabolism. The large proportion of the
1717 annotated alkaloid (9/33) and terpene-related (14/33) BGCs might be associated with
1718 the enriched production of alkaloid and terpenoid compounds in *A. fimbriata*.

1719 Polyketides are compounds that possess various critical biological activities, including
1720 antibiotic, anticancer, antifungal, immunosuppressive, and neurotoxic¹⁵⁸⁻¹⁶⁰
1721 properties, and have been considered as a useful source of pharmaceutical
1722 agents^{159,161}. The polyketides could be one of the key compounds for its
1723 pharmaceutical usage. Polyketide synthases (PKSs) are the key enzymes for the
1724 biosynthesis of a large class of pharmacologically valuable compounds, which have
1725 been classified into three types¹⁶⁰. Based on the annotated Pfam model and conserved
1726 domains, we found four genes belonging to type III PKS in *A. fimbriata*, and the
1727 Af04G164800 gene was annotated in the polyketide gene cluster (BGC #6) which
1728 also showed the highest sequence similarity to the chalcone synthase (CHS) gene
1729 AT5G13930 in *Arabidopsis* (Supplementary Fig. 6.3). Phylogenetic analysis further
1730 confirmed the close relationship of *Af04G164800* and the known *CHS* gene
1731 (AT5G13930) in *Arabidopsis* in Clade P1 (Supplementary Fig. 6.2). The sequence
1732 alignment showed that the previously characterized active sites of the Cys-His-Asn
1733 (CHN) catalytic triad and Phe in *Medicago sativa* are conserved in *A. fimbriata* and *A.*
1734 *thaliana*¹⁶² (Supplementary Fig. 6.3). Based on these results, we propose that the
1735 *Af04G164800* gene is mostly likely the candidate gene encoding chalcone synthase in
1736 *A. fimbriata*, which could be functionally tested further if interested.

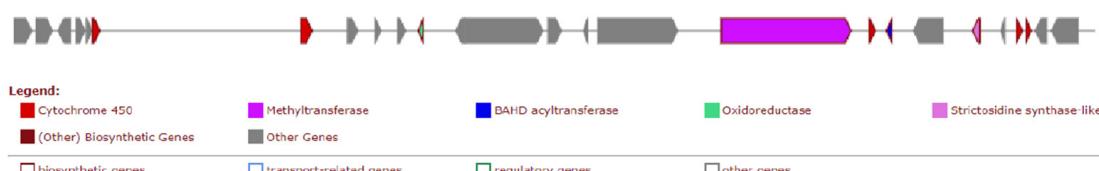
1. chr02 - Gene Cluster 1. Type = alkaloid. Location: 4804428 - 4831122 nt.



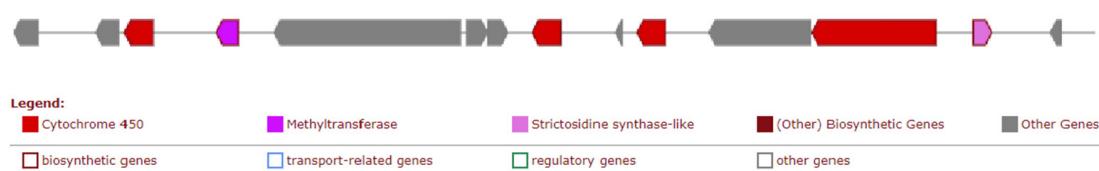
2. chr02 - Gene Cluster 2. Type = alkaloid. Location: 40032211 - 40062418 nt.



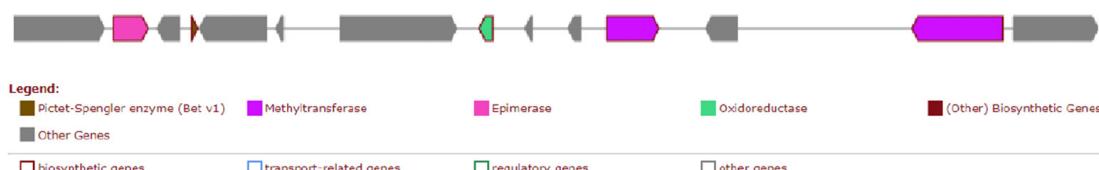
3. chr03 - Gene Cluster 3. Type = alkaloid. Location: 31357932 - 31632713 nt.



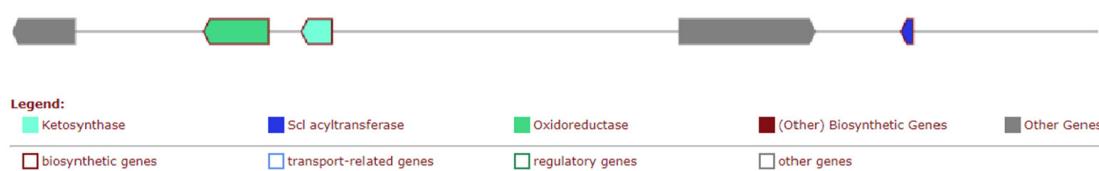
4. chr03 - Gene Cluster 4. Type = alkaloid. Location: 31671702 - 31730008 nt.



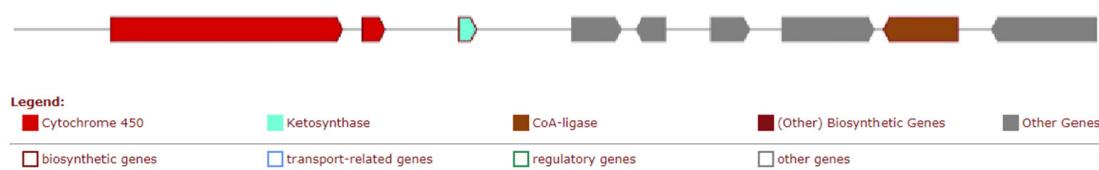
5. chr05 - Gene Cluster 5. Type = alkaloid. Location: 2443496 - 2539325 nt.



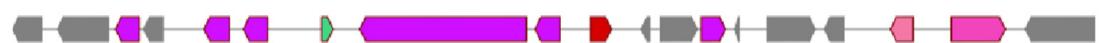
6. chr04 - Gene Cluster 6. Type = polyketide. Location: 24900798 - 24945969 nt.

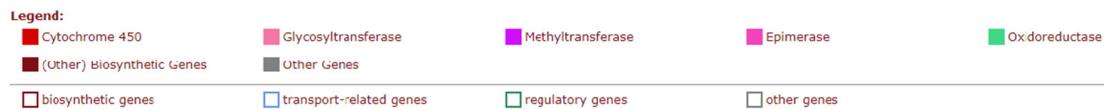


7. chr07 - Gene Cluster 7. Type = polyketide. Location: 21455618 - 21533274 nt.



8. chr02 - Gene Cluster 8. Type = saccharide. Location: 1397150 - 1485140 nt.

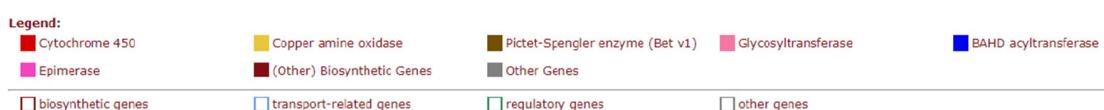




9. chr07 - Gene Cluster 9. Type = saccharide. Location: 18562708 - 18582182 nt.



10. chr02 - Gene Cluster 10. Type = saccharide-alkaloid. Location: 35476718 - 35678337 nt.



11. chr06 - Gene Cluster 11. Type = saccharide-alkaloid. Location: 4570830 - 4644656 nt.



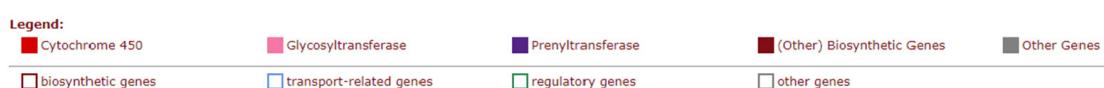
12. chr03 - Gene Cluster 12. Type = saccharide-terpene. Location: 1454222 - 1490131 nt.



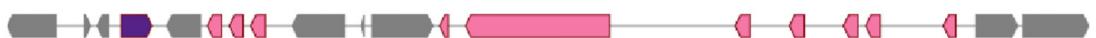
13. chr06 - Gene Cluster 13. Type = saccharide-terpene. Location: 3350754 - 3416888 nt.



14. chr06 - Gene Cluster 14. Type = saccharide-terpene. Location: 19861542 - 19948891 nt.



15. chr07 - Gene Cluster 15. Type = saccharide-terpene. Location: 19755066 - 19856792 nt.



Legend:
█ Glycosyltransferase █ Prenyltransferase █ (Other) Biosynthetic Genes █ Other Genes
█ biosynthetic genes █ transport-related genes █ regulatory genes █ other genes

16. chr01 - Gene Cluster 16. Type = terpene. Location: 5547305 - 5586781 nt.



Legend:
█ Terpene synthase █ Lipoxygenase █ (Other) Biosynthetic Genes █ Other Genes
█ biosynthetic genes █ transport-related genes █ regulatory genes █ other genes

17. chr02 - Gene Cluster 17. Type = terpene. Location: 29635093 - 29814343 nt.



Legend:
█ Cytochrome 450 █ Prenyltransferase █ PRISE enzymes █ (Other) Biosynthetic Genes █ Other Genes
█ biosynthetic genes █ transport-related genes █ regulatory genes █ other genes

18. chr02 - Gene Cluster 18. Type = terpene. Location: 31658540 - 31727527 nt.



Legend:
█ Cytochrome 450 █ Terpene synthase █ Aminotransferase █ (Other) Biosynthetic Genes █ Other Genes
█ biosynthetic genes █ transport-related genes █ regulatory genes █ other genes

19. chr02 - Gene Cluster 19. Type = terpene. Location: 32427377 - 32579422 nt.



Legend:
█ Terpene synthase █ Dioxygenase █ (Other) Biosynthetic Genes █ Other Genes
█ biosynthetic genes █ transport-related genes █ regulatory genes █ other genes

20. chr02 - Gene Cluster 20. Type = terpene. Location: 32662073 - 32803745 nt.



Legend:
█ Cytochrome 450 █ Terpene synthase █ BAHD acyltransferase █ (Other) Biosynthetic Genes █ Other Genes
█ biosynthetic genes █ transport-related genes █ regulatory genes █ other genes

21. chr05 - Gene Cluster 21. Type = terpene. Location: 2122765 - 2187644 nt.



Legend:
█ Cytochrome 450 █ Terpene synthase █ (Other) Biosynthetic Genes █ Other Genes
█ biosynthetic genes █ transport-related genes █ regulatory genes █ other genes

22. chr06 - Gene Cluster 22. Type = terpene. Location: 23934688 - 23994410 nt.





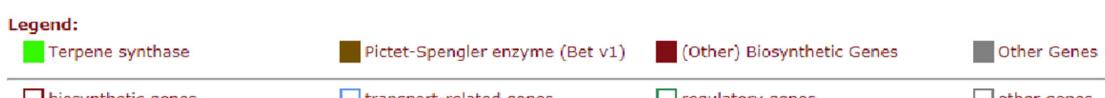
23. chr07 - Gene Cluster 23. Type = terpene. Location: 585971 - 638309 nt.



24. chr01 - Gene Cluster 24. Type = terpene-alkaloid. Location: 9719385 - 9803606 nt.



25. chr05 - Gene Cluster 25. Type = terpene-alkaloid. Location: 1938278 - 2004826 nt.



26. chr02 - Gene Cluster 26. Type = putative. Location: 126101 - 277600 nt.



27 chr02 - Gene Cluster 27 Type = putative Location: 982170 - 1088396 nt



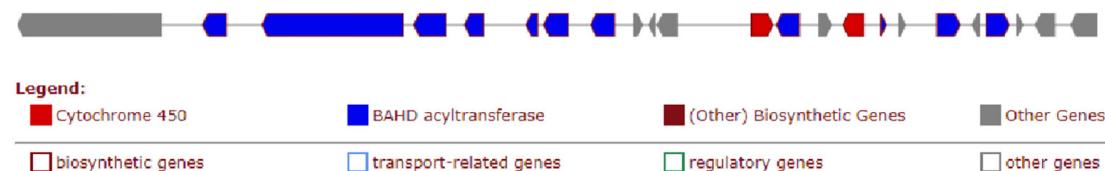
28 chr03 - Gene Cluster 28 Type = putative Location: 3558163 - 3658935 nt



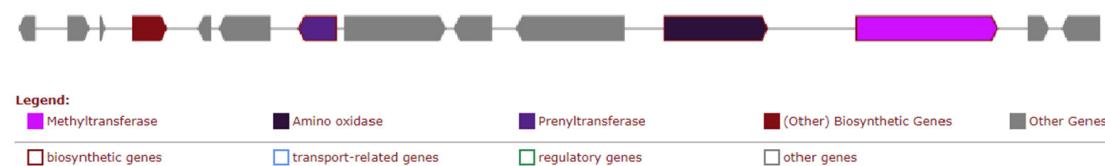
29 chr03 - Gene Cluster 29 Type = putative Location: 23666975 - 24006353 nt



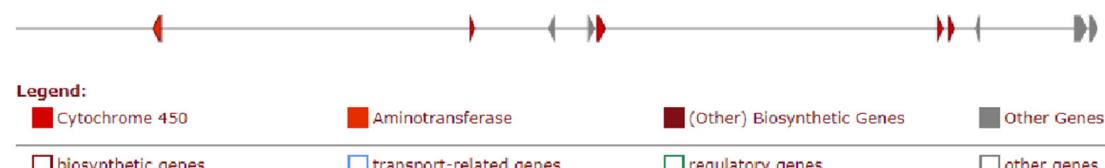
30. chr03 - Gene Cluster 30. Type = putative. Location: 33771454 - 33829925 nt.



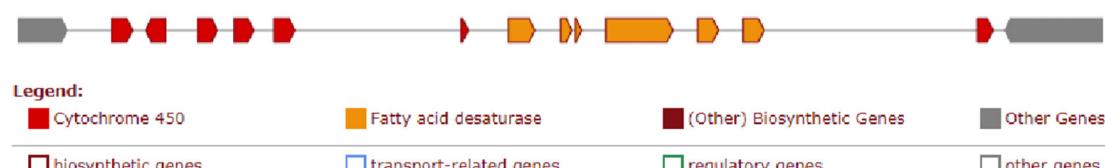
31. chr03 - Gene Cluster 31. Type = putative. Location: 36266912 - 36332658 nt.



32. chr04 - Gene Cluster 32. Type = putative. Location: 23761286 - 24120027 nt.

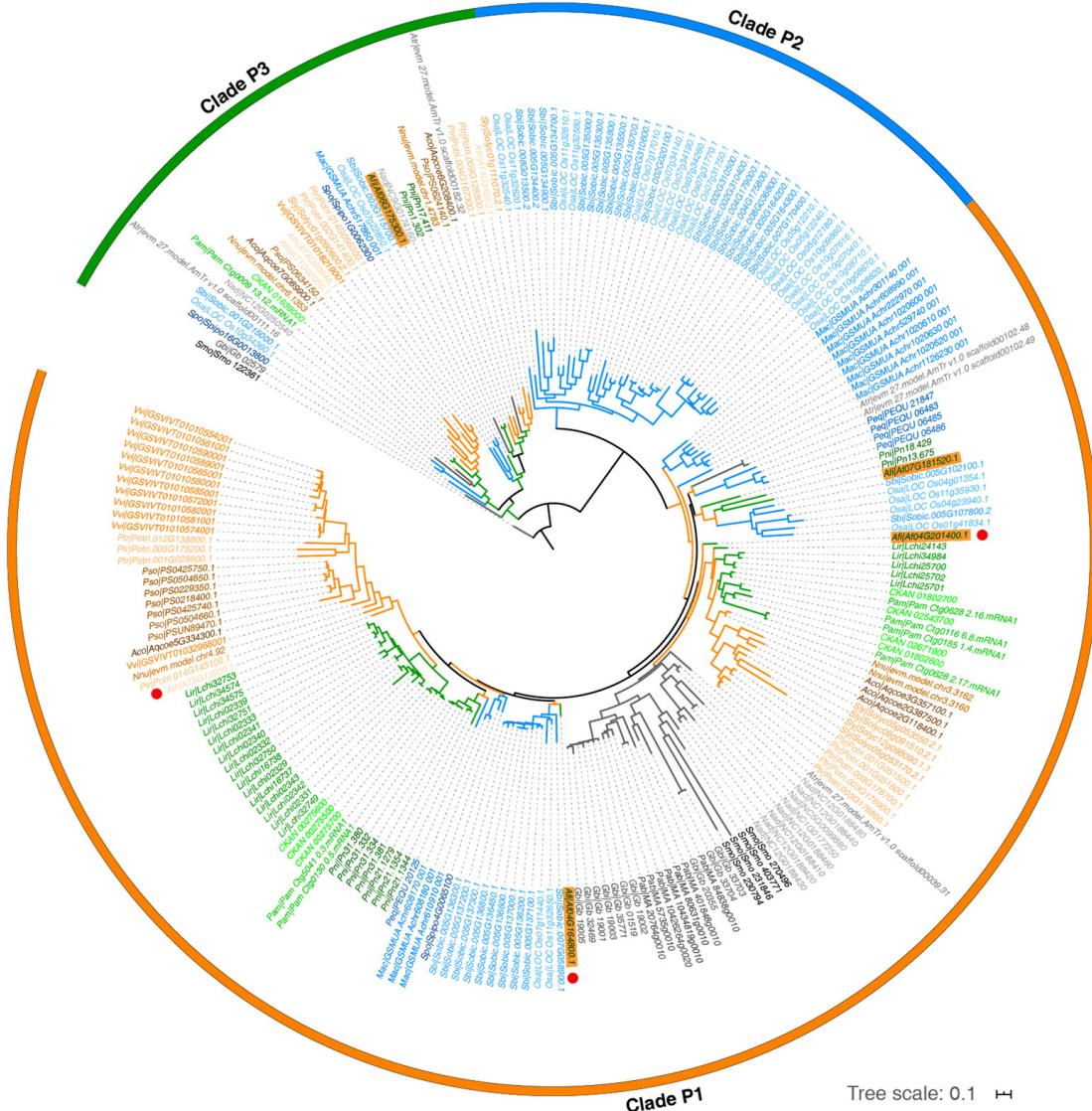


33. chr07 - Gene Cluster 33. Type = putative. Location: 21012811 - 21100590 nt.



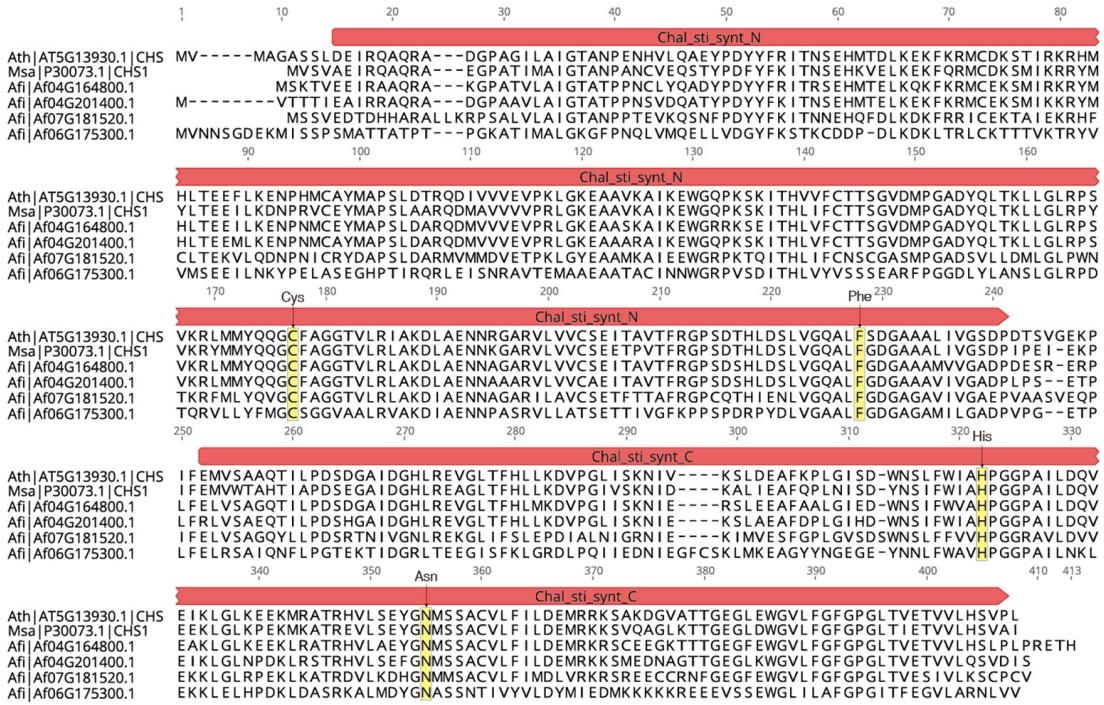
1741

1742 **Supplementary Fig. 6.1 | Visualization of 33 BGCs identified by plantiSMASH.**
1743 Various core enzymes were annotated, and the related genes were colored according
1744 to the enzymatic classes.



1745

1746 **Supplementary Fig. 6.2 | Phylogeny of the type III PKS gene family.** Branches of
 1747 the phylogeny were colored: orange-eudicots; blue-monocots; green-magnoliids;
 1748 gray-ANA grade, gymnosperms and lycophyte. The previously identified functional
 1749 gene in *A. thaliana* and the two *A. fimbriata* candidate genes are marked with red
 1750 dots. The *A. fimbriata* genes are highlighted with an orange background.



1751

1752 **Supplementary Fig. 6.3 | Multiple sequence alignment of CHS proteins in *M. sativa*, *A. thaliana*, and *A. fimbriata*.** The previously characterized active catalytic residues (Cys-His-Asn triad and Phe) are highlighted in yellow. The identified four *Aristolochia* protein sequences seem to be well-conserved compared to their orthologs in *M. sativa* and *A. thaliana*.

1757 6.2 Floral scent and terpenes

1758 Materials and Methods

1759 The complex and diverse floral scent composition is crucial for the deceptive
 1760 pollination systems in *Aristolochia*^{3,7,163}. To investigate the floral volatile production
 1761 of *A. fimbriata*, we collected the newly opened flowers for gas chromatography-mass
 1762 spectrometry (GC-MS) analysis, with the added 0.0825 µg of 3-Octanol as an internal
 1763 standard. Then, the samples were incubated at 40 °C for 30 min. The volatiles were
 1764 further extracted using SPME fiber with 50/30 µm
 1765 divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS) (Supelco Co.,
 1766 Bellefonte, PA, USA). Finally, GC-MS analysis was conducted on an Agilent 7890B
 1767 gas chromatograph coupled to a mass spectrometer (Agilent 7000D, Santa Clara, CA,
 1768 USA) with a fused silica capillary column (HP-5MS) coated with
 1769 polydimethylsiloxane (19091S-433UI) (30m × 0.25 mm internal diameter, 0.25 µm
 1770 film thickness). The oven temperature was programmed to start at 40 °C for 3 min,
 1771 and then ramped to 130 °C at a rate of 5 °C min⁻¹, followed by a second ramp to

1772 156 °C at a rate of 2 °C min⁻¹, the final ramp to 280 °C at a rate of 10 °C min⁻¹. Three
1773 biological replicates were conducted for the GC-MC analysis.

1774 Terpenes are one of the major groups of flower volatile compounds in numerous
1775 plants^{164,165}, and monoterpenoids and sesquiterpenoids have also been commonly
1776 detected in several *Aristolochia* species^{7,166,167}. Terpene synthases (TPSs) are pivotal
1777 enzymes for the biosynthesis of terpenes^{164,165}, and here we explored the number and
1778 expression of TPS genes to infer their putative roles in *A. fimbriata*. To identify TPS
1779 genes in *A. fimbriata* be consistent with the other studies, we performed profile HMM
1780 searches against the predicted protein sequences by using hmmsearch in HMMER
1781 v3.3¹¹⁴ (--domE 0.001). To be selected, putative TPS genes were required to match at
1782 least one of the Pfam models (PF01397 and PF03936), which correspond to the
1783 conserved domains localized at the N and C termini of known terpene synthases
1784 (TPSs) respectively¹⁶⁸. We also identified TPS genes in *A. trichopoda*, *N. colorata*, *O.*
1785 *sativa* and three other magnoliids (*C. kanehirae*, *P. nigrum*, and *L. chinense*) and
1786 compared them with the genes in previously published reports^{55,59,70,168}. The well-
1787 studied and annotated subfamilies of TPS genes in *A. thaliana*¹⁶⁸ were used as
1788 reference for phylogenetically classifying these TPS genes into corresponding
1789 subfamilies. In addition, the TPS genes located in the annotated biosynthetic gene
1790 clusters were further identified, and their expression patterns were investigated as in
1791 Supplementary Note 5.3.

1792 Results and Discussion

1793 GC-MS analyses revealed a complex of *A. fimbriata* flower scent, including
1794 terpenoids (sesquiterpenes and monoterpenoids), fatty acid derivatives and
1795 benzenoids. Among the terpenoids, the sesquiterpenes are the most abundant volatile
1796 compounds (Fig. 6a), while the diterpenoids were not detected in the *A. fimbriata*
1797 flower volatiles. Then, the candidate TPS genes responsible for the biosynthesis of
1798 sesquiterpenes and monoterpenoids in *A. fimbriata* were further identified based on
1799 comprehensive analyses of phylogeny, associated biosynthetic gene clusters, and
1800 spatio-temporal expression patterns.

1801 After comprehensive TPS family annotation, 41 genes were identified in the *A.*
1802 *fimbriata* genome, and 16 of them were located in the 13 terpene-related biosynthetic
1803 gene clusters noted above (Supplementary Fig. 6.1 and Supplementary Table 6.2).
1804 The TPS genes identified in four of six other species (*A. trichopoda*, *N. colorata*, *O.*
1805 *sativa* and *C. kanehirae*) were largely consistent with previous report^{55,59,70,168}. The
1806 TPS genes in the other two species (*L. chinense* and *P. nigrum*) were newly
1807 annotated. Phylogenetic analysis classified the TPS genes into five previously
1808 recognized subfamilies, representing TPS-a, TPS-b, TPS-c, TPS-e/f, and TPS-g based
1809 on the well-studied *A. thaliana* TPS genes (Fig. 6b and Supplementary Table 6.3).

1810 There are 21 genes in *A. fimbriata* clustered in subfamily of TPS-a, which are usually
1811 encoding sesquiterpene synthases¹⁶⁴. Notably, the *Af06G158900* gene in the TPS-a
1812 subfamily showed extremely high and increased expression in the utricle of anthetic
1813 flowers compared to pre-anthetic flowers and leaves, which is consistent with the
1814 abundant component of sesquiterpenes in the newly opened flower volatiles (Fig. 6c).
1815 Moreover, the *Af06G158900* gene was also identified in the terpene-related gene
1816 cluster (BGC #22), consisting of genes encoding terpene synthase, dioxygenase, and
1817 prenyltransferase (Fig. 6d). Another gene is *Af06G160300* which also showed high
1818 expression in the limbs of flowers at anthesis, while almost no expression in the pre-
1819 anthetic floral organs and leaves. Together with the previous knowledge that the
1820 osmophores in the limb and utricle may be the major floral fragrance glands emitting
1821 floral scents^{3,122,169}, the *Af06G158900* and *Af06G160300* genes, especially the former,
1822 are likely the main sesquiterpene synthase genes in *A. fimbriata*.

1823 We also examined the monoterpene synthase genes for the other main flower volatile
1824 compounds in *A. fimbriata*, and the subfamilies of TPS-g and TPS-b were proposed
1825 encoding monoterpene synthases^{164,168,170}. We found the subfamily TPS-g has recently
1826 expanded in *A. fimbriata*, resulting in 9 TPS-g genes (Fig. 6b). The *Af01G154900* and
1827 *Af06G245300* among these nine TPS-g genes showed increased expression in the
1828 utricle and limb of anthetic flowers, respectively, compared to the counterparts of pre-
1829 anthetic flowers and the leaves (Fig. 6c). Notably, the *Af01G154900* gene in the TPS-
1830 g clade was also annotated in BGC #24 (Fig. 6d). These results indicated that the
1831 *Af01G154900* gene might be one of the key candidate genes encoding monoterpene
1832 synthases in *A. fimbriata*. Another subfamily encoding monoterpene synthases is
1833 TPS-b, and 6 genes in *A. fimbriata* were identified. A pair of tandem duplicates
1834 (*Af05G030400* and *Af05G030410*) shows relatively stable expression in leaf tissue
1835 and pre-anthetic and anthetic flowers (Fig. 6c). Functional studies could be employed
1836 in the future to validate the function of these putative candidate genes, and ultimately
1837 help to better understand the deceptive pollination system in *Aristolochia*.

1838 It has been suggested that the genes in TPS-c and TPS-e/f clades are responsible for
1839 the biosynthesis of diterpenoids^{164,168}. We found that the *A. fimbriata* genes in these
1840 clades showed very weak expression at both pre-anthetic and anthetic flowers (Fig.
1841 6b,c). This finding suggests a lack of diterpenoids in *A. fimbriata* flower volatile
1842 compounds, which is consistent with the result of GC-MS analysis.

1843 **6.3 Liquid Chromatograph Mass Spectrometer (LC-MS)-based metabolomic
1844 analysis and reconstruction of aristolochic acid I biosynthesis pathway**

1845 **Materials and Methods**

1846 AA I and II were found as the most common AA analogues, and these are also
1847 responsible for the nephrotoxic and carcinogenic effects⁴. We performed an LC-MS-
1848 based metabolomic analysis for the root, stem, leaf, and fruit from one-year-old *A.*
1849 *fimbriata* plants. 50 mg of each dried tissue were processed for the HPLC-DAD-
1850 ESIMS/MS measurements. AAs were separated by UPLC (Waters, ACQUITY)
1851 equipped with an ACQUITY UPLC HSS T3 column (Waters) and detected by
1852 MS/MS using a Triple Quad Xevo TQ-S (Waters) mass spectrometer. The mobile
1853 phase consists of buffer A (5 mM ammonium acetate and 0.1% formic acid) and
1854 buffer B (100% acetonitrile). AAs were qualified using the ion mass transitions of m/z
1855 324.1/237 and 324.1/280 for AA I and m/z 329/238 and 329/268 for AA II,
1856 respectively, and the base ions were ammonium adduct ions [M + NH₄]⁺. For
1857 quantitative analysis, we used a higher abundance of the adduct ion mode. Standard
1858 curves were generated by running a concentration series of pure commercial
1859 aristolochic acids. The content of aristolochic acids in each sample was then
1860 calculated by fitting the peak areas to the standard curves.

1861 **Results and Discussion**

1862 From the LC-MS investigation, we found an AA I peak at a retention time of 3.56
1863 minute and the AA II peak at retention time of 2.99 minute in the standard reference
1864 curves. For our experimental samples, we observed the AA I peak at retention time of
1865 3.56 minute in the tissues of root, stem, leaf, and fruit of *A. fimbriata*, but no AA II
1866 peak was observed (Extended Data Fig. 9). The results indicated that *A. fimbriata*
1867 contains AA I, but likely no AA II, which is consistent with a previous report⁴.
1868 Therefore, we mainly focused on the reconstruction of the biosynthetic pathway of
1869 AA I.

1870 It has been indicated that several key intermediate such as benzylisoquinoline from
1871 norlaudanosoline to orientaline, orientalinone, orientalinol, prestephanine, stephanine
1872 may be involved in the biosynthesis of AA I^{153,171,172}. Norlaudanosoline could be
1873 formed via the condensation of dopamine and 3,4-dihydroxyphenylacetaldehyde by
1874 norcoclaurine synthase (NCS), while the incorporation of tyrosine, dopa, dopamine,
1875 and 3,4-dihydroxyphenylacetaldehyde to norlaudanosoline could be similar to the first
1876 part of the benzylisoquinoline alkaloids (BIAs) biosynthetic pathway (tyrosine
1877 metabolism pathway: ko000350 and isoquinoline alkaloid biosynthesis pathway:
1878 ko00950)¹⁷³. Through integrating the previous hypothesis^{153,171,172} and the KEGG
1879 annotations, we reconstructed the AA I biosynthesis pathway with the potential
1880 biogenetic route of L-tyrosine→L-dopa→dopamine→3,4-
1881 dihydroxyphenylacetaldehyde→(S)-norlaudanosoline→(S)-6-O-
1882 methylnorlaudanosoline→(S)-nororientaline→orientaline→orientalinone→
1883 orientalinol→prestephanine→stephanine→aristolochic acid I (Fig. 6e).

Furthermore, we identified the putative genes encoding key enzymes involved in the AA I biosynthesis pathway based on previous studies^{173,174}, including tyrosinases (TYR), tyrosine/dopa decarboxylases (TYDC), monoamine oxidases (MAO), norcoclaurine synthase (NCS), norcoclaurine 6-O-methyltransferase (6OMT), 6-O-methylnorlaudanosoline 5'-O-methyltransferase (5'OMT), coclaurine N-methyltransferase (CNMT), and possibly cytochrome P450s (CYPs). Those enzymes in other species mostly were verified to assign their functions using heterologous expression of microbial systems (Supplementary Table 6.4). Based on seed motif from the Pfam database¹¹³, we identified the candidate genes encoding key enzymes involved in the AA I biosynthesis pathway, including 7 TYRs, 2 TYDCs, 6 MAOs, 26 NCSs, 29 OMTs and 8 CNMT (Supplementary Table 6.5). In addition, three cytochrome P450 (CYPs) subfamilies including CYP80, CYP719, and CYP82 might play key oxidative roles in alkaloid metabolism¹⁷⁴⁻¹⁷⁸, which could catalyze the formation of C–C or C–O bonds in the BIA pathway^{173,179}. We speculate that they could be involved in the reactions of orientaline→orientalinone and prestephanine→stephanine in the aristolochic acid I pathway. Here, we identified 7 CYP80 genes, 11 CYP719 genes, and 23 CYP82 genes, which are potentially involved in alkaloid biosynthesis in *A. fimbriata* (Supplementary Table 6.6).

6.4 Identifying the key candidate genes based on phylogenetic classification and the sequence conservation of key residues

Materials and Methods

To predict key candidate genes encoding the associated enzymes, we integrated phylogeny, gene expression, conservation of critical enzyme residue analyses. We first performed profile hidden Markov model (HMM) searches based on the seed motif from the Pfam database¹¹³ against the annotated amino acid sequences using hmmsearch in HMMER v3.3¹¹⁴ (--domE 0.001). Then, sequence alignments and phylogenetic analysis were constructed as described in the above sections (Supplementary Note 5.1). In the following analyses, we specifically compared the candidate genes in *A. fimbriata* to opium poppy (*P. somniferum*), because *P. somniferum* has emerged as a model system to study BIA metabolism. Many functional genes and their corresponding key residues for BIA biosynthesis in *P. somniferum* have been well characterized¹⁸⁰⁻¹⁸³, and thus they were used as references to infer the associated candidate genes in *A. fimbriata*.

Expression profiles of key genes were further examined by qPCR. The samples of root, stem, leaf, flower, and fruit collected from 1-year-old *A. fimbriata* plants and one-month-old seedlings were used for gene expression profiling using qRT-PCR. Three independent plants and seedlings were used as biological replicates. Total

1921 RNAs were isolated using an RNA Extraction Kit (Qiagen, Germany), and cDNAs
1922 were synthesized using the FastKing RT Kit (TIANGEN BIOTECH, China). Primers
1923 were designed and further checked for their specificity (Supplementary Table 6.7).
1924 qPCRs were performed on a QTOWER (Jena, Germany) using Real Master Mix
1925 SYBR Green II dye (Takara, Japan) according to the manufacturer's instructions.
1926 Actin was used as the internal reference gene for normalization. Relative gene
1927 expression was performed using the comparative $2^{-\Delta\Delta C_t}$ method¹⁸⁴. Three technical
1928 replicates and three independent biological experiments were performed.

1929 Results and Discussion

1930 **Coclaurine N-Methyltransferase (CNMT).** The intermediate part of the AA
1931 biosynthetic pathway comprises several key steps from norlaudanosoline to
1932 orientaline, which seem to involve methylation by an N-methyltransferase (NMT).
1933 Coclaurine N-methyltransferases (CNMTs) are a subclass of (S)-adenosyl-L-
1934 methionine (SAM)-dependent NMTs [E.C.2.1.1.140], which can catalyze the transfer
1935 of a methyl group from SAM to the amino group of coclaurine and several
1936 structurally similar alkaloids such as (S)-norcoclaurine, (R, S)-6-O-methylcoclaurine
1937 and (R, S)-norlaudanosoline¹⁸⁵⁻¹⁸⁷. Therefore, we predicted that CNMTs may catalyze
1938 the reactions from (S)-nororientaline to orientaline.

1939 Eight putative genes encoding CNMTs were identified in the *A. fimbriata* genome,
1940 and a set of six tandem duplicates that phylogenetically clustered with a previously
1941 acknowledged functional gene of *P. somniferum* (Supplementary Fig. 6.4). Notably,
1942 these six genes were also annotated in a saccharide-type gene cluster (BGC #8). In
1943 addition, we further investigated the amino acid sequence alignment of these six *A.*
1944 *fimbriata* genes with the functional *P. somniferum* gene and found they all exhibit
1945 high conservation for the functional motif A of plant-SAM and activation sites¹⁸⁸
1946 (Supplementary Fig. 6.5). Our recent study isolated and molecularly characterized a
1947 putative CNMT from *A. fimbriata*¹⁸⁸, which is the same gene as *Af02G021500*
1948 identified here. In particular, we found that the *Af02G022300* gene has the highest
1949 sequence conservation for functional sites (Supplementary Fig. 6.5), which should be
1950 included for functional validations in the future.

1951 **Norcoclaurine 6-O-methyltransferase (6OMT).** Plant O-methyltransferases
1952 (OMTs) constitute a large family of enzymes that can methylate the oxygen atom of a
1953 variety of secondary metabolites, including phenylpropanoids, flavonoids, and
1954 alkaloids¹⁸⁹. It has been reported OMTs could be classified into two major groups,
1955 including PL-OMT I comprising the caffeoyl CoA OMTs and carboxylic acid OMTs,
1956 and PL-OMT II acting on a diverse group of metabolites such as hydroxycinnamic
1957 acids, flavonoids, and alkaloids¹⁸⁹. Here, we mainly focused on PL-OMT II group and
1958 the alkaloid-related genes that were previously characterized in BIA pathways in

1959 *Papaver*¹⁸⁰. Based on the reference genome annotation of *A. fimbriata* and the known
1960 KEGG pathways, we predicted that 6OMT [EC 2.1.1.128] might be involved in the
1961 biosynthetic pathway of AA I, which could catalyze the reaction of S-Adenosyl-L-
1962 methionine + (S)-Norlaudanosoline ⇌ S-Adenosyl-L-homocysteine + (S)-6-O-
1963 Methylnorlaudanosoline.

1964 Based on the reconstructed orthogroups and screening for the genes with conserved
1965 Pfam model, we identified 478 OMT genes belonging to PL-OMT II in the 22
1966 species, in which the *A. fimbriata* genome contains 29 genes. Phylogenetic analysis
1967 classified them into four subclades (II-A, II-B, II-C, II-D) (Supplementary Fig. 6.6),
1968 which is consistent with previous reports, and the genes in subclades were involved in
1969 different biosynthesis¹⁸⁹⁻¹⁹¹. It has been suggested that majority of BIA OMTs fall
1970 within subclade II-D (e.g., 6OMT, 4OMT, SOMT)¹⁸⁹⁻¹⁹¹. Here, we found that nine
1971 genes in *A. fimbriata* and the previously characterized 6OMT gene in *P. somniferum*
1972 were classified within the PL-OMT II-D subclade (Supplementary Fig. 6.6). Notably,
1973 three of nine genes in *A. fimbriata* were also annotated in three associated BGCs
1974 (*Af03G181700* in BGC #29; *Af03G218000* in BGC #3; *Af03G219800* in BGC #4),
1975 that were annotated as putative-, alkaloids-, and alkaloids-type clusters, respectively.

1976 In addition, the amino acid sequence alignment revealed that these nine genes all
1977 exhibit high sequence similarity to the previously characterized *P. somniferum* gene
1978 for the motif regions and the activation sites (Supplementary Fig. 6.7). In particular,
1979 the amino acid sequence, the position of motifs A, B, and C, and the activation sites
1980 (e.g., I₂₈₅, K₂₈₈, P₃₁₁) in *Af04G180700*, *Af03G181700*, *Af03G218000* and
1981 *Af03G219800* genes are highly conserved with that of the *P. somniferum*^{181,182}.
1982 Together, our results indicate that these four genes (*Af04G180700*, *Af03G181700*,
1983 *Af03G218000*, and *Af03G219800*) are likely the functional genes encoding 6OMTs.

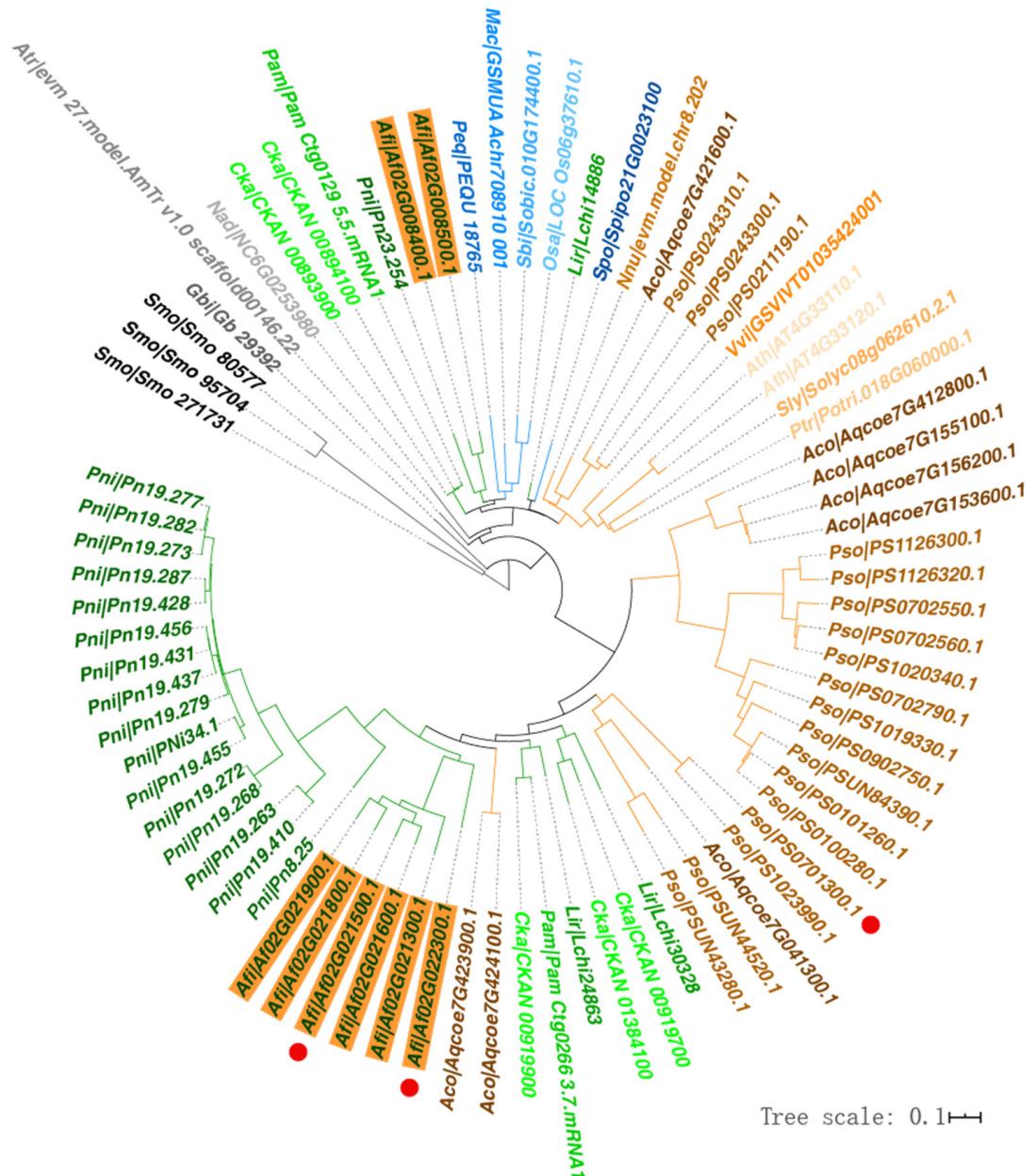
1984 **Norcoclaurine synthase (NCS).** NCS catalyzes the first committed step in the
1985 biosynthesis of BIAs in plants¹⁹². NCS genes have been acknowledged to have
1986 evolved from the pathogenesis-related 10 (PR10)/Bet v 1 ancestor, and their
1987 biochemical activities have been characterized in opium poppy^{193,194} and yellow
1988 meadow-rue (*Thalictrum flavum*)¹⁹⁵. Two subfamilies of NCSI and NCSII have been
1989 defined¹⁹⁶. The NCSII genes are universal in plants, while NCSI genes are crucial for
1990 BIA biosynthesis in certain families, including Ranunculaceae, Papaveraceae,
1991 Berberidaceae, and Nelumbonaceae^{192,196}.

1992 We found 26 genes belonging to the PR10/Bet v1 family in *A. fimbriata*, and several
1993 genes are located closely in the chromosomes and likely resulted from tandem
1994 duplications (*Af02G076800*, *Af02G077000*, *Af02G077200*, *Af02G077300*,
1995 *Af02G077500*; *Af02G263900*, *Af02G264000*; *Af05G030500*, *Af05G030600*,
1996 *Af05G030700*, *Af05G030800*). Based on phylogenetic analysis, we identified 7 NCSI,

1997 17 NCSII, and 2 PR10/Bet v1 genes (Supplementary Fig. 6.8). Consistent with
1998 previous reports, a highly conserved P-loop sequence motif could be identified in the
1999 NCSI and NCSII genes, but it is absent in PR10/Bet v1 genes^{193,196}.

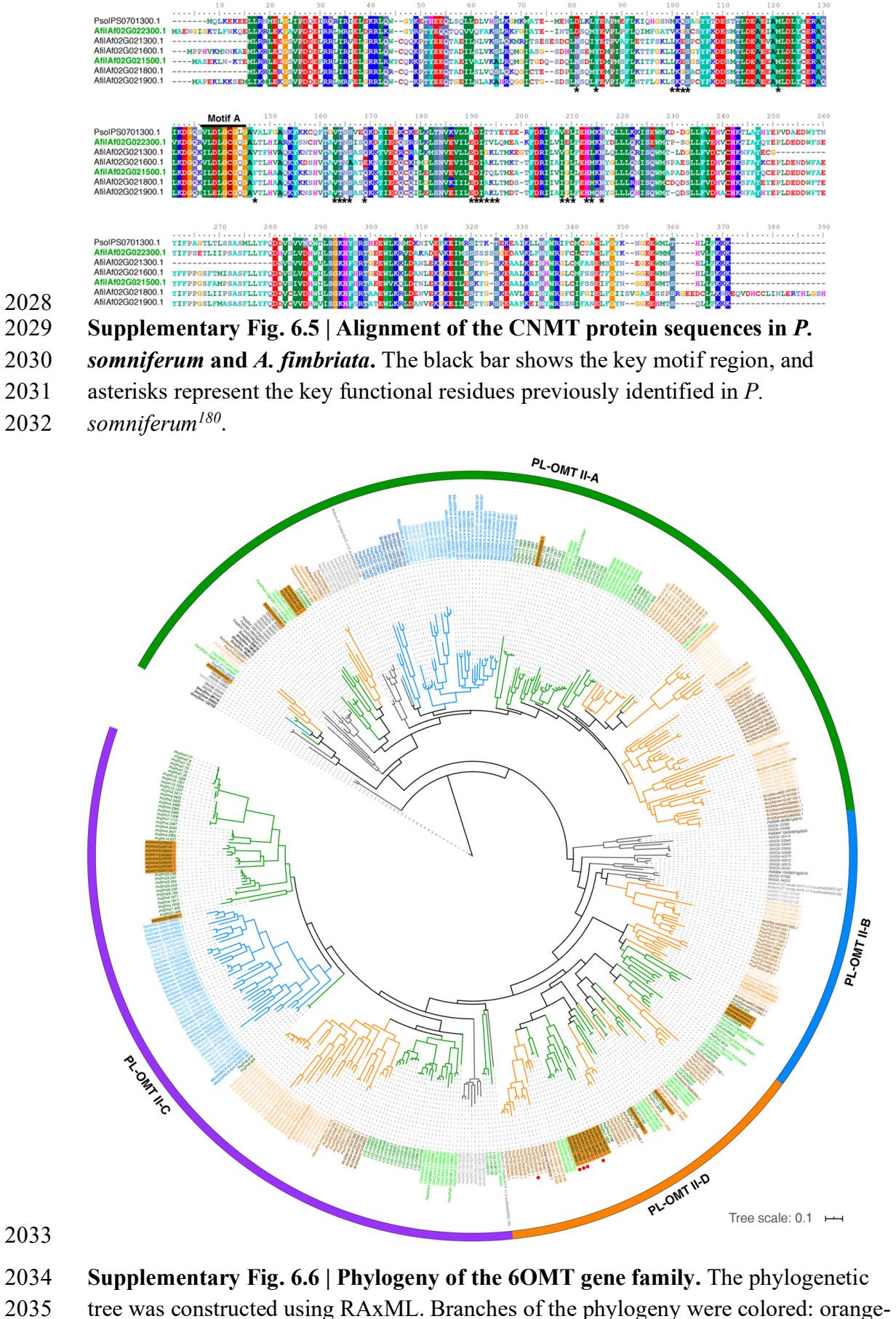
2000 NCSI genes are considered to be crucial for BIA biosynthesis. We found that the
2001 genes of *Af02G077000* and *Af02G076800* in the NCSI clade were annotated in an
2002 alkaloid-associated gene cluster (BGC #1), the *Af02G263900* and *Af02G264000* gene
2003 were annotated in a saccharide-alkaloid-associated gene cluster (BGC #10), and the
2004 *Af01G154600* and *Af05G030600* genes were located in terpene-alkaloid-associated
2005 gene clusters (BGC #24 and #25) (Supplementary Fig. 6.1 and Supplementary Table
2006 6.2). Sequence alignment of the key NCSI amino acid sequences in *A. fimbriata* and
2007 *P. somniferum* showed these three genes from *A. fimbriata* (*Af01G154600*,
2008 *Af02G077000*, and *Af02G076800*) have conserved catalytic residues with those of *P.*
2009 *somniferum*¹⁸³ (Supplementary Fig. 6.9).

2010 In addition, we performed qPCR to quantify the expression levels of the 7 NCSI
2011 genes in seedlings and the root, stem, leaf, flower, and fruit of mature plants
2012 (Extended Data Fig. 10). In general, most NCS genes have lower expression levels in
2013 seedlings than in tissues of mature plants. Two genes (*Af02G077000* and
2014 *Af01G154600*) previously identified in BGCs show high expression across root, stem,
2015 and fruit tissues in *A. fimbriata*, which seems to be consistent with the amount of their
2016 AA I production (Extended Data Fig. 10). Together, these results suggest that
2017 *Af02G077000* and *Af01G154600* are likely the main NCS candidate genes in *A.*
2018 *fimbriata*. Further integrated metabolomics, transcriptomics, and functional genetic
2019 analyses could be used to validate their roles in the aristolochic acid I biosynthesis.

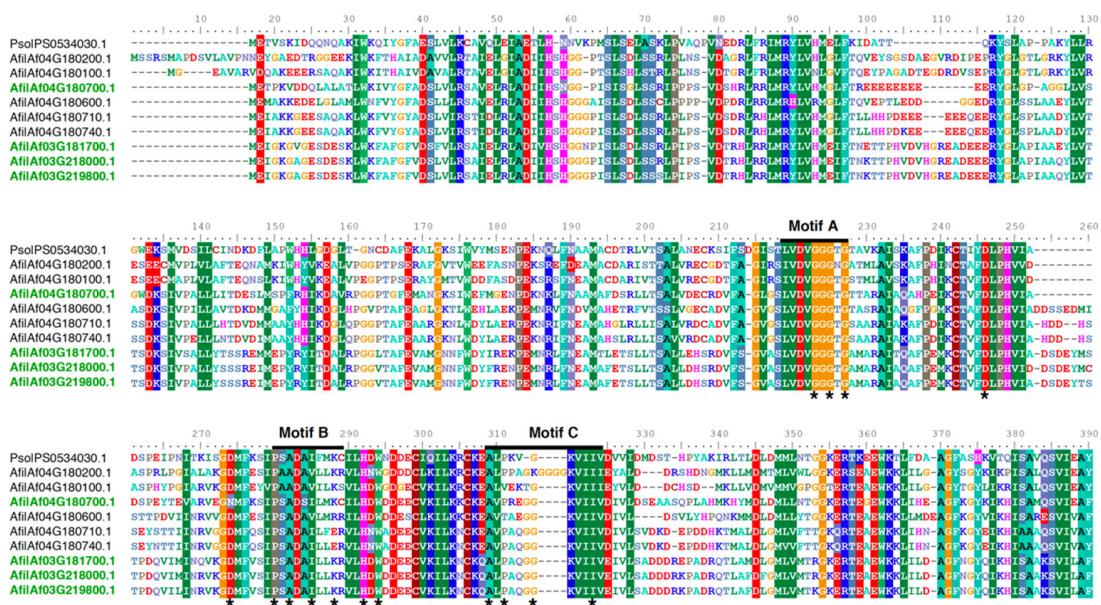


2020

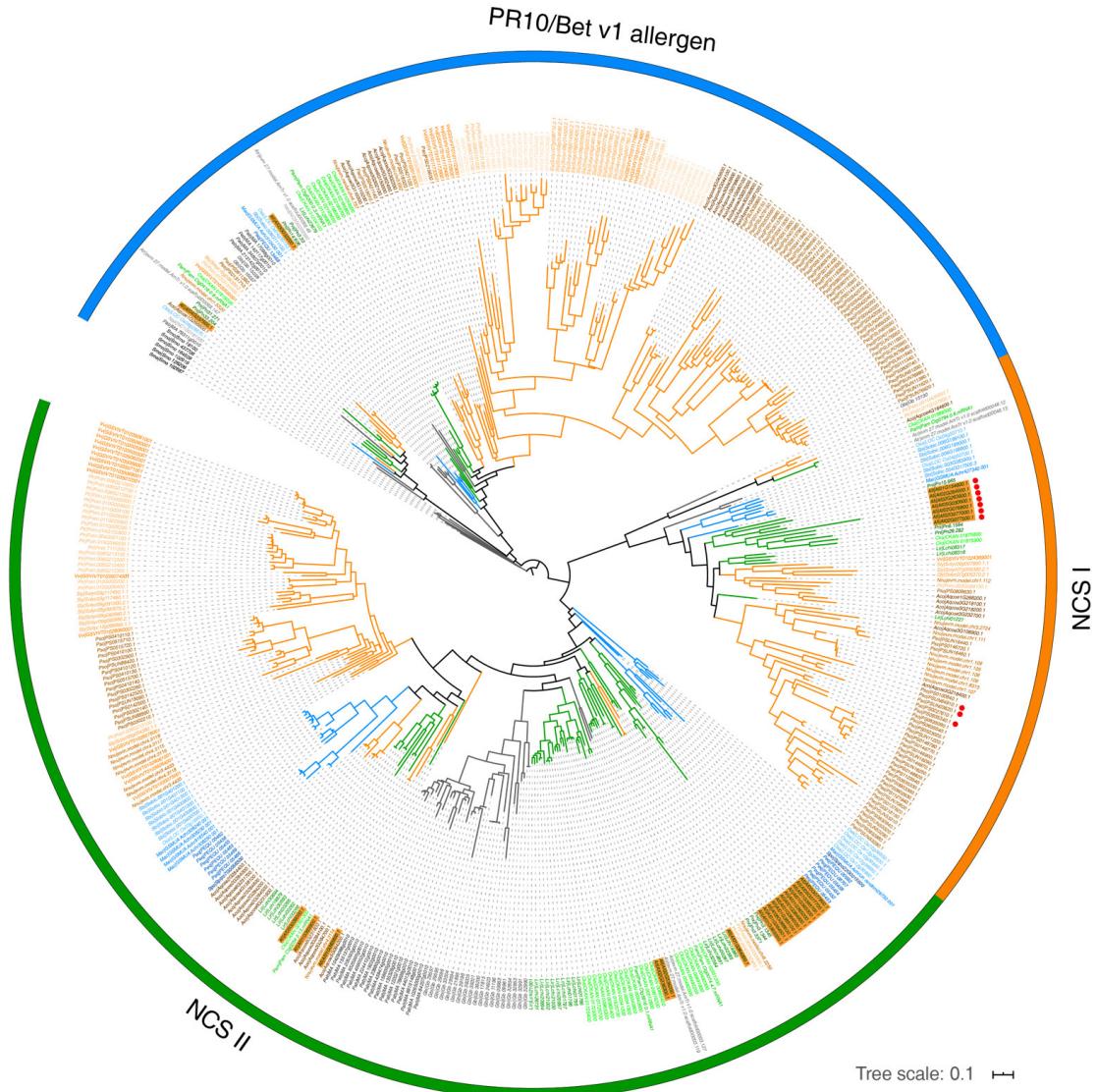
Supplementary Fig. 6.4 | Phylogeny of the CNMT gene family. The phylogenetic tree was constructed using RAxML. The *A. fimbriata* genes are highlighted in orange. The previously identified functional gene in *P. somniferum* and the two mentioned *A. fimbriata* genes in text are marked with red dots. Branches of the phylogeny were colored: orange-eudicots; blue-monocots; green-magnoliids; gray-ANA grade, gymnosperms and lycophyte. The *A. fimbriata* genes are highlighted with an orange background.



2036 eudicots; blue-monocots; green-magnoliids; gray-ANA grade, gymnosperms and
 2037 lycophyte. The previously identified functional genes in *P. somniferum* and the
 2038 closely related *A. fimbriata* genes in the PL-OMT II-D clade are marked with red
 2039 dots. The *A. fimbriata* genes are highlighted with an orange background.

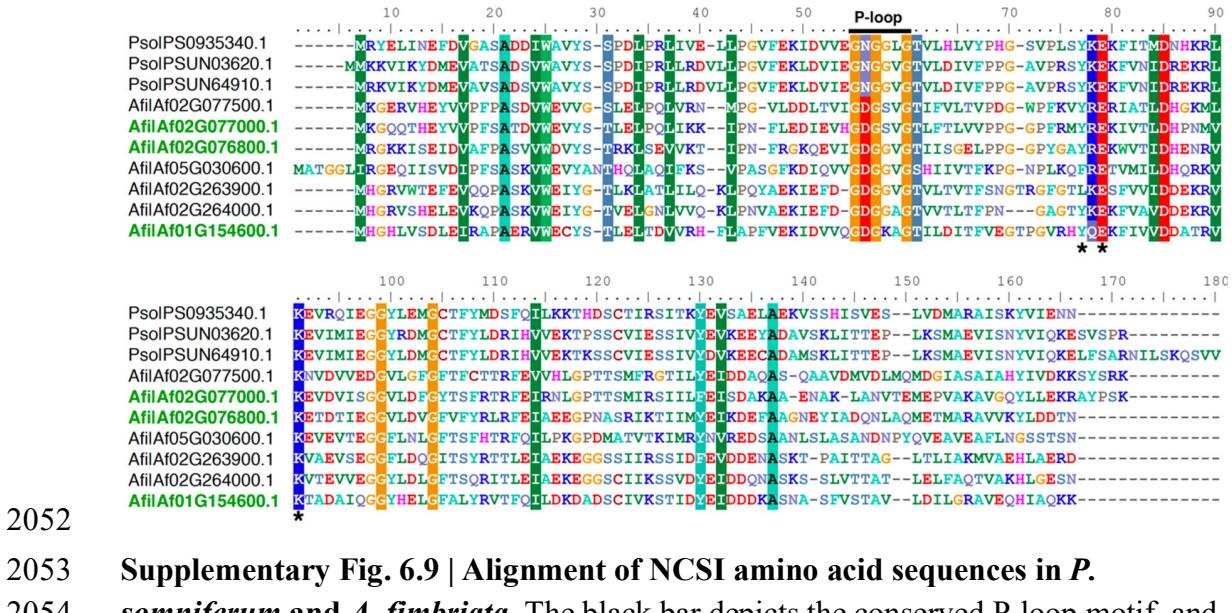


2040
 2041 **Supplementary Fig. 6.7 | Alignment of the 6OMT amino acid sequences in *P.*
 2042 *somniferum* and *A. fimbriata*.** The black bars show the regions of three key motifs,
 2043 and asterisks represent the key enzyme activation residues that were identified in *P.*
 2044 *somniferum* previously^{181,182}.



2045

2046 **Supplementary Fig. 6.8 | Phylogeny of the NCS gene family.** The phylogenetic tree
 2047 was constructed using RAxML. Branches and genes were colored as follows: orange-
 2048 eudicots; blue-monocots; green-magnoliids; gray-ANA grade, gymnosperms and
 2049 lycophyte. The *A. fimbriata* genes are highlighted with an orange background. Seven
 2050 *A. fimbriata* genes and three *P. somniferum* genes in the NCSI clade are marked with
 2051 red dots. The *A. fimbriata* genes are highlighted with an orange background.



2052

2053 **Supplementary Fig. 6.9 | Alignment of NCSI amino acid sequences in *P. somniferum* and *A. fimbriata*.** The black bar depicts the conserved P-loop motif, and
2054 asterisks represent the catalytic residues.
2055

2057 **Supplementary References:**

- 2058 1 Wagner, S. T. et al. Major trends in stem anatomy and growth forms in the
2059 perianth-bearing Piperales, with special focus on *Aristolochia*. *Ann. Bot.* **113**,
2060 1139-1154 (2014).
- 2061 2 González, F. & Stevenson, D. W. Perianth development and systematics of
2062 *Aristolochia*. *Flora* **195**, 370-391 (2000).
- 2063 3 Gonzalez, F. & Pabon-Mora, N. Trickery flowers: the extraordinary chemical
2064 mimicry of *Aristolochia* to accomplish deception to its pollinators. *New*
2065 *Phytol.* **206**, 10-13 (2015).
- 2066 4 Michl, J. et al. LC-MS- and (1)H NMR-based metabolomic analysis and in
2067 vitro toxicological assessment of 43 *Aristolochia* species. *J. Nat. Prod.* **79**, 30-
2068 37 (2016).
- 2069 5 Bliss, B. J. et al. Characterization of the basal angiosperm *Aristolochia*
2070 *fimbriata*: a potential experimental system for genetic studies. *BMC Plant*
2071 *Biol.* **13**, 13 (2013).
- 2072 6 Oelschlagel, B. et al. Spatio-temporal patterns in pollination of deceptive
2073 *Aristolochia rotunda* L. (Aristolochiaceae). *Plant Biol.* **18**, 928-937 (2016).
- 2074 7 Martin, K. R. et al. Spatial and temporal variation in volatile composition
2075 suggests olfactory division of labor within the trap flowers of *Aristolochia*
2076 *gigantea*. *Flora* **232**, 153-168 (2017).
- 2077 8 Oelschlagel, B., Gorb, S., Wanke, S. & Neinhuis, C. Structure and
2078 biomechanics of trapping flower trichomes and their role in the pollination
2079 biology of *Aristolochia* plants (Aristolochiaceae). *New Phytol.* **184**, 988-1002
2080 (2009).
- 2081 9 Heinrich, M., Chan, J., Wanke, S., Neinhuis, C. & Simmonds, M. S. Local
2082 uses of *Aristolochia* species and content of nephrotoxic aristolochic acid 1 and
2083 2--a global assessment based on bibliographic sources. *J. Ethnopharmacol.*
2084 **125**, 108-144 (2009).
- 2085 10 Nortier, J. L. et al. Urothelial carcinoma associated with the use of a Chinese
2086 herb (*Aristolochia fangchi*). *N. Engl. J. Med.* **342**, 1686-1692 (2000).
- 2087 11 Ng, A. W. T. et al. Aristolochic acids and their derivatives are widely
2088 implicated in liver cancers in Taiwan and throughout Asia. *Sci. Transl. Med.*
2089 **9**, eaan6446 (2017).
- 2090 12 Krell, D. & Stebbing, J. Aristolochia: the malignant truth. *Lancet Oncol.* **14**,
2091 25-26 (2013).
- 2092 13 Poon, S. L. et al. Genome-wide mutational signatures of aristolochic acid and
2093 its application as a screening tool. *Sci. Transl. Med.* **5**, 197ra101 (2013).
- 2094 14 Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel
2095 counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770 (2011).
- 2096 15 Vurture, G. W. et al. GenomeScope: fast reference-free genome profiling from
2097 short reads. *Bioinformatics* **33**, 2202-2204 (2017).

- 2098 16 Michael, T. P. et al. High contiguity *Arabidopsis thaliana* genome assembly
2099 with a single nanopore flow cell. *Nat. Commun.* **9**, 541 (2018).
- 2100 17 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*
2101 **34**, 3094-3100 (2018).
- 2102 18 Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy
2103 long sequences. *Bioinformatics* **32**, 2103-2110 (2016).
- 2104 19 Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo
2105 genome assembly from long uncorrected reads. *Genome Res.* **27**, 737-746
2106 (2017).
- 2107 20 Li, H. Aligning sequence reads, clone sequences and assembly contigs with
2108 BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997> (2013).
- 2109 21 Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial
2110 variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963
2111 (2014).
- 2112 22 Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-
2113 eesolution Hi-C experiments. *Cell Syst.* **3**, 95-98 (2016).
- 2114 23 Dudchenko, O. et al. The Juicebox Assembly Tools module facilitates de novo
2115 assembly of mammalian genomes with chromosome-length scaffolds for
2116 under \$1000. Preprint at <https://www.biorxiv.org/content/10.1101/254797v1>
2117 (2018).
- 2118 24 Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using
2119 Hi-C yields chromosome-length scaffolds. *Science* **356**, 92-95 (2017).
- 2120 25 Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive
2121 k-mer weighting and repeat separation. *Genome Res.* **27**, 722-736 (2017).
- 2122 26 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for
2123 Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
- 2124 27 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2.
2125 *Nat. Methods* **9**, 357-359 (2012).
- 2126 28 Li, H. et al. The sequence alignment/map format and SAMtools.
2127 *Bioinformatics* **25**, 2078-2079 (2009).
- 2128 29 Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its
2129 applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455-477 (2012).
- 2130 30 Kearse, M. et al. Geneious Basic: an integrated and extendable desktop
2131 software platform for the organization and analysis of sequence data.
2132 *Bioinformatics* **28**, 1647-1649 (2012).
- 2133 31 Hahn, C., Bachmann, L. & Chevreux, B. Reconstructing mitochondrial
2134 genomes directly from genomic next-generation sequencing reads--a baiting
2135 and iterative mapping approach. *Nucleic Acids Res.* **41**, e129 (2013).
- 2136 32 Lohse, M., Drechsel, O. & Bock, R. OrganellarGenomeDRAW (OGDRAW):
2137 a tool for the easy generation of high-quality custom graphical maps of plastid
2138 and mitochondrial genomes. *Curr. Genet.* **52**, 267-274 (2007).

- 2139 33 Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence
2140 of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
- 2141 34 Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. &
2142 Zdobnov, E. M. BUSCO: assessing genome assembly and annotation
2143 completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212
2144 (2015).
- 2145 35 Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the
2146 LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
- 2147 36 Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of
2148 repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
- 2149 37 Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive
2150 elements in genomic sequences. *Curr. Protoc. Bioinformatics* **4**, 10 (2009).
- 2151 38 Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible
2152 software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*
2153 **9**, 18 (2008).
- 2154 39 Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-
2155 length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265-W268 (2007).
- 2156 40 Ou, S. & Jiang, N. LTR_retriever: A highly accurate and sensitive program for
2157 identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**,
2158 1410-1422 (2018).
- 2159 41 Salmela, L. & Rivals, E. LoRDEC: accurate and efficient long read error
2160 correction. *Bioinformatics* **30**, 3506-3514 (2014).
- 2161 42 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large
2162 sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
- 2163 43 Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in *Drosophila*
2164 genomic DNA. *Genome Res.* **10**, 516-522 (2000).
- 2165 44 Hoff, K. J. & Stanke, M. Predicting genes in single genomes with
2166 AUGUSTUS. *Curr. Protoc. Bioinformatics* **65**, e57 (2019).
- 2167 45 Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation
2168 experiment. *Genome Res.* **10**, 547-548 (2000).
- 2169 46 Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: homology-based gene
2170 prediction utilizing intron position conservation and RNA-seq data. *Methods
2171 Mol. Biol.* **1962**, 161-177 (2019).
- 2172 47 Xu, Y., Wang, X., Yang, J., Vaynberg, J. & Qin, J. PASA--a program for
2173 automated protein NMR backbone signal assignment by pattern-filtering
2174 approach. *J. Biomol. NMR* **34**, 41-56 (2006).
- 2175 48 Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment
2176 program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875
2177 (2005).
- 2178 49 Haas, B. J. et al. Automated eukaryotic gene structure annotation using
2179 EVidenceModeler and the program to assemble spliced alignments. *Genome
2180 Biol.* **9**, R7 (2008).

- 2181 50 Conesa, A. et al. Blast2GO: a universal tool for annotation, visualization and
2182 analysis in functional genomics research. *Bioinformatics* **21**, 3674-3676
2183 (2005).
- 2184 51 Huerta-Cepas, J. et al. Fast genome-wide functional annotation through
2185 orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115-2122
2186 (2017).
- 2187 52 Camacho, C. et al. BLAST+: architecture and applications. *BMC
2188 Bioinformatics* **10**, 421 (2009).
- 2189 53 Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog
2190 groups for eukaryotic genomes. *Genome Res.* **13**, 2178-2189 (2003).
- 2191 54 Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for
2192 large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575-1584
2193 (2002).
- 2194 55 Amborella Genome Project. The *Amborella* genome and the evolution of
2195 flowering plants. *Science* **342**, 1241089 (2013).
- 2196 56 Moore, M. J., Bell, C. D., Soltis, P. S. & Soltis, D. E. Using plastid genome-
2197 scale data to resolve enigmatic relationships among basal angiosperms. *Proc.
2198 Natl. Acad. Sci. USA* **104**, 19363-19368 (2007).
- 2199 57 Soltis, D. E. et al. Angiosperm phylogeny: 17 genes, 640 taxa. *Am. J. Bot.* **98**,
2200 704-730 (2011).
- 2201 58 Jansen, R. K. et al. Analysis of 81 genes from 64 plastid genomes resolves
2202 relationships in angiosperms and identifies genome-scale evolutionary
2203 patterns. *Proc. Natl. Acad. Sci. USA* **104**, 19369-19374 (2007).
- 2204 59 Zhang, L. et al. The water lily genome and the early evolution of flowering
2205 plants. *Nature* **577**, 79-84 (2020).
- 2206 60 Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of
2207 gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
- 2208 61 Wang, J. et al. An overlooked paleotetraploidization in Cucurbitaceae. *Mol.
2209 Biol. Evol.* **35**, 16-26 (2018).
- 2210 62 Wang, J. et al. Hierarchically aligning 10 legume genomes establishes a
2211 family-level genomics platform. *Plant Physiol.* **174**, 284-300 (2017).
- 2212 63 Nei, M. & Gojobori, T. Simple methods for estimating the numbers of
2213 synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**,
2214 418-426 (1986).
- 2215 64 Jiao, Y. et al. Ancestral polyploidy in seed plants and angiosperms. *Nature*
2216 **473**, 97-100 (2011).
- 2217 65 Jiao, Y. & Paterson, A. H. Polyploidy-associated genome modifications during
2218 land plant evolution. *Philos. Trans. R. Soc. B Biol. Sci.* **369**, 20130355 (2014).
- 2219 66 Hu, L. et al. The chromosome-scale reference genome of black pepper
2220 provides insight into piperine biosynthesis. *Nat. Commun.* **10**, 4702 (2019).
- 2221 67 Samuel, R. Chromosome numbers in *Piper*. *Kew Bulletin* **42**, 465-470 (1987).

- 2222 68 Rendon-Anaya, M. et al. The avocado genome informs deep angiosperm
2223 phylogeny, highlights introgressive hybridization, and reveals pathogen-
2224 influenced gene space adaptation. *Proc. Natl. Acad. Sci. USA* **116**, 17081-
2225 17089 (2019).
- 2226 69 Chen, J. et al. *Liriodendron* genome sheds light on angiosperm phylogeny and
2227 species-pair differentiation. *Nat. Plants* **5**, 18-25 (2019).
- 2228 70 Chaw, S. M. et al. Stout camphor tree genome fills gaps in understanding of
2229 flowering plant genome evolution. *Nat. Plants* **5**, 63-73 (2019).
- 2230 71 Soltis, D. E. & Soltis, P. S. Isozyme evidence for ancient polyploidy in
2231 primitive angiosperms. *Syst. Bot.* **15**, 328-337 (1990).
- 2232 72 Cui, L. et al. Widespread genome duplications throughout the history of
2233 flowering plants. *Genome Res.* **16**, 738-749 (2006).
- 2234 73 Chen, Y. C. et al. The *Litsea* genome and the evolution of the laurel family.
2235 *Nat. Commun.* **11**, 1675 (2020).
- 2236 74 Shang, J. et al. The chromosome-level wintersweet (*Chimonanthus praecox*)
2237 genome provides insights into floral scent biosynthesis and flowering in
2238 winter. *Genome Biol.* **21**, 200 (2020).
- 2239 75 Wang, J. et al. Recursive paleohexaploidization shaped the durian genome.
2240 *Plant Physiol.* **179**, 209-219 (2019).
- 2241 76 Jiao, Y. et al. A genome triplication associated with early diversification of the
2242 core eudicots. *Genome Biol.* **13**, R3 (2012).
- 2243 77 Jiao, Y., Li, J., Tang, H. & Paterson, A. H. Integrated syntenic and
2244 phylogenomic analyses reveal an ancient genome duplication in monocots.
2245 *Plant Cell* **26**, 2792-2802 (2014).
- 2246 78 Tang, H., Bowers, J. E., Wang, X. & Paterson, A. H. Angiosperm genome
2247 comparisons reveal early polyploidy in the monocot lineage. *Proc. Natl. Acad.*
2248 *Sci. USA* **107**, 472-477 (2010).
- 2249 79 Singh, R. et al. Oil palm genome sequence reveals divergence of interfertile
2250 species in Old and New worlds. *Nature* **500**, 335-339 (2013).
- 2251 80 Guo, L. et al. The opium poppy genome and morphinan production. *Science*
2252 **362**, 343-347 (2018).
- 2253 81 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and
2254 high throughput. *Nucleic Acids Res.* **32**, 1792-1797 (2004).
- 2255 82 Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein
2256 sequence alignments into the corresponding codon alignments. *Nucleic Acids*
2257 *Res.* **34**, W609-612 (2006).
- 2258 83 Yang, Y. et al. Prickly waterlily and rigid hornwort genomes shed light on
2259 early angiosperm evolution. *Nat. Plants* **6**, 215-222 (2020).
- 2260 84 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-
2261 analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

- 2262 85 Mirarab, S. & Warnow, T. ASTRAL-II: coalescent-based species tree
2263 estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*
2264 **31**, i44-i52 (2015).
- 2265 86 Smith, S. A., Moore, M. J., Brown, J. W. & Yang, Y. Analysis of
2266 phylogenomic datasets reveals conflict, concordance, and gene duplications
2267 with examples from animals and plants. *BMC Evol. Biol.* **15**, 150 (2015).
- 2268 87 Sayyari, E. & Mirarab, S. Fast coalescent-based computation of local branch
2269 support from quartet frequencies. *Mol. Biol. Evol.* **33**, 1654-1668 (2016).
- 2270 88 Sayyari, E. & Mirarab, S. Testing for polytomies in phylogenetic species trees
2271 using quartet frequencies. *Genes* **9**, 132 (2018).
- 2272 89 One Thousand Plant Transcriptomes Initiative. One thousand plant
2273 transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679-685
2274 (2019).
- 2275 90 Yang, L. et al. Phylogenomic insights into deep phylogeny of angiosperms
2276 based on broad nuclear gene sampling. *Plant Comm.* **1** (2020).
- 2277 91 Wickett, N. J. et al. Phylotranscriptomic analysis of the origin and early
2278 diversification of land plants. *Proc. Natl. Acad. Sci. USA* **111**, E4859-4868
2279 (2014).
- 2280 92 Stull, G. W., Soltis, P. S., Soltis, D. E., Gitzendanner, M. A. & Smith, S. A.
2281 Nuclear phylogenomic analyses of asterids conflict with plastome trees and
2282 support novel relationships among major lineages. *Am. J. Bot.* **107**, 790-805
2283 (2020).
- 2284 93 Li, H. T. et al. Origin of angiosperms and the puzzle of the Jurassic gap. *Nat.*
2285 *Plants* **5**, 461-470 (2019).
- 2286 94 Gitzendanner, M. A., Soltis, P. S., Wong, G. K., Ruhfel, B. R. & Soltis, D. E.
2287 Plastid phylogenomic analysis of green plants: a billion years of evolutionary
2288 history. *Am. J. Bot.* **105**, 291-301 (2018).
- 2289 95 Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., Soltis, D. E. & Burleigh, J.
2290 G. From algae to angiosperms-inferring the phylogeny of green plants
2291 (Viridiplantae) from 360 plastid genomes. *BMC Evol. Biol.* **14**, 23 (2014).
- 2292 96 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol.*
2293 *Evol.* **24**, 1586-1591 (2007).
- 2294 97 Kenrick, P. & Crane, P. R. The origin and early evolution of plants on land.
2295 *Nature* **389**, 33-39 (1997).
- 2296 98 Miller, C. N. Implications of fossil conifers for the phylogenetic relationships
2297 of living families. *Bot. Rev.* **65**, 239-277 (1999).
- 2298 99 Doyle, J. A. & Hotton, C. L. in *Pollen and Spores, Patterns of Diversification*
2299 (eds Blackmore, S. & Barnes, S. H.) 169-195 (Clarendon Press, Oxford,
2300 1991).
- 2301 100 Doyle, J. A. & Robbins, E. I. Angiosperm pollen zonation of the continental
2302 cretaceous of the Atlantic coastal plain and its application to deep wells in the
2303 Salisbury embayment. *Palynology* **1**, 43-78 (1977).

- 2304 101 Hickey, L. J. & Doyle, J. A. Early cretaceous fossil evidence for angiosperm
2305 evolution. *Bot. Rev.* **43**, 3-104 (1977).
- 2306 102 Doyle, J. A. & Hickey, L. J. in *Origin and Early Evolution of Angiosperms*
2307 (eds Beck, C. B.) 139-206 (Columbia University Press, New York, 1976).
- 2308 103 Mohr, B. A. R. & Bernardes-de-Oliveira, M. E. C. *Endressinia brasiliiana*, a
2309 magnoliaceous angiosperm from the Lower Cretaceous Crato Formation. *Int. J.
2310 Plant Sci.* **165**, 1121-1133 (2004).
- 2311 104 De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a
2312 computational tool for the study of gene family evolution. *Bioinformatics* **22**,
2313 1269-1271 (2006).
- 2314 105 Gramzow, L. & Theissen, G. A hitchhiker's guide to the MADS world of
2315 plants. *Genome Biol.* **11**, 214 (2010).
- 2316 106 Becker, A., Winter, K. U., Meyer, B., Saedler, H. & Theissen, G. MADS-box
2317 gene diversity in seed plants 300 million years ago. *Mol. Biol. Evol.* **17**, 1425-
2318 1434 (2000).
- 2319 107 De Bodt, S., Raes, J., Van de Peer, Y. & Theissen, G. And then there were
2320 many: MADS goes genomic. *Trends Plant Sci.* **8**, 475-483 (2003).
- 2321 108 Ng, M. & Yanofsky, M. F. Function and evolution of the plant MADS-box
2322 gene family. *Nat. Rev. Genet.* **2**, 186-195 (2001).
- 2323 109 Smaczniak, C., Immink, R. G., Angenent, G. C. & Kaufmann, K.
2324 Developmental and evolutionary diversity of plant MADS-domain factors:
2325 insights from recent studies. *Development* **139**, 3081-3098 (2012).
- 2326 110 Theissen, G. & Saedler, H. Floral quartets. *Nature* **409**, 469-471 (2001).
- 2327 111 Parenicova, L. et al. Molecular and phylogenetic analyses of the complete
2328 MADS-box transcription factor family in *Arabidopsis*: new openings to the
2329 MADS world. *Plant Cell* **15**, 1538-1551 (2003).
- 2330 112 Arora, R. et al. MADS-box gene family in rice: genome-wide identification,
2331 organization and expression profiling during reproductive development and
2332 stress. *BMC Genomics* **8**, 242 (2007).
- 2333 113 Finn, R. D. et al. The Pfam protein families database: towards a more
2334 sustainable future. *Nucleic Acids Res.* **44**, D279-285 (2016).
- 2335 114 Potter, S. C. et al. HMMER web server: 2018 update. *Nucleic Acids Res.* **46**,
2336 W200-W204 (2018).
- 2337 115 Jones, P. et al. InterProScan 5: genome-scale protein function classification.
2338 *Bioinformatics* **30**, 1236-1240 (2014).
- 2339 116 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software
2340 version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**,
2341 772-780 (2013).
- 2342 117 Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for
2343 automated alignment trimming in large-scale phylogenetic analyses.
2344 *Bioinformatics* **25**, 1972-1973 (2009).

- 2345 118 Chaudhary, S., Jabre, I., Reddy, A. S. N., Staiger, D. & Syed, N. H.
2346 Perspective on alternative splicing and proteome complexity in plants. *Trends*
2347 *Plant Sci.* **24**, 496-506 (2019).
- 2348 119 Filichkin, S., Priest, H. D., Megraw, M. & Mockler, T. C. Alternative splicing
2349 in plants: directing traffic at the crossroads of adaptation and environmental
2350 stress. *Curr. Opin. Plant Biol.* **24**, 125-135 (2015).
- 2351 120 Lee, J. H. et al. Regulation of temperature-responsive flowering by MADS-
2352 box transcription factor repressors. *Science* **342**, 628-632 (2013).
- 2353 121 Lin, F., Zhang, Y. & Jiang, M. Y. Alternative splicing and differential
2354 expression of two transcripts of nicotine adenine dinucleotide phosphate
2355 oxidase B gene from *Zea mays*. *J. Integr. Plant Biol.* **51**, 287-298 (2009).
- 2356 122 Pabón-Mora, N., Suárez-Baron, H., Ambrose, B. A. & González, F. Flower
2357 development and perianth identity candidate genes in the basal angiosperm
2358 *Aristolochia fimbriata* (Piperales: Aristolochiaceae). *Front. Plant Sci.* **6**, 1095
2359 (2015).
- 2360 123 Rogers, M. F., Thomas, J., Reddy, A. S. & Ben-Hur, A. SpliceGrapher:
2361 detecting patterns of alternative splicing from RNA-Seq data in the context of
2362 gene models and EST data. *Genome Biol.* **13**, R4 (2012).
- 2363 124 Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-
2364 level expression analysis of RNA-seq experiments with HISAT, StringTie and
2365 Ballgown. *Nat. Protoc.* **11**, 1650-1667 (2016).
- 2366 125 Wellmer, F., Graciet, E. & Riechmann, J. L. Specification of floral organs in
2367 *Arabidopsis*. *J. Exp. Bot.* **65**, 1-9 (2014).
- 2368 126 Theißen, G., Melzer, R. & Rümpler, F. MADS-domain transcription factors
2369 and the floral quartet model of flower development: linking plant development
2370 and evolution. *Development* **143**, 3259-3271 (2016).
- 2371 127 Irish, V. The ABC model of floral development. *Curr. Biol.* **27**, R887-R890
2372 (2017).
- 2373 128 Specht, C. D. & Howarth, D. G. Adaptation in flower form: a comparative
2374 evodevo approach. *New Phytol.* **206**, 74-90 (2015).
- 2375 129 Aida, M., Ishida, T., Fukaki, H., Fujisawa, H. & Tasaka, M. Genes involved in
2376 organ separation in *Arabidopsis*: an analysis of the cup-shaped cotyledon
2377 mutant. *Plant Cell* **9**, 841-857 (1997).
- 2378 130 Baker, C. C., Sieber, P., Wellmer, F. & Meyerowitz, E. M. The early extra
2379 petals1 mutant uncovers a role for microRNA miR164c in regulating petal
2380 number in *Arabidopsis*. *Curr. Biol.* **15**, 303-315 (2005).
- 2381 131 Endress, P. K. The immense diversity of floral monosymmetry and asymmetry
2382 across angiosperms. *Bot. Rev.* **78**, 345-397 (2012).
- 2383 132 Horn, S., Pabón-Mora, N., Theuss, V. S., Busch, A. & Zachgo, S. Analysis of
2384 the CYC/TBI class of TCP transcription factors in basal angiosperms and
2385 magnoliids. *Plant J.* **81**, 559-571 (2015).

- 2386 133 Dilcher, D. Toward a new synthesis: major evolutionary trends in the
2387 angiosperm fossil record. *Proc. Natl. Acad. Sci. USA* **97**, 7030-7036 (2000).
- 2388 134 Martin-Trillo, M. & Cubas, P. TCP genes: a family snapshot ten years later.
Trends Plant Sci. **15**, 31-39 (2010).
- 2390 135 Pabón-Mora, N. et al. Evolution of Class II TCP genes in perianth bearing
2391 Piperales and their contribution to the bilateral calyx in *Aristolochia*. *New*
2392 *Phytol.* **228**, 752-769 (2020).
- 2393 136 Luo, D., Carpenter, R., Vincent, C., Copsey, L. & Coen, E. Origin of floral
2394 asymmetry in *Antirrhinum*. *Nature* **383**, 794-799 (1996).
- 2395 137 Luo, D. et al. Control of organ asymmetry in flowers of *Antirrhinum*. *Cell* **99**,
2396 367-376 (1999).
- 2397 138 Nath, U., Crawford, B. C., Carpenter, R. & Coen, E. Genetic control of surface
2398 curvature. *Science* **299**, 1404-1407 (2003).
- 2399 139 Hileman, L. C. Bilateral flower symmetry--how, when and why? *Curr. Opin.*
2400 *Plant Biol.* **17**, 146-152 (2014).
- 2401 140 Ishida, T., Kurata, T., Okada, K. & Wada, T. A genetic regulatory network in
2402 the development of trichomes and root hairs. *Annu. Rev. Plant Biol.* **59**, 365-
2403 386 (2008).
- 2404 141 Yang, C. & Ye, Z. Trichomes as models for studying plant cell differentiation.
2405 *Cell. Mol. Life Sci.* **70**, 1937-1948 (2013).
- 2406 142 Liu, X., Bartholomew, E., Cai, Y. & Ren, H. Trichome-related mutants
2407 provide a new perspective on multicellular trichome initiation and
2408 development in cucumber (*Cucumis sativus* L.). *Front. Plant Sci.* **7**, 1187
2409 (2016).
- 2410 143 Yang, C. et al. A regulatory gene induces trichome formation and embryo
2411 lethality in tomato. *Proc. Natl. Acad. Sci. USA* **108**, 11836-11841 (2011).
- 2412 144 Tanaka, Y. & Ohmiya, A. Seeing is believing: engineering anthocyanin and
2413 carotenoid biosynthetic pathways. *Curr. Opin. Biotechnol.* **19**, 190-197
2414 (2008).
- 2415 145 Albert, N. W. et al. A conserved network of transcriptional activators and
2416 repressors regulates anthocyanin pigmentation in eudicots. *Plant Cell* **26**, 962-
2417 980 (2014).
- 2418 146 Dixon, R. A., Liu, C. & Jun, J. H. Metabolic engineering of anthocyanins and
2419 condensed tannins in plants. *Curr. Opin. Biotechnol.* **24**, 329-335 (2013).
- 2420 147 Olsen, K. M. et al. Identification and characterisation of *CYP75A31*, a new
2421 flavonoid 3'5'-hydroxylase, isolated from *Solanum lycopersicum*. *BMC Plant*
2422 *Biol.* **10**, 21 (2010).
- 2423 148 Holton, T. A. et al. Cloning and expression of cytochrome P450 genes
2424 controlling flower colour. *Nature* **366**, 276-279 (1993).
- 2425 149 Da, L. et al. AppleMDO: a multi-dimensional omics database for apple co-
2426 expression networks and chromatin states. *Front. Plant Sci.* **10**, 1333 (2019).

- 2427 150 Maes, L., Inzé, D. & Goossens, A. Functional specialization of the
2428 *TRANSPARENT TESTA GLABRA1* network allows differential hormonal
2429 control of laminal and marginal trichome initiation in *Arabidopsis* rosette
2430 leaves. *Plant Physiol.* **148**, 1453-1464 (2008).
- 2431 151 Jaramillo, M. A. & Kramer, E. M. *APETALA3* and *PISTILLATA* homologs
2432 exhibit novel expression patterns in the unique perianth of *Aristolochia*
2433 (Aristolochiaceae). *Evol. Dev.* **6**, 449-458 (2004).
- 2434 152 Mara, C. D. & Irish, V. F. Two GATA transcription factors are downstream
2435 effectors of floral homeotic gene action in *Arabidopsis*. *Plant Physiol.* **147**,
2436 707-718 (2008).
- 2437 153 Schutte, H. R., Orban, U. & Mothes, K. Biosynthesis of aristolochic acid. *Eur.*
2438 *J. Biochem.* **1**, 70-72 (1967).
- 2439 154 Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A. & Medema, M. H.
2440 plantiSMASH: automated identification, annotation and expression analysis of
2441 plant biosynthetic gene clusters. *Nucleic Acids Res.* **45**, W55-W63 (2017).
- 2442 155 Wang, X., Hui, F., Yang, Y. & Yang, S. Deep sequencing and transcriptome
2443 analysis to identify genes related to biosynthesis of aristolochic acid in
2444 *Asarum heterotropoides*. *Sci. Rep.* **8**, 17850 (2018).
- 2445 156 Nutzmann, H. W. & Osbourn, A. Gene clustering in plant specialized
2446 metabolism. *Curr. Opin. Biotechnol.* **26**, 91-99 (2014).
- 2447 157 Nutzmann, H. W., Huang, A. & Osbourn, A. Plant metabolic clusters - from
2448 genetics to genomics. *New Phytol.* **211**, 771-789 (2016).
- 2449 158 Austin, M. B. & Noel, J. P. The chalcone synthase superfamily of type III
2450 polyketide synthases. *Nat. Prod. Rep.* **20**, 79-110 (2003).
- 2451 159 Xie, L. et al. Phylogeny and expression analyses reveal important roles for
2452 plant PKS III Family during the conquest of land by plants and angiosperm
2453 diversification. *Front. Plant Sci.* **7**, 1312 (2016).
- 2454 160 Shimizu, Y., Ogata, H. & Goto, S. Type III polyketide synthases: functional
2455 classification and phylogenomics. *ChemBioChem.* **18**, 50-65 (2017).
- 2456 161 Pandith, S. A., Ramazan, S., Khan, M. I., Reshi, Z. A. & Shah, M. A.
2457 Chalcone synthases (CHSs): the symbolic type III polyketide synthases.
Planta **251**, 15 (2019).
- 2459 162 Ferrer, J. L., Jez, J. M., Bowman, M. E., Dixon, R. A. & Noel, J. P. Structure
2460 of chalcone synthase and the molecular basis of plant polyketide biosynthesis.
2461 *Nat. Struct. Biol.* **6**, 775-784 (1999).
- 2462 163 Oelschlagel, B. et al. The betrayed thief - the extraordinary strategy of
2463 *Aristolochia rotunda* to deceive its pollinators. *New Phytol.* **206**, 342-351
2464 (2015).
- 2465 164 Chen, F., Tholl, D., Bohlmann, J. & Pichersky, E. The family of terpene
2466 synthases in plants: a mid-size family of genes for specialized metabolism that
2467 is highly diversified throughout the kingdom. *Plant J.* **66**, 212-229 (2011).

- 2468 165 Tholl, D. Terpene synthases and the regulation, diversity and biological roles
2469 of terpene metabolism. *Curr. Opin. Plant Biol.* **9**, 297-304 (2006).
- 2470 166 Johnson, S. D. & Jürgens, A. Convergent evolution of carrion and faecal scent
2471 mimicry in fly-pollinated angiosperm flowers and a stinkhorn fungus. *S. Afr. J. Bot.* **76**, 796-807 (2010).
- 2473 167 Wang, X. et al. A new sesquiterpene, a new monoterpenes and other
2474 constituents with anti-inflammatory activities from the roots of *Aristolochia debilis*. *Nat. Prod. Res.* **34**, 351-358 (2020).
- 2476 168 Jiang, S. Y., Jin, J., Sarojam, R. & Ramachandran, S. A comprehensive survey
2477 on the terpene synthase gene family provides new insight into its evolutionary
2478 patterns. *Genome Biol. Evol.* **11**, 2078-2098 (2019).
- 2479 169 Tolke, E. D. et al. Osmophores and floral fragrance in *Anacardium humile* and
2480 *Mangifera indica* (Anacardiaceae): an overlooked secretory structure in
2481 Sapindales. *AoB Plants* **10**, plv062 (2018).
- 2482 170 Dudareva, N. et al. (E)-beta-ocimene and myrcene synthase genes of floral
2483 scent biosynthesis in snapdragon: function and expression of three terpene
2484 synthase genes of a new terpene synthase subfamily. *Plant Cell* **15**, 1227-1241
2485 (2003).
- 2486 171 Comer, F., Tiwari, H. P. & Spenser, I. D. Biosynthesis of aristolochic acid.
2487 *Can. J. Chem.* **47**, 481-487 (1969).
- 2488 172 Sharma, V., Jain, S., Bhakuni, D. & Kapil, R. Biosynthesis of aristolochic acid.
2489 *J. Chem. Soc. Perkin Trans.* **1**, 1153-1155 (1982).
- 2490 173 Hagel, J. M. & Facchini, P. J. Benzylisoquinoline alkaloid metabolism: a
2491 century of discovery and a brave new world. *Plant Cell Physiol.* **54**, 647-672
2492 (2013).
- 2493 174 Ziegler, J. & Facchini, P. J. Alkaloid biosynthesis: metabolism and trafficking.
2494 *Annu. Rev. Plant Biol.* **59**, 735-769 (2008).
- 2495 175 Takemura, T., Ikezawa, N., Iwasa, K. & Sato, F. Molecular cloning and
2496 characterization of a cytochrome P450 in sanguinarine biosynthesis from
2497 *Eschscholzia californica* cells. *Phytochemistry* **91**, 100-108 (2013).
- 2498 176 Ikezawa, N., Iwasa, K. & Sato, F. Molecular cloning and characterization of
2499 CYP80G2, a cytochrome P450 that catalyzes an intramolecular C-C phenol
2500 coupling of (S)-reticuline in magnoflorine biosynthesis, from cultured *Coptis*
2501 *japonica* cells. *J. Biol. Chem.* **283**, 8810-8821 (2008).
- 2502 177 Ikezawa, N. et al. Molecular cloning and characterization of CYP719, a
2503 methylenedioxy bridge-forming enzyme that belongs to a novel P450 family,
2504 from cultured *Coptis japonica* cells. *J. Biol. Chem.* **278**, 38557-38565 (2003).
- 2505 178 Ikezawa, N., Iwasa, K. & Sato, F. Molecular cloning and characterization of
2506 methylenedioxy bridge-forming enzymes involved in stylopine biosynthesis in
2507 *Eschscholzia californica*. *FEBS J.* **274**, 1019-1035 (2007).
- 2508 179 Mizutani, M. & Ohta, D. Diversification of P450 genes during land plant
2509 evolution. *Annu. Rev. Plant Biol.* **61**, 291-315 (2010).

- 2510 180 Ounaroon, A., Decker, G., Schmidt, J., Lottspeich, F. & Kutchan, T. M. (R,S)-
2511 Reticuline 7-O-methyltransferase and (R,S)-norcoclaurine 6-O-
2512 methyltransferase of *Papaver somniferum* - cDNA cloning and
2513 characterization of methyl transfer enzymes of alkaloid biosynthesis in opium
2514 poppy. *Plant J.* **36**, 808-819 (2003).
- 2515 181 Meelaph, T. et al. Coregulation of biosynthetic genes and transcription factors
2516 for aporphine-type alkaloid production in wounded lotus provides insight into
2517 the biosynthetic pathway of nuciferine. *ACS omega* **3**, 8794-8802 (2018).
- 2518 182 Robin, A. Y., Giustini, C., Graindorge, M., Matringe, M. & Dumas, R. Crystal
2519 structure of norcoclaurine-6-O-methyltransferase, a key rate-limiting step in
2520 the synthesis of benzylisoquinoline alkaloids. *Plant J.* **87**, 641-653 (2016).
- 2521 183 Li, J., Lee, E. J., Chang, L. & Facchini, P. J. Genes encoding norcoclaurine
2522 synthase occur as tandem fusions in the Papaveraceae. *Sci. Rep.* **6**, 39256
2523 (2016).
- 2524 184 Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data
2525 using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.
2526 *Methods* **25**, 402-408 (2001).
- 2527 185 Yang, L., Zhu, J., Sun, C., Deng, Z. & Qu, X. Biosynthesis of plant
2528 tetrahydroisoquinoline alkaloids through an imine reductase route. *Chem. Sci.*
2529 **11**, 364-371 (2020).
- 2530 186 Choi, K. B., Morishige, T. & Sato, F. Purification and characterization of
2531 coclaurine N-methyltransferase from cultured *Coptis japonica* cells.
2532 *Phytochemistry* **56**, 649-655 (2001).
- 2533 187 Malakar, D., Chaudhuri, P., Dutta, T. & Ghosh, A. K. S-adenosyl-L-
2534 methionine, trehalose and oleanolic acid in few plants. *Health* **2**, 968-972
2535 (2010).
- 2536 188 Ali, R. et al. In silico identification and structure function analysis of a
2537 putative coclaurine N-methyltransferase from *Aristolochia fimbriata*. *Comput.*
2538 *Biol. Chem.* **85**, 107201 (2020).
- 2539 189 Morris, J. S. & Facchini, P. J. Molecular origins of functional diversity in
2540 benzylisoquinoline alkaloid methyltransferases. *Front. Plant Sci.* **10**, 1058
2541 (2019).
- 2542 190 Salim, V., Jones, A. D. & DellaPenna, D. *Camptotheca acuminata* 10-
2543 hydroxycamptothecin O-methyltransferase: an alkaloid biosynthetic enzyme
2544 co-opted from flavonoid metabolism. *Plant J.* **95**, 112-125 (2018).
- 2545 191 Lam, K. C., Ibrahim, R. K., Behdad, B. & Dayanandan, S. Structure, function,
2546 and evolution of plant O-methyltransferases. *Genome* **50**, 1001-1013 (2007).
- 2547 192 Lee, E. J. & Facchini, P. Norcoclaurine synthase is a member of the
2548 pathogenesis-related 10/Bet v1 protein family. *Plant Cell* **22**, 3489-3503
2549 (2010).

- 2550 193 Liscombe, D. K., Macleod, B. P., Loukanina, N., Nandi, O. I. & Facchini, P. J.
2551 Evidence for the monophyletic evolution of benzylisoquinoline alkaloid
2552 biosynthesis in angiosperms. *Phytochemistry* **66**, 1374-1393 (2005).
- 2553 194 Samanani, N. & Facchini, P. J. Isolation and partial characterization of
2554 norcoclaurine synthase, the first committed step in benzylisoquinoline alkaloid
2555 biosynthesis, from opium poppy. *Planta* **213**, 898-906 (2001).
- 2556 195 Samanani, N. & Facchini, P. J. Purification and characterization of
2557 norcoclaurine synthase: the first committed enzyme in benzylisoquinoline
2558 alkaloid biosynthesis in plants. *J. Biol. Chem.* **277**, 33878-33883 (2002).
- 2559 196 Vimolmangkang, S. et al. Evolutionary origin of the *NCSI* gene subfamily
2560 encoding norcoclaurine synthase is associated with the biosynthesis of
2561 benzylisoquinoline alkaloids in plants. *Sci. Rep.* **6**, 26323 (2016).