

<https://doi.org/10.1038/s42003-025-07683-y>

Structure and evolution of the Forsythieae genome elucidated by chromosome-level genome comparison of *Abeliophyllum distichum* and *Forsythia ovata* (Oleaceae)



Hoyeol Jang¹, Haneul Kim¹, Ara Cho^{1,2}, Hee-Ju Yu³, Sun Mi Huh³, Hyuk-Jin Kim², Dong-Kab Kim², Joonhyung Jung⁴, Joo-Hwan Kim⁴ & Jeong-Hwan Mun¹ 

Abeliophyllum distichum and *Forsythia ovata* are two closely related ornamental species of the tribe Forsythieae (Oleaceae) native to Korea. Here we report their genomic characteristics, highlighting genetic differences contributing to variations in corolla coloration, genomic variations associated with heterostyly, and the reconstruction of their ancestral karyotypes. Genome comparison revealed that *A. distichum* had a more compact organization of gene space than *F. ovata*. Centromeres of both species were enriched in Forsythieae-specific satellite repeats, hAT-Ac and MuLE-MuDR DNA transposons, and OTA-Athila Ty3/Gypsy retrotransposons. Transcriptome analysis revealed spatially differential expression of carotenoid biosynthesis-related genes in *A. distichum*, with downregulation in the white lobe and upregulation in the yellow base. Genome-wide analysis of structural variation in *A. distichum* identified retrotransposon insertions in the promoter region of an *AGAMOUS* homolog in the thrum plant, which showed significant downregulation of the gene compared to the pin plant. Evolutionary analyses suggested that the Oleaceae genomes evolved from 13 ancestral karyotypes via lineage-specific genomic events, including chromosome recombination, rearrangement, and whole-genome duplication followed by diploidization. The divergence of *A. distichum* and *F. ovata* was estimated to have occurred 13.87 million years ago during the Miocene epoch.

Decoding the genomic basis of plant taxa with unique characteristics has been a focus of plant science and evolutionary studies. Oleaceae is a medium-sized family consisting of approximately 700 species in 28 genera. Oleaceae species are found as shrubs, trees or a few lianas in tropical and temperate regions of the world, except Antarctica¹. Many species of the Oleaceae have been widely cultivated for their economic value. Well-known species of this family are the olive trees of the genus *Olea*, which are grown for their fruit and oil, while many species of the genera *Forsythia*, *Ligustrum*, *Jasminum*, *Syringa*, and *Fraxinus* are ornamental plants or sources of fragrance and wood. Most classifications of the Oleaceae divide the family into two subfamilies: Oleoideae and Jasminoideae. Members of the subfamily Oleoideae form a monophyletic group within the tribe Oleae, with the haploid chromosome numbers (x) of 23 and rarely $x = 24$. Cytological studies suggest that the tribe Oleae has a polyploid origin and the basic

chromosome number of the family is $x = 11^{2,3}$. The subfamily Jasminoideae is heterogeneous and paraphyletic with $x = 11$, 13, and 14 and is divided into four tribes: Myxopyreae, Fontanesieae, Jasmineae, and Forsythieae^{2,4}. The tribe Forsythieae includes two genera: *Abeliophyllum* and *Forsythia*. Forsythieae species grow naturally in the mountainous regions of Eurasia, with the greatest diversity in East Asia, and are economically important as ornamentals and traditional herbal medicines. In particular, some Forsythieae species, such as *A. distichum*, *F. koreana*, *F. ovata*, and *F. nakaii*, are endemic to Korea⁵. *A. distichum* is the only species of the monotypic genus *Abeliophyllum* and is endangered in the wild. The genus *Forsythia* consists of 11 to 13 species, most of which are native to East Asia, with only *F. europaea* being native to Albania in Southeast Europe. Distinct phenotypic differences between *Abeliophyllum* and *Forsythia* species include flower color and fruit morphology. *A. distichum* has overall white corollas (white

¹Department of Bioscience and Bioinformatics, Myongji University, Yongin, 17058, Korea. ²Division of Forest Biodiversity, Korea National Arboretum, Pocheon, 11186, Korea. ³Department of Medical and Biological Sciences, The Catholic University of Korea, Bucheon, 14662, Korea. ⁴Department of Life Science, Gachon University, Seongnam, 13120, Korea. ✉e-mail: munjh@mju.ac.kr

lobe and pale yellow base) and winged samara-type fruits, whereas *Forsythia* species are characterized by yellow corollas and capsule fruits. The tribe Forsythieae is considered monophyletic⁶ based on chromosome cytology, which shows the same number of chromosomes ($x = 14$) and consistent karyotype patterns across the species². In addition, Forsythieae species are characterized by dimorphic flowers (heterostyly), which are also recognized in the genera *Jasminum*, *Phillyrea*, and *Forestiera* of the Oleaceae⁷. The morphological and functional divergences between plants with different flower morphs (pin and thrum types) in Forsythieae species remain little known. The reticulate surface of pollen morphology⁸ and molecular phylogenies based on nuclear and chloroplast genes^{6,9} suggest that *Forsythia* may have been derived from *Abeliophyllum* or its ancestor. Considering that Forsythieae species have relatively small but compact genomes (approximately 1 Gb or less), they may retain the ancestral chromosome structure of the Oleaceae in their 14 chromosomes¹⁰.

Recent advances in genome sequencing technologies and bioinformatics tools are shedding light on the evolutionary history and diversification patterns of the Oleaceae genomes. Over the past decade, chromosome-level genome assemblies have been reported for five species of the tribe Oleaeae, including *Olea europaea*^{11–13}, *Fraxinus excelsior*¹⁴, *Osmanthus fragrans*¹⁵, *Fraxinus pennsylvanica*¹⁶, and *Syringa oblata*^{13,17,18}. In addition, assemblies have been reported for *Jasminum sambac*^{19,20} of the tribe Jasmineae, and *Forsythia suspensa*^{21,22} of the tribe Forsythieae. Genome analyses of these species revealed that the genomes of the tribe Oleaeae contain homologous chromosomal segments throughout each genome, suggesting a duplication of the entire genome, known as a whole genome duplication (WGD). Molecular dating of orthologs based on sequence divergence estimated that the WGD may have occurred approximately 28 million years ago (MYA)^{11,13}. In contrast, few duplicated chromosomes have been identified in the genomes of *J. sambac* ($x = 13$) and *F. suspensa* ($x = 14$), suggesting that the genomes of the tribes Jasmineae and Forsythieae have not undergone recent episodes of genome doubling. Therefore, WGD is thought to be shared only by the tribe Oleaeae, contributing to an increase in the number of chromosome pairs ($x = 23$) in these species compared to those of the tribes in the subfamily Jasminoideae. However, it remains unclear whether the WGD resulted from autopolyploidy or allopolyploidy, and what the basal chromosome number of the Oleaceae family is.

In plant genomes, repetitive sequences can be interspersed throughout the genome or exist as locally tandem repeats. Interspersed repeats mainly consist of transposable elements (TEs), including DNA transposons and retrotransposons (RTs), whereas tandem repeats are mainly localized at or near centromeres and telomeres. Insertion of repeated sequences into the genome can generate structural variations (SVs), which contribute to genetic and epigenetic diversity in phenotypes or traits. A number of plant species have shown phenotypic variation due to SVs, such as fruit shape in tomato²³ and peach²⁴, fruit color in apple²⁵ and peach²⁶, fruit texture in apple²⁷, leaf shape in rapeseed²⁸, resistance to biotic stress in maize²⁹, sorghum³⁰ and rice³¹, and abiotic stress in *Arabidopsis*³². Satellite repeats (SRs) are long tandem repeat sequences with a repeat unit ranging from a few base pairs (bp) to thousands of bp in length. They are mainly found in the centromeric and pericentromeric heterochromatin of chromosomes. In plant genomes with monocentric tandem repeat architecture, centromeric satellite arrays are typically localized to a single chromosomal region, with SR unit lengths ranging from 44 bp to 5 kb in highly ordered repeats that are typically at least kilobases long. Examples include *A. thaliana* (178-bp *CEN178*), *A. lyrata* (168-bp *CEN168* and 179-bp *CEN179*), *Brassica rapa* (176-bp *CentBr*), *Erianthus rufipilus* (137-bp *CEN137*), *Glycine max* (92-bp *GmCent*), *Oryza* species (155-bp *CentO*), *Solanum tuberosum* (979-bp *St24* and 5.4-kb *St3-294*), tribe Fabaeae species (from 33 bp to 2,979 bp satellites), *Vigna unguiculata* (455-bp *CEN455*, 721-bp *CEN721*, and 1600-bp *CEN1600*), and *Vitis vinifera* (107-bp satellites)³³. Additionally, the 177-bp *CentRs* in radish has been identified in radish centromeres^{34,35}. In the Oleaceae species, four *CentFs* satellites (356-bp *CentFs356*, 365-bp *CentFs365*, 530-bp *CentFs530*, and 732-bp *CentFs732*) have been reported from centromeres of *F. suspensa*; however, their unit sequences are not

available²². Some species, such as einkorn wheat, do not have high-copy centromere-specific SRs³⁶. Satellite DNAs exhibit a high degree of polymorphism due to random mutations and rearrangements between satellite arrays. The hypervariable nature of SRs among related organisms makes them excellent markers for characterizing genomes and conducting diversity-related studies, as well as for mapping and marker-assisted breeding in crop species³⁷. Our understanding of centromere structure and centromeric tandem repeat arrays in Oleaceae genomes has been limited. However, as more near-complete genomes become available, it will be feasible to analyze the diversity of SRs in these species.

Comparison of genomes with contiguous assemblies provides a detailed picture of chromosome-level synteny and large-scale rearrangements that differentiate genomes. Moreover, syntenic relationships between species have been instrumental in unraveling genome evolution and reconstructing ancestral genomes. In a previous study, we reported the chromosome-level assemblies and annotations of the *A. distichum* and *F. ovata* genomes (NCBI accession numbers JBFOLK000000000 and JBFOLJ000000000, respectively) using a combination of PacBio sequencing and Hi-C chromosome conformation capture for scaffolding¹⁰. The maximum haploid genome sizes of these species were estimated to be 844.58 Mb for *A. distichum* and 937.12 Mb for *F. ovata* based on k-mer analysis of homozygous reads. The assembled genomes of *A. distichum* and *F. ovata* were 795.7 and 1108.5 Mb in size, with scaffold N50 lengths of 53.1 Mb and 69.0 Mb, BUSCO completeness (Embryophyta) of 97.7% and 98.8%, and 49,414 and 57,587 protein-coding genes, respectively. The 14 pseudo-chromosomes of *A. distichum* and *F. ovata* covered 93.8% and 89.1% of the genome assemblies, respectively, demonstrating sufficient contiguity for chromosome-to-chromosome comparisons. This work represents the structural characterization of the *A. distichum* and *F. ovata* pseudo-chromosome sequences. We identified the repetitive sequences, including DNA TEs, RTs, and SRs in the chromosomes, differentially expressed genes related to carotenoid biosynthesis, and SVs between pin and thrum plants of *A. distichum*. In addition, a comprehensive comparative genome analysis with the sequenced Oleaceae genomes allowed the establishment of syntenic genome blocks and the reconstruction of ancestral karyotypes (AKs), providing novel insights into the evolutionary history of the tribe Forsythieae.

Results

General features of the *A. distichum* and *F. ovata* chromosome assemblies

The 14 longest pseudo-chromosomes were selected from our previous genome assemblies of *A. distichum* and *F. ovata* for detailed analysis. The chromosome assemblies measured 746.4 Mb (3235 contigs and 323.4 kb gaps) for *A. distichum* and 987.8 Mb (1331 contigs and 149.0 kb gaps) for *F. ovata*, respectively (Table 1). The *A. distichum* chromosomes contained 3.8% fewer repetitive sequences, 2491 fewer protein-coding genes, and 1436 fewer RNA genes compared to *F. ovata*. The average gene size in *A. distichum* was also 1932 bp smaller. Comparison of the chromosome assemblies showed that *A. distichum* has a gene density of 16.3 kb per gene, approximately 26% and 37% higher than *F. ovata* (20.5 kb per gene) and *F. suspensa* (22.3 kb per gene), respectively, indicating the most compact gene space organization among the sequenced species of the tribe Forsythieae.

Repetitive sequence analysis revealed that approximately 44–45% and 12–14% of the chromosome assemblies consisted of LTRs and DNA TEs, respectively, with short interspersed nuclear elements (SINEs) being very rare (0.0–0.1%) (Table 2). The dominant DNA TEs and LTRs varied in length between species: in *A. distichum*, MuLE-MuDR and Ty1/Copia were dominant, similar to *F. suspensa*, whereas in *F. ovata*, hAT-Ac and Ty3/Gypsy were dominant. A total of 11,129 full-length long terminal repeat retrotransposons (FL-LTR-RTs) with an average length of 8.5 kb were identified in the chromosomes of *A. distichum* and *F. ovata* (Table S1). OTA-Tat was the predominant Ty3/Gypsy FL-LTR-RT in both genomes, while Tork and SIRE families were the principal Ty1/Copia FL-LTR-RTs in *A. distichum* and *F. ovata*, respectively. A proteome comparison between the two genomes identified 18,026 orthologous groups, comprising 27,775

genes in *A. distichum* and 31,198 genes in *F. ovata*. The number of hypothetical or uncharacterized unique genes was 2629 (6%) in *A. distichum* and 4356 (9%) in *F. ovata*. The distribution of genes and repetitive sequences, together with expression and methylation levels per 100-kb bin along the chromosomes, is shown in Fig. 1.

Syntenic analysis showed a one-to-one chromosome collinear relationship between homologous chromosomes, albeit with several mismatched regions due to widely distributed repetitive sequences. The mismatched regions were aligned with centromeres of recently published *F. suspensa* chromosome assemblies²³ (Fig. 2). Comparison with *J. sambac* also revealed long-range mismatches in these regions (Fig. S1). In-depth sequence analysis of homologous chromosomes identified heterochromatic regions characterized by clusters of repetitive sequences, DNA TEs, and FL-

LTR-RTs. These regions (average length of 3.6 Mb) had 1.3- and 1.2-fold higher repeat coverage, 1.1- and 1.3-fold higher CG and CHG methylation levels, and 1.6- and 1.8-fold lower gene density compared to other chromosome regions in *A. distichum* and *F. ovata*, respectively (Table S2). Consequently, these regions were identified as putative centromere regions.

Satellite repeats and centromere structure

Long tandem repeat sequences dispersed throughout the chromosome assemblies were identified using repeat-finding algorithms and manual curation. Satellite Repeat Finder detected multiple arrays of SRs in *A. distichum*, *F. ovata*, and other sequenced Oleaceae genomes. After filtering SR arrays shorter than 1 kb, those with >75 bp unit length and more than 20 copies were named Oleaceae Satellite Repeats (*OISRs*) and classified into 105 groups based on their repeat unit sequences (Supplementary Data 1). The repeat unit sizes of the *OISRs* ranged from 200 bp (*OISR103*) to 4023 bp (*OISR68*), with AT content varying from 34.8% (*OISR45*) to 85.1% (*OISR87*). None of the *OISRs* were conserved across all seven sequenced Oleaceae genomes; instead, 74 out of 105 *OISRs* (70.5%) were species-specific, indicating unique amplification of these sequences in each Oleaceae genome. The most abundant SR was *OISR102*, with 21,512 copies specifically identified at 52 loci in the *O. europaea* genome. In contrast, *J. sambac* contains only 187 copies of *OISR36*, which is uniquely present at three loci in its genome.

A total of 35 SR groups were identified in the tribe Forsythieae, including 7 species-specific SRs (*OISR1–OISR7*), 7 SRs shared between two species (*OISR8–OISR14*), and 21 SRs commonly shared among three species (*OISR15–OISR35*). *A. distichum* has 1.8- and 2.4-fold fewer SRs in its genome compared to *F. suspensa* and *F. ovata*, respectively. Interestingly, the AT-rich *OISR30–35* (with AT content ranging from 70.6% to 73.7%) were the six dominant *OISRs* common to all three Forsythieae species. These repeats were highly similar (56.9%–84.2% sequence identity) and typically occurred together in a linear arrangement, occasionally overlapping. Among them, *OISR35* was the most abundant SR in the tribe Forsythieae genomes, present as 4536 copies in 67 loci in *A. distichum*, 9462 copies in 146 loci in *F. ovata*, and 6990 copies in 98 loci in *F. suspensa*. Although most *OISRs* in *A. distichum* and *F. ovata* were distributed along the chromosome arms, 35.6% and 22.3% of them were located in the putative centromere regions, respectively (Supplementary Data 1). In contrast to other plant species, few centromere-specific tandem arrays of SRs were found in the Forsythieae species, as well as in other Oleaceae species.

Table 1 | Chromosome assembly statistics for *A. distichum*, *F. ovata*, and *F. suspensa*

Chromosome	<i>A. distichum</i>	<i>F. ovata</i>	<i>F. suspensa</i> ^a
Number	14	14	14
Total length (bp)	746,393,055	987,838,149	712,936,073
Gaps (bp)	323,400	149,000	80,500
GC (%)	33.9	33.8	33.7
Repetitive sequences (bp)	358,459,201	512,990,529	319,603,139
RNA genes	3814	5250	4915
Protein-coding genes	45,792	48,283	31,919
Gene density (kb/gene)	16.3	20.5	22.3
Average size (bp)	3168.7	5100.7	-
Number of exon per gene	4.2	4.6	-
Average exon size (bp)	253.0	230.6	-
Average intron size (bp)	662.8	1126.2	-

^aStatistics for *F. suspensa* chromosome assembly were calculated using the report by Li et al. (2023), excluding protein-coding genes due to lack of annotation information.

Table 2 | Repetitive sequences in the chromosomes of *A. distichum*, *F. ovata*, and *F. suspensa*

Species Class	<i>A. distichum</i>			<i>F. ovata</i>			<i>F. suspensa</i>		
	Number	Length (bp)	Chromosome ratio (%)	Number	Length (bp)	Chromosome ratio (%)	Number	Length (bp)	Chromosome ratio (%)
DNA TE	238,427	88,888,358	11.9	366,402	134,213,134	13.6	268,831	82,330,870	11.6
hAT-Ac	42,574	18,488,782	2.5	71,514	29,782,695	3.0	62,248	18,534,947	2.6
MuLE-MuDR	42,615	19,055,635	2.6	63,505	28,577,816	2.9	78,816	23,345,156	3.3
hAT-Tip100	43,884	17,431,712	2.3	37,824	12,964,247	1.3	22,527	8,368,879	1.2
CMC-EnSpm	21,092	9,906,238	1.3	42,169	20,983,061	2.1	36,003	15,559,372	2.2
PIF-Harbinger	26,683	9,341,352	1.3	25,509	7,842,627	0.8	20,012	6,100,977	0.9
Others	61,579	14,664,639	2.0	125,881	34,062,688	3.4	49,225	10,421,539	1.5
LINE	19,227	9,681,181	1.3	73,057	25,800,450	2.6	31,952	10,850,297	1.5
Helitron	32,312	7,585,879	1.0	25,109	14,388,219	1.5	15,695	7,480,815	1.1
SINE	24	2103	0.0	3741	724,451	0.1	1877	220,456	0.0
LTR	306,921	330,895,650	44.3	429,979	440,130,777	44.6	380,686	306,743,499	43.0
Ty1/Copia	137,873	162,483,101	21.8	195,626	205,880,512	20.8	165,435	147,306,245	20.7
Ty3/Gypsy	156,306	156,763,725	21.0	215,475	218,149,659	22.1	184,339	144,834,978	20.3
Others	12,742	11,648,824	1.6	18,878	16,100,606	1.6	30,912	14,602,276	2.0

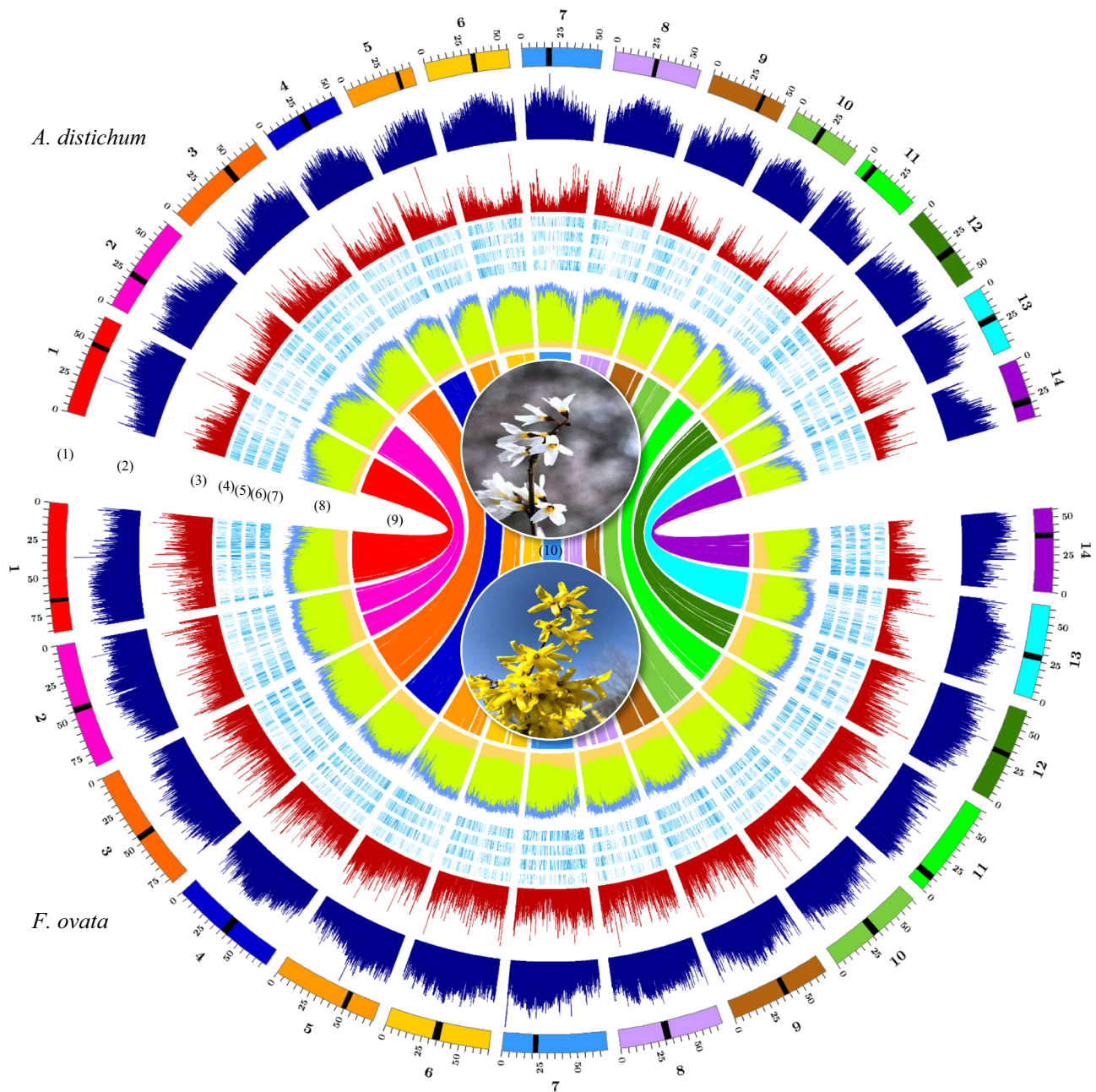


Fig. 1 | Genomic features of the *Abeliophyllum distichum* and *Forsythia ovata* chromosome-level assemblies. From the outside circle to the center: (1) Chromosomes are depicted in 14 different colors, with putative centromere positions indicated by black bands. Size bars represent 25 Mb intervals with 5 Mb steps. (2) Nucleotides of repetitive sequences per 100-kb bin. (3) Density of protein-coding genes per 100-kb bin. Gene expression levels in (4) leaf, (5) corolla, (6) carpel, and (7)

stamen tissues, measured as average TMM values (range: 0–100) per 100-kb bin. (8) Whole-genome bisulfite methylation profile in leaf tissue, showing nucleotides per 100-kb bin, with CG sites in light blue, CHG sites in light green, and CHH sites in yellow. (9) Connection of homologous chromosome segments between the genomes. (10) Photographs of *A. distichum* (top) and *F. ovata* (bottom) plants.

The co-localization of DNA TEs, RTs, FL-LTR-RTs, SRs, and genes characterized the putative centromere regions of *A. distichum* and *F. ovata* chromosomes (Fig. 3a). In these regions, DNA TEs were more widely inserted than FL-LTR-RTs and other RTs, showing a 2.0- and 1.2-fold higher localization in centromeres, respectively (Table 3 and Fig. S2). Approximately 4% of DNA TEs were localized in centromeres of both species, with hAT-Ac, MuLE-MuDR, and CMC-EnSpm being abundantly distributed. In contrast, relatively few insertions of FL-LTR-RTs were found in centromere regions. The dominant FL-LTR-RTs in these regions were OTA-Athila of Ty3/Gypsy in both species, which is the second most abundant FL-LTR-RT in the entire genome. Notably, no

SINEs were inserted into the centromere regions of *A. distichum*. A total of 18 and 22 groups of OISRs were distributed in the centromere regions of 12 *A. distichum* and 13 *F. ovata* chromosomes, respectively (Table S3). However, *A. distichum* chromosomes 4 and 14 and *F. ovata* chromosome 9 lacked OISRs in their centromeres. OISR30–35 were the six dominant OISRs localized in the centromere regions of both species, with an average of 2637 copies in *A. distichum* and 3311 copies in *F. ovata*. These SRs were particularly tandemly clustered in the centromere regions of *A. distichum* chromosome 11 and *F. ovata* chromosomes 3, 6, 8, and 11, where the nested insertion of OISRs formed clusters ranging in size from 22.9 kb to 654.3 kb (Figs. 3b, S3, and S4). In contrast, OISR1–3

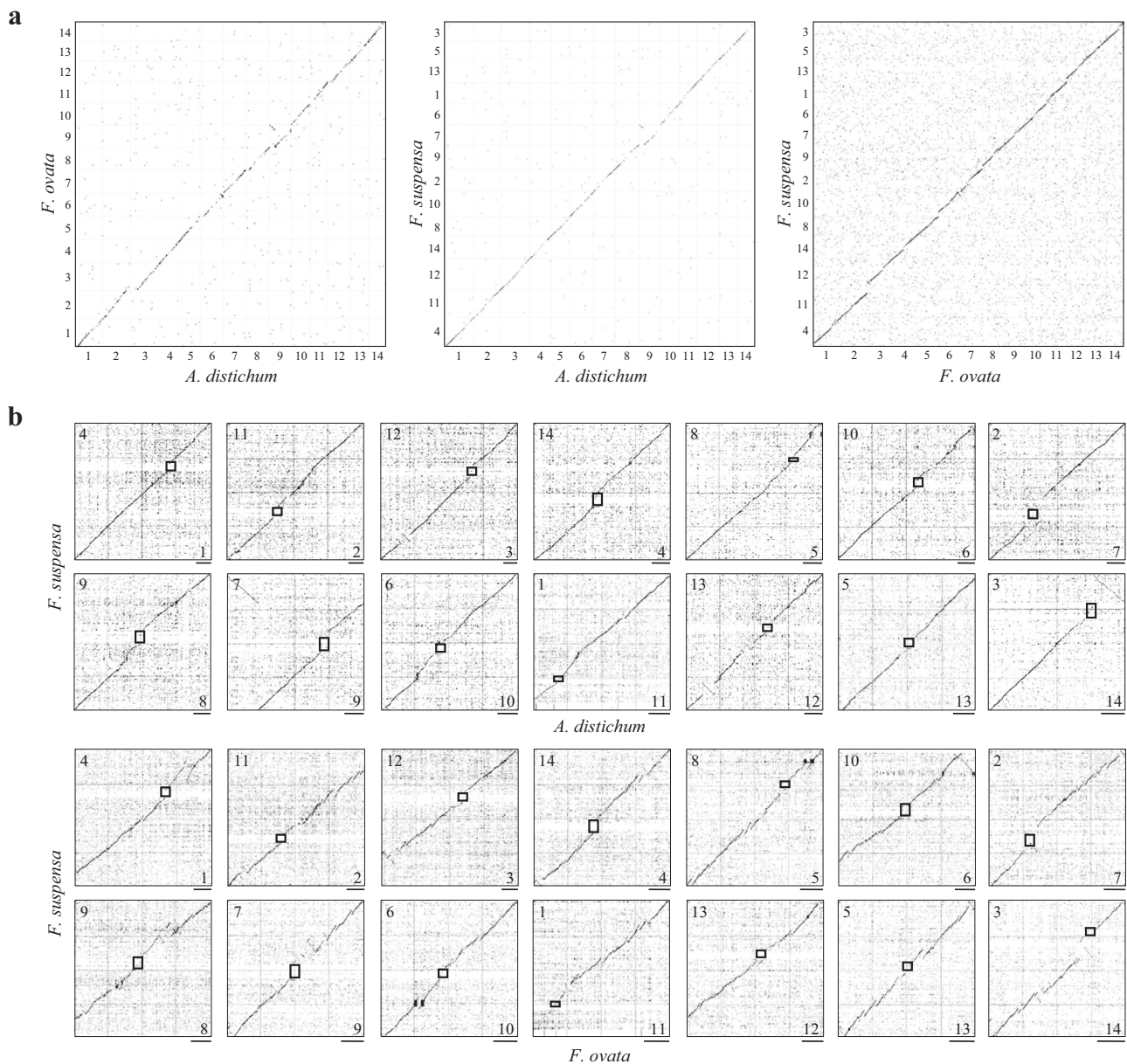


Fig. 2 | Dot-plot comparison of the *A. distichum*, *F. ovata*, and *F. suspensa* chromosome assemblies. a Dot-plots of the assembly-to-assembly comparison. Dots closest to the diagonal line reflect collinearity of genes between the two assemblies. **b** Pairwise alignment of homologous chromosomes, highlighting

putative centromere positions on the mismatched heterochromatic regions. Centromere regions of *F. suspensa* chromosomes are shown in black boxes. Bars represent 5 Mb.

(chromosome 11) and *OISR10* (chromosome 10) in *A. distichum* and *OISR13* (chromosome 4) in *F. ovata* were inserted exclusively into the centromere region of a single chromosome. The centromeres of *F. suspensa* showed a similar clustered distribution of *OISR30–35* with an average of 2720 copies. (Table S3). Moreover, *OISR30–35* clusters were localized to the *CentFs* regions of *F. suspensa* chromosome 2 and 9 centromeres (Fig. S5).

Expression of genes related to carotenoid biosynthesis in the corolla

The bell-shaped corolla of the tribe Forsythieae is deeply four-lobed, with the petals joined at the base. The yellow floral pigmentation of *Forsythia* species is mainly due to the accumulation of carotenoids. In contrast, the corolla of *A. distichum* is mostly white with four lobes and a pale yellow base, indicating a loss of carotenoid pigmentation in the lobes (Fig. 4a). The deposition of carotenoids in plastids is associated with the expression of

genes within their biosynthesis pathways (Fig. 4b). We identified genes in the carotenoid biosynthesis pathway and their regulatory genes in the MBW complex (composed of R2R3-MYB, bHLH, and WD40) in the genomes of *A. distichum* and *F. ovata*. A total of 21 and 29 carotenoid biosynthesis-related genes and 10 and 8 MBW complex genes were identified in the genomes of *A. distichum* and *F. ovata*, respectively. To compare the expression of these genes, we performed RNA-seq analyses in four tissues: leaf, corolla, carpel, and stamen. Transcriptome data indicated that most of the carotenoid biosynthesis-related genes were differentially expressed in the four tissues (Fig. 4c). In the corolla, the expression of genes encoding lycopene beta cyclase (LCY-B) and zeaxanthin epoxidase (ZEP) was downregulated, whereas those encoding beta-carotene isomerase (DWARF27), capsanthin/capsorubin synthase (CCS), and 9-*cis*-epoxycarotenoid dioxygenase (NCED) were upregulated in *A. distichum* compared to *F. ovata*. In *F. ovata* corolla, the genes encoding LCY-B, LCY-E, and ZEP showed higher expression than in *A. distichum*, while genes related to

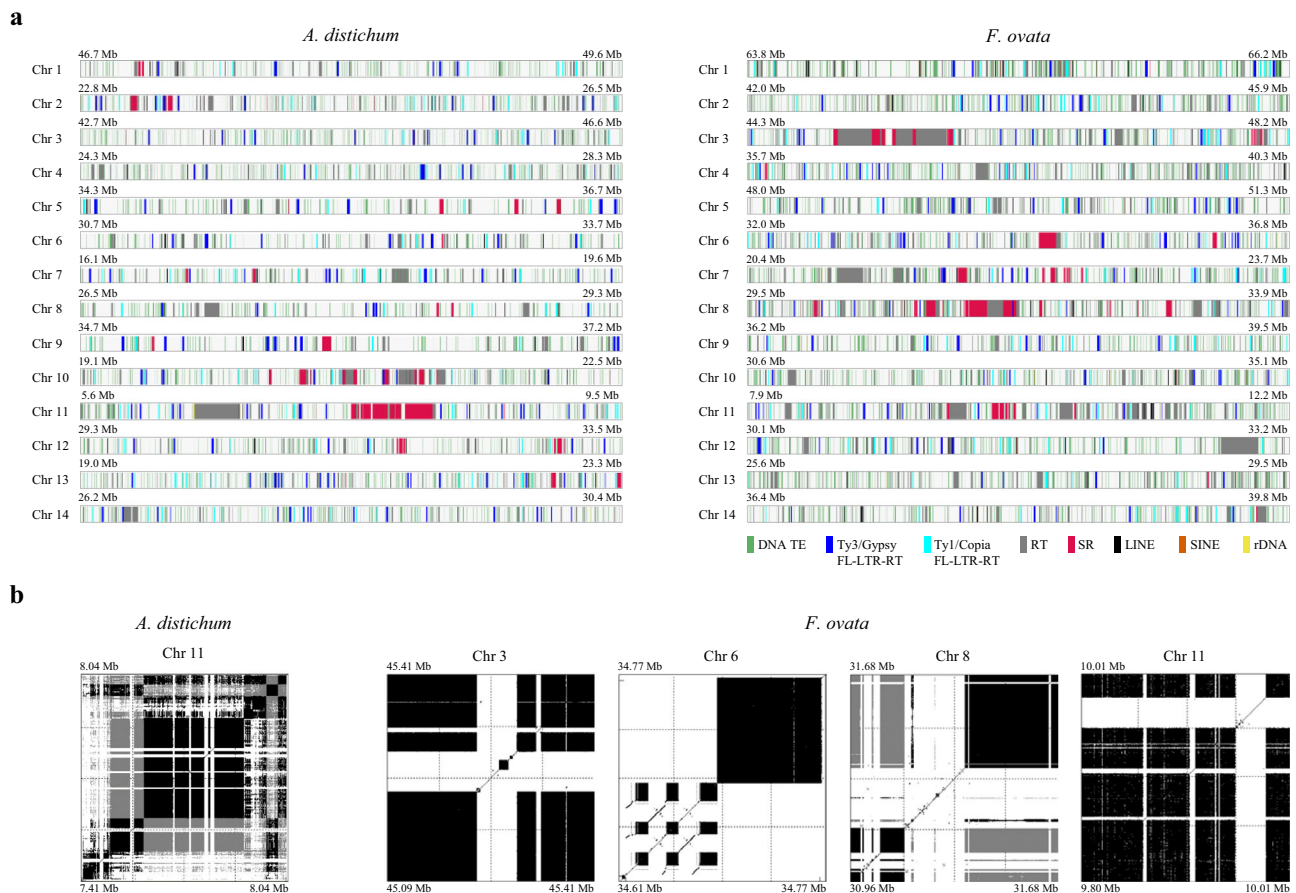


Fig. 3 | Characterization of putative centromere regions of *A. distichum* and *F. ovata* chromosomes. a Distribution of different repetitive sequences, including DNA TEs, RTs, SRs, LINEs, SINEs, and rDNAs within 3.6 Mb centromere regions.

b Self-comparison of the SR-enriched regions in centromeres, showing extensions of repetitive sequence blocks reaching several hundred kilobases.

carotenoid modification were transcriptionally downregulated. Furthermore, the expression levels of genes encoding bHLH and WD40 were higher in *F. ovata* corolla than in *A. distichum*.

Further RNA-seq (Fig. 4d) and quantitative PCR (qPCR) analyses (Fig. 4e) of the base and lobe of the *A. distichum* corolla revealed that the carotenoid biosynthesis-related genes encoding phytoene synthase (PSY), 15-*cis*-zeta-carotene isomerase (Z-ISO), zeta-carotene desaturase (ZDS), polycopene isomerase (CRTISO), lycopene beta cyclase (HYD-B), lycopene epsilon cyclase (HYD-E), and CCS, as well as bHLH genes, were abundantly expressed in the pale yellow base. Notably, qPCR showed that the expression levels of *PSY* (Ad07.20820) and *CRTISO* (Ad02.34180) were 187- and 28-fold higher, respectively, in the base compared to the lobe. In contrast, no or considerably lower expression levels of these genes were observed in the white lobe. Conversely, genes encoding NCED and ethylene responsive factors (ERF), which are negative regulators of carotenoid biosynthesis, showed increased expression in the white lobe. These results are consistent with the observed corolla color of *A. distichum*. The downregulation of bHLH genes in the MBW complex and genes involved in beta-carotene and xanthophyll synthesis, together with the upregulation of negative regulatory transcription factor genes and genes involved in carotenoid modification, are likely to reduce the accumulation of carotenoids in the lobes, resulting in the white corolla of *A. distichum*. Unfortunately, little correlation was found between DNA methylation and gene expression in the two tissues. The differences in ^{13}C levels for the genic regions of the above-mentioned genes (including the -1 kb upstream promoter, gene body, and 1 kb downstream region) between the two tissues were mostly less than 2% (Fig. 4d).

Structural variations in the *A. distichum* genome

For the genome comparison between the pin and thrum plants of *A. distichum*, we obtained approximately 157.9 Gb of HiFi sequence data with an N50 length of 8.4 kb and a Phred quality score of 31. The average read mapping rate and depth to the reference chromosome sequences were 99.2% and 275.7-fold, respectively. The total length of the assembled genome was 1.49 Gb in 12,841 contigs for the pin plant and 1.25 Gb in 2088 contigs for the thrum plant (Tables S4 and S5). The initial variation calls identified 70,811 SVs by DeBreak, of which 2310 SVs were homozygous variants. Most of the homozygous SVs were intergenic, except for 197 SVs (8.5%) in the genic region. Genic SVs were presence/absence variations located in 183 genes, mostly distributed in promoters (117), introns (51), or 3'-untranslated regions (26), and rarely in exons (3) (Supplementary Data 2). The average size of the genic SVs was 1557.2 bp and 82.2% of them were smaller than 1000 bp in size. It is noteworthy that the SVs in 133 genes (72.7%) were caused by the insertion of repetitive sequences. The most frequently inserted repetitive sequences were LTRs (264), followed by DNA TEs (143) and simple sequence repeats (136). Gene ontology analysis of 183 SV genes revealed that 9 transcription factor genes were included in the 183 SV genes, of which 8 genes were related to stress response, while one gene encoded a floral homeotic MADS-box transcription factor (Ad11.09110, *AdAG*) homologous to Arabidopsis *AGAMOUS* (*AtAG*). The protein identity between *AdAG* and *AtAG* was 65%. For further analysis, we BLASTN searched *AdAG* against the HiFi contigs of pin and thrum plants and identified two independent contigs (contigs 360 F and 214 F for pin plant and contigs 70 F and 256 F for thrum plant) containing *AdAG* for each genotype. Contig 70 F was identical to the corresponding region of the

Table 3 | Sequence characteristics of the centromere regions of *A. distichum* and *F. ovata* chromosomes

Species	Chr.	Centromere length (bp)	Number OISRs	DNA TE		Ty1/Copia FL-LTR-RT					Ty3/Gypsy FL-LTR-RT					LINE	SINE	rDNA		Gene			
				hAT-Ac	CMC-EnSpm	MuLE-MuDR	PIF-Harbinger	hAT-Tag1	SIRE	Angela	Tork	Ale	TAR	OTA-Athila	OT-A-Tat			CRM	Tek-ay		Reina	45S	5S
A.distichum	1	2,932,795	337	93	95	124	61	41	0	1	2	2	0	1	3	4	0	0	58	0	1	0	82
	2	3,726,212	1675	153	86	89	81	26	2	1	4	5	6	6	3	0	1	0	68	0	8	0	121
	3	3,948,855	80	176	117	159	103	50	0	1	2	1	2	5	4	1	1	1	75	0	11	0	135
	4	3,930,359	0	172	148	158	103	57	3	0	1	3	3	8	1	1	1	0	87	0	6	1	169
	5	2,481,776	653	108	109	76	51	19	0	0	1	2	1	1	6	0	0	1	31	0	2	0	104
	6	2,957,085	427	98	98	86	67	33	0	1	1	1	1	3	3	4	1	0	84	0	0	0	94
	7	3,508,844	382	156	91	144	58	20	3	2	1	1	6	5	6	0	2	1	70	0	0	0	139
	8	2,782,583	248	113	80	116	98	33	2	2	1	1	1	1	3	2	0	0	56	0	0	0	89
	9	2,439,019	692	99	86	73	41	13	0	3	0	1	1	5	4	0	1	0	38	0	0	0	99
	10	3,404,644	2065	147	117	231	50	29	2	3	3	0	0	0	5	3	1	1	36	0	13	2	124
	11	3,926,280	8218	110	120	198	59	20	4	2	1	2	2	11	4	1	1	0	55	0	6	2	143
	12	4,177,421	1431	203	74	149	114	63	1	1	3	3	1	4	1	2	1	0	106	0	2	0	137
	13	4,314,939	1092	215	120	169	126	26	6	2	6	4	0	6	10	1	0	2	67	0	3	0	215
	14	4,254,259	0	224	122	215	104	40	1	2	3	6	3	12	5	2	1	0	67	0	11	0	189
F.ovata	1	2,456,292	389	176	206	112	39	45	2	1	0	4	1	2	2	0	2	0	151	8	2	1	85
	2	3,865,346	126	251	146	237	94	84	2	2	4	6	1	4	6	0	2	0	297	19	4	2	190
	3	3,808,853	4113	173	140	110	40	28	7	6	0	2	0	6	2	1	2	2	201	12	2	0	88
	4	4,626,475	141	262	251	194	68	41	4	1	4	2	3	3	4	2	4	2	248	11	0	1	120
	5	3,351,558	65	180	184	153	37	9	3	0	2	0	0	7	2	4	1	0	145	2	0	1	81
	6	4,838,040	3305	182	340	175	42	50	1	3	1	6	3	5	6	4	9	1	206	11	2	3	120
	7	3,353,888	2349	195	105	107	67	33	4	3	0	1	1	1	3	2	4	1	194	9	2	6	88
	8	4,414,342	8155	199	172	177	39	20	1	0	1	1	3	5	3	2	0	1	175	10	0	1	82
	9	3,236,852	0	229	112	151	50	30	3	2	3	2	4	3	4	0	1	1	158	16	0	1	93
	10	4,495,998	216	236	245	213	73	37	3	1	3	4	0	5	1	1	2	0	213	27	4	5	163
	11	4,279,319	4198	185	156	216	38	30	5	1	0	1	3	3	5	2	4	1	239	4	13	6	133
	12	3,160,642	78	108	237	146	29	29	3	2	4	2	2	5	1	1	0	0	147	9	0	0	74
	13	3,908,867	115	302	128	196	65	46	1	0	0	3	0	2	0	1	0	0	193	11	0	2	68
	14	3,352,738	122	147	189	140	58	29	2	4	1	2	0	2	2	2	0	1	202	9	2	0	58

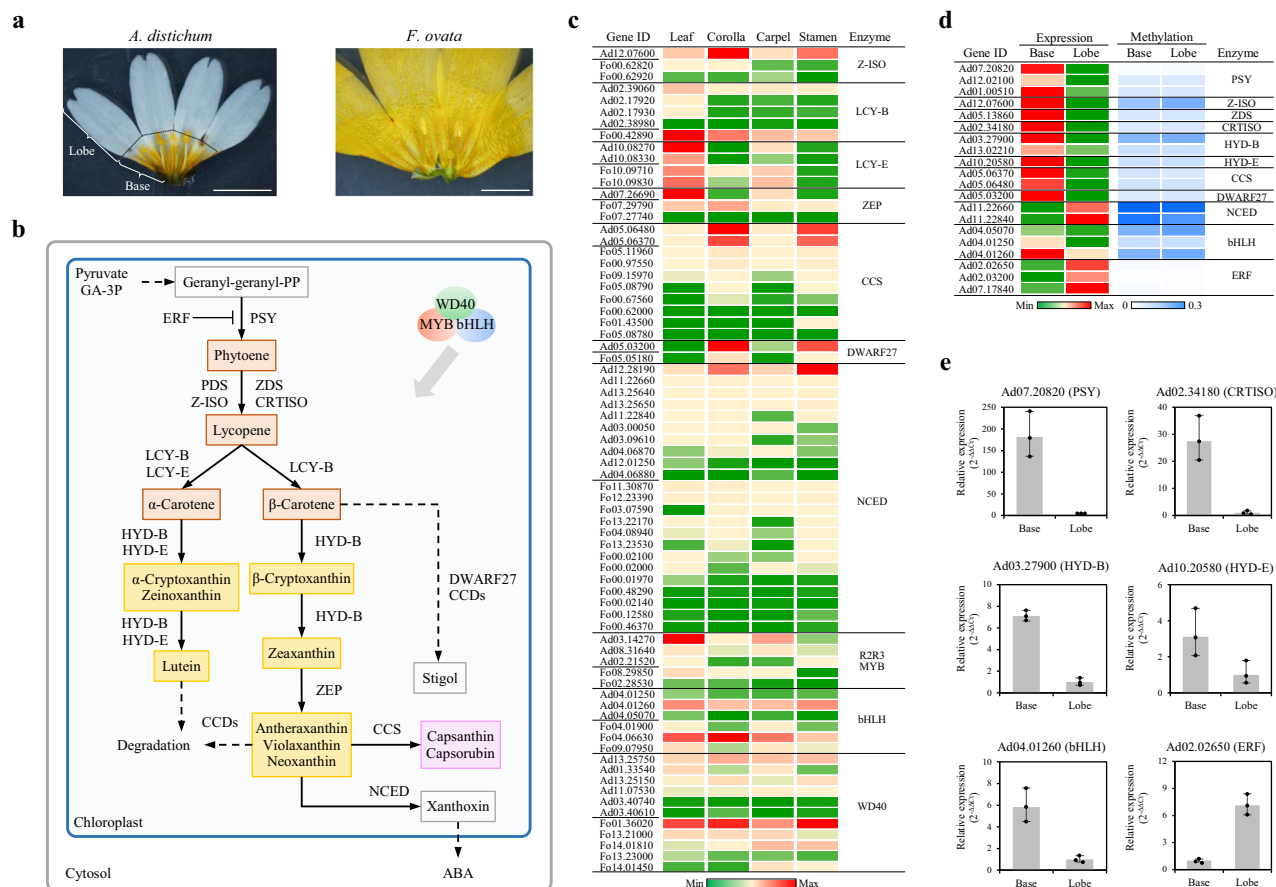


Fig. 4 | Expression of carotenoid biosynthesis-related genes in *A. distichum* (Ad) and *F. ovata* (Fo). **a** Photographs of *A. distichum* (left) and *F. ovata* (right) flowers. The lobe and base of the corolla are indicated. Black dotted lines represent the sampling boundary for transcriptome analysis. White bars represent 5 mm. **b** Diagram of the biosynthesis, modification, and degradation pathways of carotenoid, together with their transcriptional regulators: PSY, phytoene synthase; PDS, phytoene desaturase; Z-ISO, 15-*cis*-zeta-carotene isomerase; ZDS, zeta-carotene desaturase; CRTISO, polycycopene isomerase; LCY-B, lycopene beta cyclase; LCY-E, lycopene epsilon cyclase; HYD-B, beta-ring hydroxylase; HYD-E, carotene epsilon-monoxygenase; ZEP, zeaxanthin epoxidase; DWARF27, beta-carotene isomerase; CCD, carotenoid cleavage dioxygenase; CCS, capsanthin/capsorubin synthase; NCED, 9-*cis*-epoxycarotenoid dioxygenase. Transcriptional regulators include ERF (ethylene response factor), MYB (myeloblastosis family

transcription factor), bHLH (basic Helix-Loop-Helix transcription factor), and WD40 (WD40 repeat protein). **c** Heat map showing differences in the expression of selected genes in four tissues (leaf, corolla, carpel, and stamen) based on expression values (average normalized TMM counts) determined by RNA-seq. **d** Expression and methylation patterns of carotenoid biosynthesis-related genes in the base and lobe of *A. distichum* corolla. Expression levels range from minimum (green) to maximum (red) for each gene. The methylation heat map indicates whether a gene is hypomethylated (white) or hypermethylated (blue) -1 kb upstream from the transcription initiation codon (promoter) to the transcription termination region (gene body). **e** qPCR analysis of selected carotenoid biosynthesis-related genes in the base and lobe of the *A. distichum* corolla. Expression values are presented relative to the less expressed genes. Error bars represent standard errors of three independent replicates per sample.

reference genome, whereas the other contigs were haplotigs of different sequences.

In depth analysis of the contigs revealed that the coding sequences (CDSs) of *AdAG* (762 bp in length) in the contigs are almost identical (99.5%–99.9% identity), with two nonsynonymous mismatches (683 A/G and 720 T/A) resulting in different encoded amino acids (228 Q/R and 240 D/E) at the C-terminus. Of particular importance, the promoter region of *AdAG* was highly polymorphic, showing SVs mainly due to the presence or absence of LTRs and SSRs (Fig. 5a). In the thrum contigs, approximately 7.5 kb and 10.5 kb sequences consisting of Tork FL-LTR-RT (contig 256 F) and fragmented reverse transcriptases of the TAR element (contig 70 F) were inserted 2.7 kb upstream of the transcription initiation codon. In contrast, no insertion sequences were identified in the corresponding promoter region of the pin contigs. An additional Ale FL-LTR-RT insertion far from the transcription initiation codon at 13.2 kb upstream was identified from a pin contig (contig 214 F). We examined the expression profile of *AdAG* in different tissues of *A. distichum* by RNA-seq and qPCR (Fig. 5b). RNA-seq revealed that *AdAG*, as a C class gene of the ABC model for floral organ development, was mainly expressed in the carpel and slightly in the

stamen. Notably, its expression was 11.6-fold higher in the pin type carpel than in the thrum type carpel. The expression of a *AdAG* paralog (Ad11.23870) was very low in all tissues examined. Further analysis of eight independent genotypes by qPCR showed significantly higher expression levels of *AdAG* in pin type carpels than in thrum type carpels ($p < 0.01$). No differential expression (>2 fold-change) of A (*AdAP1* and *AdAP2*) and B (*AdAP3* and *AdPI*) class genes of the ABC model was observed between flower tissues of pin and thrum plants (Fig. S6). Therefore, it is likely that the transcriptional repression of *AdAG* in the thrum plant is caused by the insertion of repetitive sequences into its promoter region.

Chromosome-level synteny of the Oleaceae species

Alignment of homologous pseudo-chromosomes among the *A. distichum*, *F. ovata*, and *F. suspensa* assemblies revealed global one-to-one collinear chromosome matches across the 14 chromosomes (Figs. 2a and 6). Chromosome collinearity was more pronounced between *F. ovata* and *F. suspensa* than between either *A. distichum* and *F. ovata* or *A. distichum* and *F. suspensa*. Nevertheless, the overall collinearity patterns remained fundamentally similar across the species. Several regions showed mismatches due

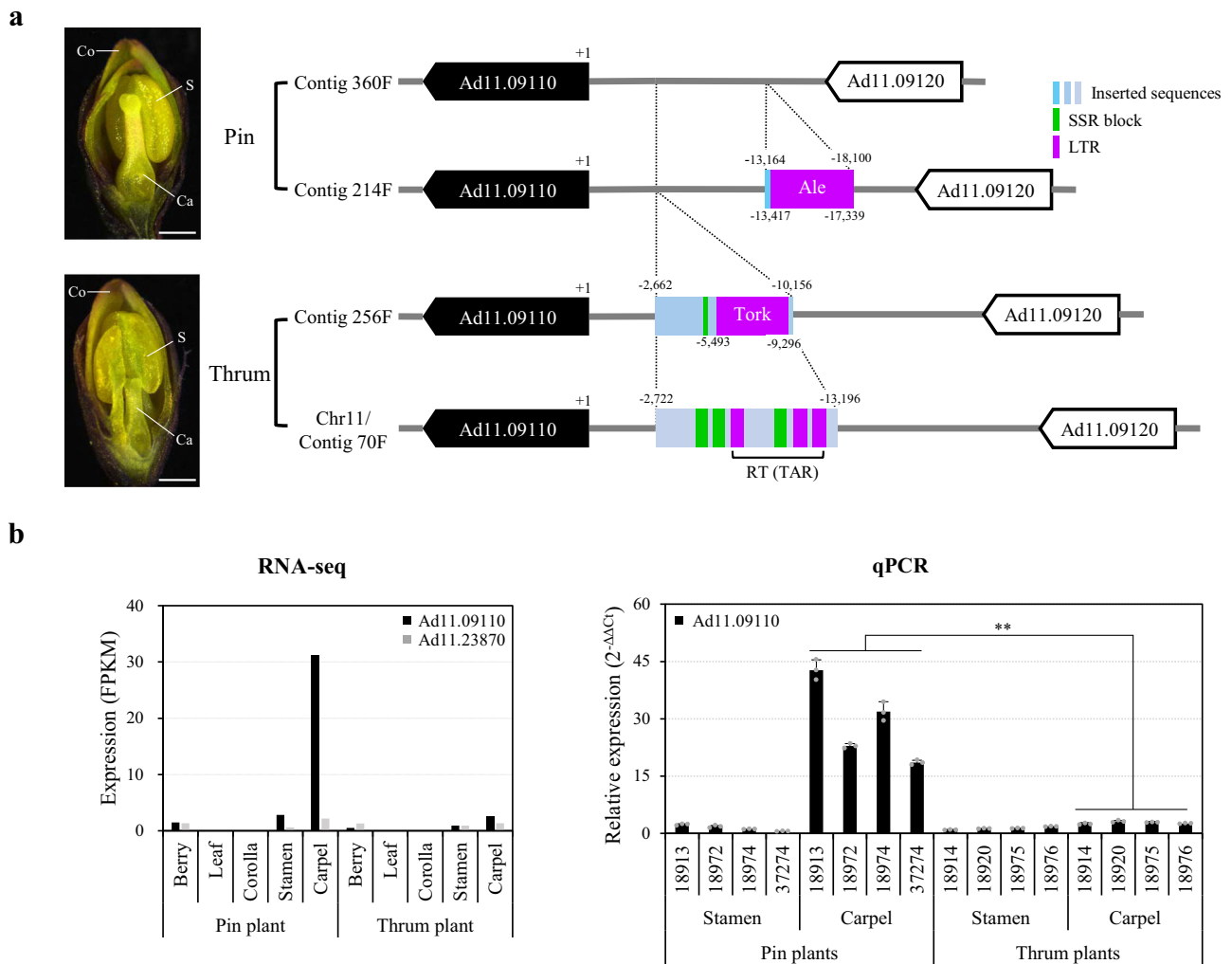


Fig. 5 | SV in the promoter region of *AdAG*. **a** The 7.5 kb and 10.5 kb sequences are inserted 2.7 kb upstream of the *AdAG* initiation codon of a thrum plant; the corresponding regions of a pin plant have no inserted sequences. Pre-flowering buds were harvested in late March. Ca, carpel; Co, corolla; S, stamen. Bars are 0.5 mm. Numbers are positions upstream from the initiation codon (+1). Characteristics of the inserted sequences are shown in colors. Ale and Tork are FL-LTR-RTs, whereas

RT is reverse transcriptase fragments of TAR element. **b** Expression levels of *AdAG* in different tissues as determined by RNA-seq and qPCR. Values represent FPKM for RNA-seq and relative expression based on the comparative cycle threshold method ($2^{-\Delta\Delta C_t}$) using plant ID 18914 as a reference for qPCR. Error bars depict the standard deviations of three independent biological replicates. Asterisks indicate statistical significance (** $p < 0.01$) between pin and thrum plants by t-test.

to inversions or insertions/deletions between the syntenic counterpart chromosomes. However, no translocations and duplications of chromosome blocks were detected. These results suggest that *Abeliophyllum* and *Forsythia* species are closely related sibling species with nearly similar genome structures, likely derived from a common ancestor within the tribe Forsythieae.

A six-way comparison of protein-coding genes from *A. distichum* and *F. ovata* with those from four Oleaceae species (*J. sambac* from the tribe Jasmineae and *F. excelsior*, *O. europaea*, and *S. oblata* from the tribe Oleaeae) identified 11,803 gene families (comprising 166,356 genes) shared by all six species. Chromosome-to-chromosome synteny and collinearity comparisons among the six Oleaceae species based on the orthologous gene pairs identified homologous chromosomes with conserved syntenic blocks (Fig. S7). The proteome comparison revealed one-to-one conserved synteny matches at the chromosome level for 13 chromosome pairs between *A. distichum*/*F. ovata* and *J. sambac*, except for chromosome 8 of *A. distichum*/*F. ovata*, which had no single corresponding chromosome in *J. sambac*. Instead, it showed partial synteny with restricted regions of chromosomes 7 and 9 in *J. sambac*. This finding suggested that chromosome 8 in *A. distichum*/*F. ovata* may have arisen by

chromosome recombination in their lineage (Figs. 6 and S7). Additionally, each chromosome of *A. distichum*/*F. ovata*/*J. sambac* generally corresponded to two chromosomes each from *F. excelsior*, *O. europaea*, and *S. oblata*. Recombination between two or three homologous chromosomes in *F. excelsior*, *O. europaea*, and *S. oblata*, such as chromosomes 1, 9, and 19 in *S. oblata*, 1, 7, and 17 in *F. excelsior*, and 10, 18, and 23 in *O. europaea*, was also detected. These results are consistent with previous reports indicating that species in the tribes Jasmineae and Forsythieae do not share an episode of the WGD event that is common to species of the tribe Oleaeae.

Evolutionary analysis of the Forsythieae genomes

The divergence of Forsythieae genomes within the Oleaceae family was inferred from the Ks of homologous genes and Bayesian evolutionary analysis. *A. distichum* and *F. ovata* shared a single peak in the Ks mode with other Oleaceae (*J. sambac*, *F. excelsior*, *O. europaea*, and *S. oblata*) and with outgroup species (*Sesamum indicum*, *Coffea canephora*, and *Solanum lycopersicum*), with peaks between *A. distichum* and *F. ovata* being nearly identical (Figs. 7 and S8). The Ks distribution for orthologs among these Oleaceae species indicated that *A. distichum*/*F. ovata* diverged successively from *J. sambac*, with a peak Ks of 0.37, and from species within the tribe

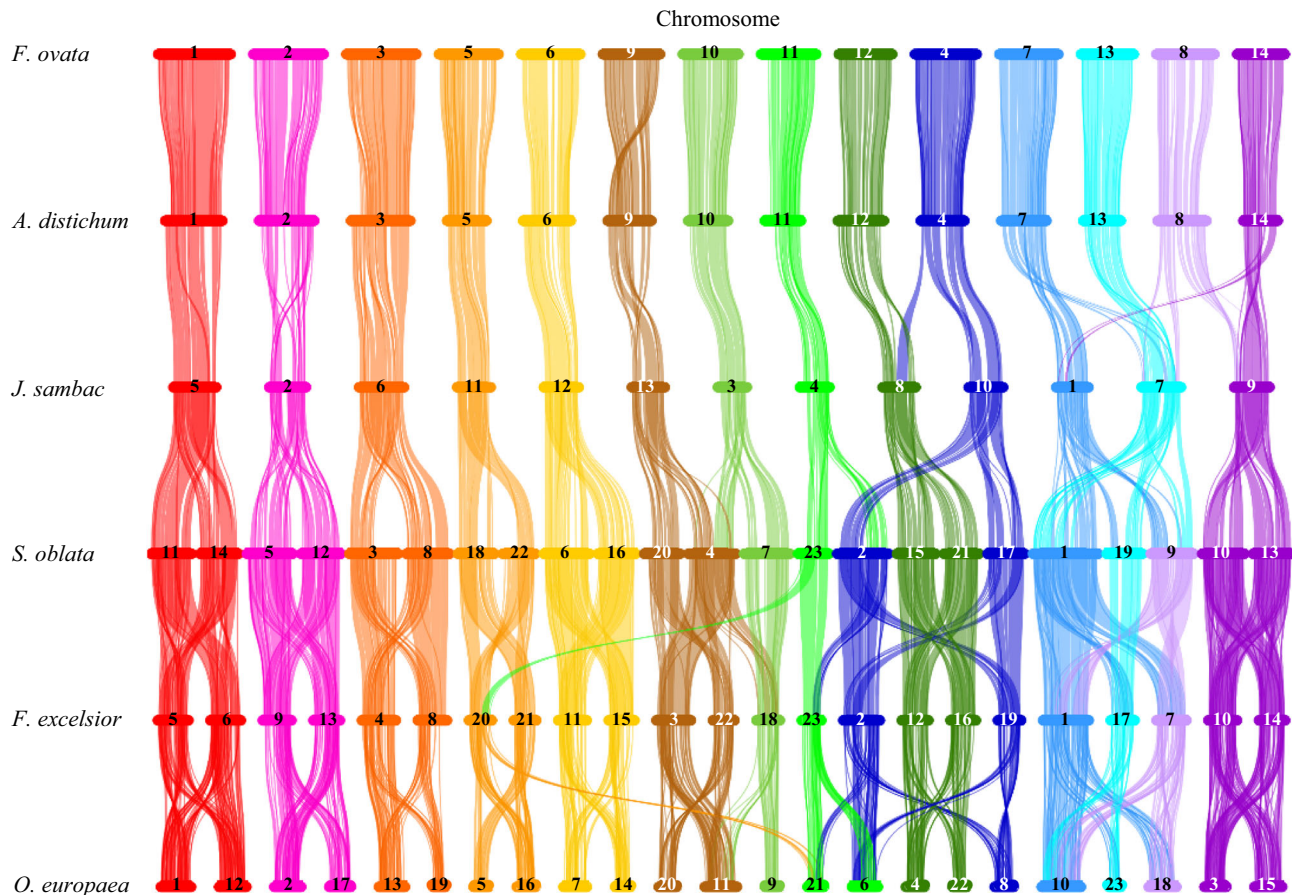


Fig. 6 | Composite chromosome-to-chromosome synteny plot comparing six Oleaceae genomes. The plot was generated using SynVisio, with syntenic blocks in homologous chromosomes represented by the same colors.

Oleaceae, with a peak K_s ranging from 0.21 to 0.25. The most recent divergence was between *A. distichum* and *F. ovata*, with a peak K_s of 0.07 (Fig. 7a). These results were in good agreement with the Bayesian evolutionary analysis and molecular dating based on 118 conserved single-copy genes, which estimated the Forsythieae/Oleaceae-Jasmineae split at 48.44 MYA, the Forsythieae-Oleaceae split at 39.61 MYA, and the *A. distichum*-*F. ovata* split at 13.87 MYA (Fig. 7b). Using the molecular clock hypothesis, which assumes a constant rate of nucleotide changes within phylogenetic lineages, the rate of synonymous substitution per site per year (R) was calculated as $R = (K_s \text{ peak value})/2t$, where t is the speciation time between Oleaceae species and *S. indicum* (Pedaliaceae). For the Oleaceae, this resulted in $R = 4.19 \times 10^{-9}$ substitutions per synonymous site per year based on the average peak K_s value of the orthologs (0.732) and the split time of 93.45 MYA. This rate is approximately 1.96 times lower than that of the Brassicaceae (8.22×10^{-9} substitutions per synonymous site per year), probably due to the longer generation time of the Oleaceae, which are mostly shrubs and trees. Notably, the ancient WGD event, which marked a peak K_s of 0.24–0.25 for paralogs, was retained only in the genomes of the tribe Oleaceae (Fig. 7a). These results support the chromosome-to-chromosome synteny relationships between species and indicate that the WGD, estimated to have occurred 28.64–29.83 MYA, likely arose in the tribe Oleaceae lineage after its divergence from the tribe Forsythieae.

To define an ancestral Oleaceae genome that predates the hypothesized tribe Oleaceae-specific WGD event, we constructed AK genome building blocks based on the shared orthologous gene content and order among the homologous chromosomes of six Oleaceae species (Fig. S9). Since the tribes Forsythieae and Jasmineae may have less fractionated genomes compared to the tribe Oleaceae due to their lack of a WGD event and subsequent genome diploidization, we initially focused on identifying the syntenic blocks in *A.*

distichum/*F. ovata* and *J. sambac*. In cases where no syntenic blocks were detected in specific chromosome regions between these species, we searched for additional syntenic blocks in the chromosomes of Oleaceae species. The mapped blocks were then compared for their order across the chromosomes of the six Oleaceae species and between homologous chromosomes (Figs. 6 and S7). As a result, 12 chromosomes of *J. sambac* were well conserved with at least one entire corresponding chromosome in *A. distichum*, *F. ovata*, *S. oblata*, *F. excelsior*, and *O. europaea*, suggesting a common ancestry. Chromosome 1 of *J. sambac* had no single counterpart chromosome in the other species; instead, it showed synteny with segments of two or three chromosomes, indicating its unique recombination into other chromosomes across the five species. These findings allowed us to reconstruct the AK of six species, consisting of 13 chromosomes labeled A to M (Ancestral Karyotype I). The structure of AK blocks revealed patterns of recombination, duplication, and chromosome loss within each lineage (Fig. 8 and Table S6). In the Forsythieae lineage, 10 AK chromosomes (A to I and L) were retained as single chromosomes, while 5 AK chromosomes (J, K, L, and M) were recombined into 4 chromosomes, specifically chromosomes 4 (I/J), 7 (K/L), 8 (L/M), and 14 (K/M). In the Oleaceae lineage, 5 AK chromosomes (A, B, E, I, and M) were duplicated to form double homologous chromosomes of different sizes, while 6 AK chromosomes (C, D, F, G, H, and J) were conserved either as a single chromosome or as part of one or two recombinant chromosomes (C/F/G, D/H, and H/J). In particular, the K and L chromosomes were not retained as a single chromosome, but were recombined into three chromosomes in *S. oblata* (1, 9, and 19), *F. excelsior* (1, 7, 17), and *O. europaea* (10, 18, 23). The recombination patterns of AK chromosomes in the tribe Oleaceae were fundamentally the same in all species. Based on these data, we hypothesize that the ancestral Forsythieae genome likely consisted of 13 chromosomes. Similarly, the ancestral genome of the

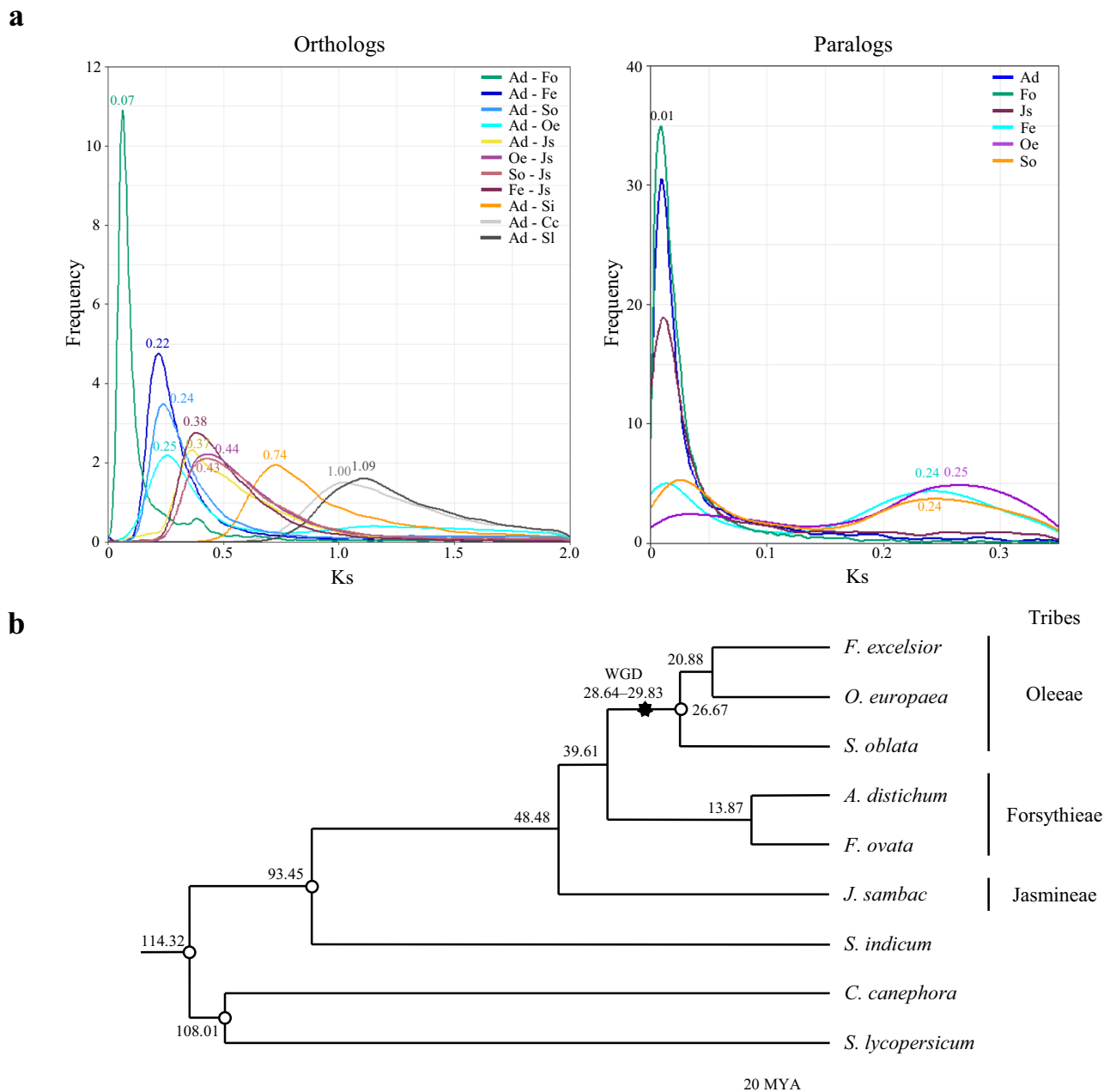


Fig. 7 | Evolutionary analysis of the Oleaceae genomes. a Distribution of Ks values obtained from comparison of orthologous gene sets between two species (left) or paralogs within each species (right). Ad, *A. distichum*; Fo, *F. ovata*; Js, *J. sambac*; Fe, *F. excelsior*; Oe, *O. europaea*; So, *S. oblata*; Si, *S. indicum*; Cc, *C. canephora*; Sl, *S. lycopersicum*. **b** Genome evolution of the Oleaceae species. A maximum likelihood tree with 1000 bootstrap replications was constructed using 118 conserved single-copy gene sets, with speciation dates (MYA) estimated using Bayesian evolutionary

analysis. Open circles represent calibration points for crown ages. A black asterisk indicates a lineage-specific WGD event in Oleaceae. NCBI and China National Center for Bioinformation accession numbers are: *C. canephora*, GCA_900059795; *F. excelsior*, GCA_900149125; *J. sambac*, GWHZHY000000000; *O. europaea*, GCF_002742605; *S. indicum*, GCF_000512975; *S. lycopersicum*, GCF_000188115; *S. oblata*, GWHBHRY000000000.

Oleaceae lineage (Ancestral Karyotype II) also consisted of 13 chromosomes that were duplicated and subsequently rearranged into the current configuration of 23 chromosomes before splitting into *S. oblata*, *F. excelsior*, and *O. europaea* through processes of chromosome recombination and the loss of redundant sub-genomic fractions.

Discussion

Wild plants often exhibit diverse adaptations to their environment and serve as sources of novel genes and biochemicals. Sequencing the genomes of wild plants can provide valuable insights into genetic diversity within and between species, which is essential for conservation efforts, enhancing

economic potential, and unraveling evolutionary processes that illuminate the complex history of their evolution. In a previous study, we obtained chromosome-level genome sequences of *A. distichum* and *F. ovata*, which were assembled into 14 chromosome pseudomolecules, recovering over 99% of the gene space for each genome¹⁰. In the current study, we characterized not only the gene space but also the highly heterochromatic regions of these assemblies at the sequence level, which have rarely been analyzed in previously reported Oleaceae genomes. Comparison of homologous chromosome pairs based on proteome and sequence similarity of the gene space revealed an almost linear extensive syntenic alignment between the genomes of *A. distichum*, *F. ovata*, and *F. suspensa*. This finding

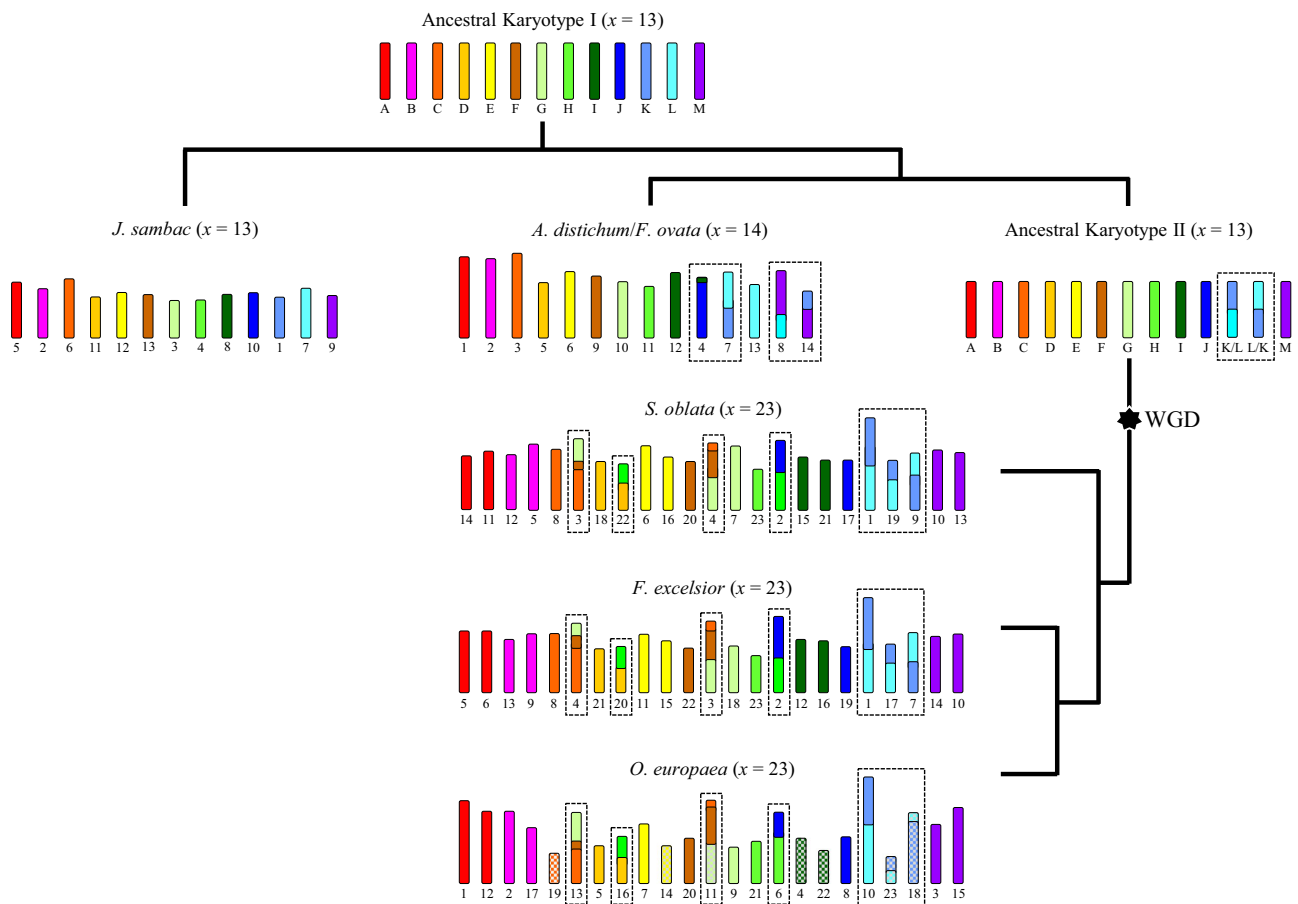


Fig. 8 | Genome rearrangement and evolution of the Oleaceae genomes. Ancestral chromosomes (A to M) were constructed through chromosome-to-chromosome synteny comparisons among six Oleaceae genomes. Blocks are color-coded according to ancestral chromosomes, and numbers indicate the chromosomes of

each species. Dotted boxes represent chromosomes generated by the recombination of ancestral chromosomes. An asterisk indicates a WGD event that occurred in the Oleaceae lineage. Ancestral chromosomes are not to scale, while chromosomes of the Oleaceae species are scaled according to the assembly size.

supports the close relationship of *Abeliophyllum* and *Forsythia* as sister genera within the tribe Forsythieae, despite their clear morphological differences. Moreover, chromosome synteny comparisons, combined with analyses of gene and repeat distribution and whole-genome methylation profiling, enabled us to propose the possible locations of centromeres in the Oleaceae family. Structural analysis of putative centromere regions revealed that *A. distichum* and *F. ovata* centromeres share a similar composition of repetitive sequences. They are rich in hAT-AC, MuLE-MuDR, CMC-EnSpm DNA transposons, OTA-Athila Gypsy elements as dominant FL-LTR-RTs, and ribosomal DNA sequences. Notably, nested tandem arrays of SRs, mostly consisting of the Forsythieae-specific *OISR30–35*, extend up to several hundred kilobases within the centromeres of six or seven chromosomes in each genome. The localization of *OISR30–35* in the *CentFs* satellites of *F. suspensa* suggests that *OISR30–35* are core satellite repeat units that are abundantly distributed in the centromeres of the Forsythieae species. The highly repetitive nature of these SRs and their surrounding transposable elements (TEs) may facilitate the binding of specific centromere proteins or contribute to centromere stability and kinetochore formation³⁸. Alternatively, given their additional distribution in chromosome arms, they may be a source of genome instability leading to chromosome recombination and genetic abnormalities³⁹. To identify centromere sequences that directly interact with the centromeric histone H3 (CENH3), we performed chromatin immunoprecipitation sequencing (ChIP-seq) using an *A. thaliana* anti-CENH3 antibody. However, our results showed non-specific alignments of ChIP-seq reads across both centromeres and chromosome arms, where TEs, LTRs, and other repetitive sequences are distributed (Fig. S10). Considering the 56.3% amino acid identity of CENH3 between *A.*

distichum/F. ovata and *A. thaliana*, the use of an *A. distichum*- or *F. ovata*-specific anti-CENH3 antibody could accurately identify the CENH3-interacting centromere sequences.

Carotenoid biosynthesis genes play a crucial role in determining the yellow and orange colors of flower petals⁴⁰. A key taxonomic character used to classify *A. distichum* and *Forsythia* species is flower color; *Abeliophyllum* typically has white or light pink flowers, whereas *Forsythia* has yellow flowers. Reconstruction of carotenoid biosynthesis pathways and transcriptome analysis have provided a genetic basis for understanding these differences in flower color. We found that the structural genes of the carotenoid biosynthesis pathway are well organized in the *A. distichum* genome and expressed in different tissues, similar to those in *F. ovata*. In particular, the pale yellow base of the *A. distichum* corolla showed expression levels of these genes comparable to the yellow corolla of *F. ovata*. In contrast, the expression of the carotenoid biosynthesis genes *PSY*, *Z-ISO*, *ZDS*, *CRTISO*, *HYD-B*, and *HYD-E* and their putative transcriptional regulators *bHLH* were significantly reduced, while the expression of the carotenoid degradation gene *NCED* was increased in the white lobes of the *A. distichum* corolla. These genes showed little correlation between expression and DNA methylation levels. These results suggest that the differential transcript levels of carotenoid biosynthesis genes in the *A. distichum* corolla are spatially regulated, and the downregulation of these genes in the corolla lobes may play a crucial role in creating the two-colored appearance of the *A. distichum* corolla. In *Arabidopsis*, *PSY* is known to be highly co-expressed with photosynthesis-related genes and its expression is influenced by gibberellic acid, brassinosteroids, and phytochrome-interacting factor⁴¹. Additionally, the post-transcriptional levels of *PSY* are regulated by ORANGE proteins⁴².

Therefore, we anticipate that carotenoid biosynthesis genes are controlled by compartmentalization and spatial regulation within the corolla of *A. distichum*. However, it remains unclear whether developmental signals, cell type-specific transcription factors, phytohormones, or post-transcriptional regulators act as the primary transcriptional regulators. Further studies using single-cell RNA sequencing combined with mass spectrometry imaging of carotenoids may reveal novel cell type-specific regulators involved in the absence of yellow pigmentation in the lobes of *A. distichum* corolla.

Natural variation in plant genomes contributes to the adaptation of plants to different environments as well as to phenotypic traits by providing genetic diversity^{43–45}. Heterostyly is a genetic polymorphism found in approximately 28 angiosperm families that favors outbreeding⁴⁶. This adaptation is characterized by the reciprocal positioning of anthers and stigmas in different floral morphs. A genetic determinant of heterostyly is the S locus, which harbors the causal genes. The known heterostyly determinant gene is *GLO^T* identified from *Primula vulgaris*, showing that thrum plants were hemizygous and pin plants were absent for *GLO^T*⁴⁷. This result indicates that SV of a floral homeotic gene results in floral heteromorphy in *Primula*. Heterostyly is not common in all species of the Oleaceae, but it is particularly common in the Forsythieae species. Genome comparison revealed that the *Primula* S locus and *GLO^T* are not conserved in the Forsythieae genomes, suggesting that heterostyly in the Forsythieae species is a result of convergent evolution. Meanwhile, we have identified SVs in 183 genes between pin and thrum plants of *A. distichum*, most of which are caused by repetitive sequence insertions. Differential insertion of repetitive sequences into the genome can be a major source of genetic differences between genotypes. Among the SV genes, *AdAG* was a unique gene associated with floral organ development. We observed a significant down-regulation of *AdAG* in *A. distichum* thrum plants, presumably due to the insertion of LTR-related large insertion sequences in its promoter region. These results are consistent with previous reports that TE and repetitive sequence insertions can lead to gene loss, affect the expression of nearby genes, or induce epigenetic modifications^{44,48,49}; however, it is still unclear how LTRs insertions alter the expression of *AdAG*. Interestingly, *AdAG* is a C class gene of the ABC model of floral organ development, whereas *GLO^T* is a B class gene in *Primula*. Considering that the genetic loci harboring these genes showed SVs, alteration of *AdAG* expression resulting from LTR insertions may lead to the development of dimorphic flowers in *A. distichum*. Further studies should clarify this hypothesis.

Syntenic relationships between closely related species have been instrumental in dissecting genome evolution, discovering genes and loci, and facilitating information transfer across genomes. Comparative genomic studies of six Oleaceae species with chromosome-level assemblies, representing the tribes Jasmineae (*J. sambac*), Forsythieae (*A. distichum* and *F. ovata*), and Oleae (*F. excelsior*, *O. europaea*, and *S. oblata*), have revealed the dynamic evolutionary history of each lineage and enabled the reconstruction of the 13 AK block structure of their common ancestor (Ancestral Karyotype I). The syntenic relationships between these species provided clear evidence for the absence of a recent WGD event in the Jasmineae and Forsythieae lineages. Although *J. sambac* and *A. distichum*/*F. ovata* have chromosomes with extensive homology, the extent of rearrangements and recombination distinguishes the Jasmineae and Forsythieae genomes. The *J. sambac* genome shows a more ancestral structure than those of *A. distichum* and *F. ovata*, as its chromosomes have undergone fewer recombination, with the least amplification of repetitive sequences such as SRs. In the tribe Jasmineae lineage, most species of the genus *Jasminum* have 13 chromosome pairs, except for a few species with abnormal chromosome numbers, such as *J. ilicifolium* (2n = 39) and *J. nudiflorum* (2n = 52)³. These species probably arose by triploidization or tetraploidization of 13 AKs. However, the basal chromosome number of the Jasmineae lineage requires further analysis, as most species in the genus *Menodora* contain 11 chromosome pairs, suggesting that the genome of *Menodora* may harbor the simplest organization of the AK structure. The genome sequences of *Menodora* species would provide fundamental information to address this concern. Of particular interest, the genomes of Forsythieae and Oleae shared the

recombinant chromosomes consisting of K and L blocks. This finding suggests that recombination of the K and L blocks of the AK may have led to the divergence of the Forsythieae and Oleae lineages from the Jasmineae. The split between Forsythieae and Oleae may have resulted from additional recombination of K, L, and M blocks within the Forsythieae lineage. Furthermore, the common WGD event in the Oleae lineage was likely an allopolyploidy involving the Ancestral Karyotype II, given the variation in size and gene content between the duplicated chromosome pairs in each Oleae species. Subsequent recombination and genome diploidization of the duplicated genome, including the loss of redundant genes and sub-genome blocks, likely led to the formation of an ancestral Oleae genome consisting of 23 chromosomes, from which the current genomes of Oleae species diverged. The same AK block organization of chromosomes among the Oleae species strongly supports this hypothesis. Meanwhile, the split between *A. distichum* and *F. ovata* is estimated to have occurred at 13.87 MYA, during the Middle Miocene epoch, which was characterized by significant climate change and geographic restructuring⁵⁰. We hypothesize that the unique organization of repeat elements and the tissue-specific regulation of gene expression would have contributed to the evolution of species-specific traits in the Forsythieae. Consequently, the divergence of the genera *Abeliophyllum* and *Forsythia* likely involved adaptations to their respective environments, resulting in distinct evolutionary paths despite their genetic proximity in East Asia.

In conclusion, the comparison of gene space and heterochromatic regions using chromosome-level genome assemblies revealed that the *Abeliophyllum* and *Forsythia* genomes retain the ancestral genome structure of the Oleaceae. They are closely related but differ in the organization of repetitive sequences which may have contributed to genome divergence. With ongoing advances in genomic technologies and bioinformatics software, the genomic data from this study will serve as a platform for the development of pan-genomes for Forsythieae species and their biological and conservation studies.

Methods

Plant materials

We selected *A. distichum* accessions KNKB198505000390 (pin plant) and KNKB198505000391 (thrum plant) and *F. ovata* accession KNKB202402200001 (thrum plant) at the Korea National Arboretum (KNA) for this study. The accessions were originally collected from the shady slopes of Goesan-gun, Chungcheongbuk-do (*A. distichum*) and Sokcho-si, Gangwon-do (*F. ovata*), Korea, and then planted at KNA.

Genome sequences and annotations

The genome assemblies, annotations, transcriptome, and methylome data for accessions KNKB198505000391 (*A. distichum*) and KNKB202402200001 (*F. ovata*) were deposited in the NCBI database under Bioproject IDs PRJNA1086675 and PRJNA1086660¹⁰. We extracted the pseudo-chromosome sequences and relevant annotations of genes and repetitive sequences from the genome assemblies for downstream analysis. Sequences and annotations of the Oleaceae genomes were downloaded from NCBI and China National Center for Bioinformation under the following accession numbers: *F. excelsior*, GCA_900149125; *F. suspensa*, JAHHPY000000000; *J. sambac*, GWHZHY000000000; *O. europaea*, GCF_002742605; *S. oblata*, GWHBHY000000000.

Repeat analysis

Chromosome sequences were searched using BLASTN with a cutoff of 1E⁻¹⁰ against 5S rDNA and 45S rDNA sequences. RTs were identified using LTR_FINDER v.1.05⁵¹. The resulting sequences with long terminal repeats were then searched using BLASTX (cutoff of 1E⁻¹⁰ and >70% coverage) against mobile genetic elements in RepeatExplorer⁵² to extract FL-LTR-RTs. FL-LTR-RTs were classified into different families based on their similarity to plant RTs from RepeatExplorer. DNA TEs were identified using RepeatMasker v.4.0.5 (<http://www.repeatmasker.org>) and RepeatModeler v.1.0.8 (<http://www.repeatmasker.org>). SR arrays longer than 1 kb were

searched using KMC 3 v.3.2.4 with parameters $-fm -k75 -ci20 -cs100000$ ⁵³ and repeat units were reconstructed using the Satellite Repeat Finder⁵⁴. Pairwise alignments of the heterochromatic regions were performed using DNAMAN v.10.0 (Lynnon Biosoft, San Ramon, CA, USA) with parameters of 100–200 bp window size and 1–2 bp mismatches. Tandem repeats in the heterochromatic regions were identified using Dot2dot⁵⁵ with parameters of MinMotifLen 50, MaxMotifLen 1000, MinMatch 0.5, MaxGaps 8, MaxInsert 1, Tolerance 0.2, and MinTRLen 20. Consensus sequence units of tandem repeats were determined by multiple alignment of repeat sequences using Geneious Prime 2023.1.2 (Biomatters, Auckland, New Zealand) with default parameters.

Transcriptome analysis and bisulfite sequencing

Gene expression analysis in different tissues was performed using our previously generated RNA-seq reads deposited in the Sequence Read Archive under accession numbers of SRR28349245–SRR28349254 for *A. distichum* and SRR28365181–SRR28365185 for *F. ovata*. For *A. distichum* corolla lobe and base transcriptome and methylome data, flowers were collected in early April from the *A. distichum* accession KNKB198505000391 and lobe and base tissues were separated. Total RNA was isolated from each tissue using the cetyltrimethylammonium bromide method⁵⁶ and used for library construction with an Illumina TruSeq Stranded Total RNA Library Prep Plant Kit (Illumina, San Diego, CA, USA). Sequencing was performed by Macrogen (Seoul, Korea). Quality filtering of RNA-seq reads (2×101 bp) was performed using Trimmomatic v.0.36⁵⁷ with parameters of LEADING:20, TRAILING:20, SLIDINGWINDOW:4:20, and MINLEN:36. Paired-end reads from three biological replicates were aligned end-to-end to the CDSs of protein-coding genes and RNA genes using STAR v.2.7.9⁵⁸ with default parameters. The mapped reads for each gene model were normalized and the average trimmed mean of M (TMM) and fragments per kilobase of transcripts per million mapped reads (FPKM) values were calculated. Genes in the metabolic pathway network were annotated using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database (www.genome.jp/kegg/pathway.html) and their expression levels were analyzed based on the TMM counts. For qPCR analysis, first-strand cDNA was synthesized from RNA extracted from flower tissues using the Revert Aid First-Strand cDNA Synthesis Kit (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer's instructions. cDNAs were diluted 10-fold and qPCR was performed using PowerUp SYBR Green Master Mix (Thermo Fisher Scientific), primer pairs designed to amplify the target genes (Table S7) and the StepOne Plus™ Real-Time PCR System (Thermo Fisher Scientific). The $2^{-\Delta\Delta Ct}$ method was used to quantify gene expression with Ad01.05870, encoding glyceraldehyde 3-phosphate dehydrogenase, as a reference. For whole-genome bisulfite sequencing (BS), fragmented DNA with a size of 200 bp was treated with bisulfite using an EZ DNA Methylation-Gold Kit (Zymo Research, Irvine, CA, USA) and sequencing libraries were constructed using an Accel-NGS Methyl-Seq DNA Library Kit (Swift Biosciences, Inc., Ann Arbor, MI, USA). A total of 93.8 Gb of quality-filtered Illumina BS sequences were aligned to the genome assemblies using BSMAP v2.9⁵⁹ with default parameters. Only uniquely mapped reads were selected and PCR duplicates were removed using SAMBAMBA v0.5.9⁶⁰. The methylation ratio of each cytosine position was extracted from the mapping results. Methylation patterns were partitioned by context (CG, CHG, and CHH) and the distribution of the mean cytosine methylation levels was calculated for each chromosome using a sliding window of 1–100-kb.

Long read sequencing and identification of structural variation

High-molecular-weight genomic DNA was extracted from leaf tissue of *A. distichum* accessions KNKB198505000390 (pin plant) and KNKB198505000391 (thrum plant), respectively, using a nuclear isolation method for long-read sequencing⁶¹. For PacBio Revio sequencing, single-molecule real-time DNA sequencing libraries with an insert size of 20 kb were constructed using the HiFi prep kit (Pacific Biosciences, Menlo Park,

CA, USA). Sequencing was performed by Macrogen (Seoul, Korea). The HiFi reads were assembled into contigs using the FALCON assembler v2.1.4⁶² with the following parameters: length cutoff, 10 kb; max difference, 80; max coverage, 80; and minimum coverage, 2. In parallel, the HiFi reads were mapped to the *A. distichum* reference genome using minimap2 v2.26-r1175⁶³ with default parameters. SV events in the mapping results were detected using DeBreak v1.0.2⁶⁴ with parameters of $-rescue_large_ins$ and $-rescue_dup$. SVs supported by more than 80% homozygous reads were extracted and visualized using the IGV v2.15.4 program⁶⁵.

Comparative genome analysis

Comparison of the annotated genes between *A. distichum* and *F. ovata* was performed using all-against-all BLASTP analysis (cutoff of $1E^{-10}$ and >70% coverage). Alignment of homologous pseudo-chromosomes from the *A. distichum* and *F. ovata* assemblies was analyzed using MUMMER⁶⁶ and SyMAP⁶⁷ with default options. Orthologous gene sets among *A. distichum*, *F. ovata*, *F. excelsior*, *J. sambac*, *O. europaea*, and *S. oblata* were identified by a six-way comparison using OrthoFinder⁶⁸. We excluded *F. suspensa* because only the chromosome pseudo-molecule sequences, excluding gene annotations, are deposited in GenBank. Genome-to-genome synteny and collinearity among six Oleaceae species, based on orthologous gene pairs, was determined using McScanX⁶⁹, and the results were visualized using SynVisio (<https://github.com/kiranbandi/synvisio>). Frequency distributions of synonymous substitutions (Ks) between homologous genes were calculated from the pairwise alignments of orthologs and paralogs. Protein sequences were aligned using ClustalW v.2.1⁷⁰. After gap removal, protein sequence alignments were converted to codon alignments using PAL2NAL v.14.0⁷¹ with a parameter $-nogap$. The Ks values from pairwise alignments of homologous sequences were determined using the CODEML program implemented in the PAML v.4.9j package⁷². The Ks distribution was plotted using ggplot2⁷³.

Molecular dating

To estimate the divergence times of the Oleaceae genomes, single-copy orthologs conserved in six Oleaceae species and three outgroup species (sesame, coffee, and tomato) were selected using reciprocal BLASTP (cutoff of $1E^{-10}$ and >70% coverage). Bayesian evolutionary analysis of the multiple aligned CDS sequences was performed using BEAST v1.10.4⁷⁴, following our previous report⁷⁵ with four calibration points. We constrained the crown ages of Lamiids to 110.84 MYA, Rubiaceae–Solanaceae to 110.09 MYA, and Oleaceae–Pedaliaceae to 94.91 MYA, following Zhang et al.⁷⁶. In addition, we applied a calibration point for the crown age of *Syringa–Osmanthus* at 27.61 MYA, based on Wang et al.¹³. Results were visualized using FigTree v1.4.2 (<http://tree.bio.ed.ac.uk/software/figtree>). Evolution of the orthologous gene family was analyzed using CAFE v4.2.1⁷⁷ with a cutoff of $P < 0.05$.

Statistics and reproducibility

To statistically test the difference in gene expression by qPCR, the independent t-test was performed using Excel. All the values were given as the mean $\pm$ SE. In Fig. 4c–e, Fig. 5b, and Supplementary Fig. S6, we used $n = 3$ biologically independent samples.

Reporting summary

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

Data availability

All sequencing data and genome assemblies of *A. distichum* and *F. ovata* used in this study are available in the NCBI database under the Bioproject IDs PRJNA1086675 for *A. distichum* and PRJNA1086660 for *F. ovata*. The assembled genomes have been deposited in the NCBI Assembly database (accession numbers JBFOLK000000000 and JBFOLJ000000000). The genome assemblies and annotation data are also available in figshare

(<https://doi.org/10.6084/m9.figshare.25539493>). The sequence read data are available in the Sequence Read Archive with accession numbers SRP495333 for *A. distichum* and SRP495335 for *F. ovata*. Numerical source data for Figs. 4e and 5b can be found in the Supplementary Data 3 file. All other data are available from the corresponding author on reasonable request.

Code availability

The software and parameters used in this study are described in the Methods section. No specific custom codes or scripts were utilized. Data processing was conducted according to the manuals and protocols provided with the respective software.

Received: 18 September 2024; Accepted: 5 February 2025;

Published online: 18 February 2025

References

- Dupin, J. et al. Resolving the phylogeny of the olive family (Oleaceae): Confronting information from organellar and nuclear genomes. *Genes (Basel)* **11**, 1508 (2020).
- Taylor, H. Cyto-taxonomy and phylogeny of the Oleaceae. *Brittonia* **5**, 337–367 (1945).
- George, K., Geethamma, S. & Ninan, A. Chromosome evolution in Oleaceae. *J. Cytol. Genet.* **24**, 71–77 (1989).
- Wallander, E. & Albert, V. Phylogeny and classification of Oleaceae based on *rps16* and *trnL-F* sequence data. *Am. J. Bot.* **87**, 1827–1841 (2000).
- Chung, G. et al. A checklist of endemic plants on the Korean Peninsula II. *Korean J. Pl. Taxon.* **53**, 79–101 (2023).
- Kim, D. & Kim, J. Molecular phylogeny of tribe Forsythieae (Oleaceae) based on nuclear ribosomal DNA internal transcribed spacers and plastid DNA *trnL-F* and *matK* gene sequences. *J. Plant Res.* **124**, 339–347 (2011).
- Hong, S.-P. & Han, M.-J. The floral dimorphism in the rare endemic plant, *Abeliophyllum distichum* Nakai (Oleaceae). *Flora* **197**, 317–325 (2002).
- Lee, S. Palynological contributions to the taxonomy of family Oleaceae, with special emphasis on genus *Forsythia* (tribe Forsythieae). *Korean J. Pl. Taxon.* **41**, 175–181 (2011).
- Ha, Y., Kim, C., Choi, K. & Kim, J. Molecular phylogeny and dating of Forsythieae (Oleaceae) provide insight into the Miocene history of eurasian temperate shrubs. *Front. Plant Sci.* **9**, 299304 (2018).
- Jang, H. et al. Chromosome-level assemblies of the endemic Korean species *Abeliophyllum distichum* and *Forsythia ovata*. *Sci. Data* **11**, 1372 (2024).
- Unver, T. et al. Genome of wild olive and the evolution of oil biosynthesis. *Proc. Natl. Acad. Sci. USA* **114**, E9413–E9422 (2017).
- Rao, G. et al. De novo assembly of a new *Olea europaea* genome accession using nanopore sequencing. *Hortic. Res.* **8**, 64 (2021).
- Wang, Y. et al. A chromosome-level genome of *Syringa oblata* provides new insights into chromosome formation in Oleaceae and evolutionary history of lilacs. *Plant J.* **111**, 836–848 (2022).
- Sollars, E. et al. Genome sequence and genetic diversity of European ash trees. *Nature* **541**, 212–216 (2017).
- Yang, X. et al. The chromosome-level quality genome provides insights into the evolution of the biosynthesis genes for aroma compounds of *Osmanthus fragrans*. *Hortic. Res.* **5**, 72 (2018).
- Huff, M. et al. A high-quality reference genome for *Fraxinus pennsylvanica* for ash species restoration and research. *Mol. Ecol. Res.* **22**, 1284–1302 (2022).
- Ma, B. et al. Lilac (*Syringa oblata*) genome provides insights into its evolution and molecular mechanism of petal color change. *Commun. Biol.* **5**, 686 (2022).
- Chen, L. et al. *Syringa oblata* genome provides new insights into molecular mechanism of flower color differences among individuals and biosynthesis of its flower volatiles. *Front. Plant Sci.* **13**, 1078677 (2022).
- Chen, G. et al. The Jasmine (*Jasminum sambac*) genome provides insight into the biosynthesis of flower fragrances and jasmonates. *Genom. Proteom. Bioinf.* **21**, 127–149 (2023).
- Xu, S. et al. A high-quality genome assembly of *Jasminum sambac* provides insight into floral trait formation and Oleaceae genome evolution. *Mol. Ecol. Resour.* **22**, 724–739 (2022).
- Li, Y. et al. The updated weeping forsythia genome reveals the genomic basis for the evolution and the forsythidin and forsythoside A biosynthesis. *Hortic. Plant J.* **9**, 1149–1161 (2023).
- Cui, J. et al. The gap-free genome of *Forsythia suspensa* illuminates the intricate landscape of centromeres. *Hortic. Res.* **11**, uhae185 (2024).
- Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. & van der Knaap, E. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science* **319**, 1527–1530 (2008).
- Guan, J. et al. Genome structure variation analyses of peach reveal population dynamics and a 1.67 Mb causal inversion for fruit shape. *Genome Biol.* **22**, 13 (2021).
- Espley, R. et al. Multiple repeats of a promoter segment causes transcription factor autoregulation in red apples. *Plant Cell* **21**, 168–183 (2009).
- Guo, J. et al. An integrated peach genome structural variation map uncovers genes associated with fruit traits. *Genome Biol.* **21**, 258 (2020).
- Wu, M., Luo, Z. & Cao, S. Promoter variation of the key apple fruit texture related gene *MdPG1* and the upstream regulation analysis. *Plants (Basel)* **12**, 1452 (2023).
- Hu, L. et al. Promoter variations in a homeobox gene, *BnA10.LMI1*, determine lobed leaves in rapeseed (*Brassica napus* L.). *Theor. Appl. Genet.* **131**, 2699–2708 (2018).
- Zuo, W. et al. A maize wall-associated kinase confers quantitative resistance to head smut. *Nat. Genet.* **47**, 151–157 (2015).
- Gobena, D. et al. Mutation in sorghum *LOW GERMINATION STIMULANT 1* alters strigolactones and causes *Striga* resistance. *Proc. Natl. Acad. Sci. USA* **114**, 4471–4476 (2017).
- Ashikawa, I. et al. Two adjacent nucleotide-binding site-leucine-rich repeat class genes are required to confer *Pikm*-specific rice blast resistance. *Genetics* **180**, 2267–2276 (2008).
- Roquis, D. et al. Genomic impact of stress-induced transposable element mobility in Arabidopsis. *Nucleic Acids Res.* **49**, 10431–10447 (2021).
- Naish, M. & Henderson, I. The structure, function, and evolution of plant centromeres. *Genome Res.* **34**, 161–178 (2024).
- Jeong, Y.-M. et al. Elucidating the triplicated ancestral genome structure of radish based on chromosome-level comparison with the *Brassica* genomes. *Theor. Appl. Genet.* **129**, 1357–1372 (2016).
- Cho, A. et al. An improved *Raphanus sativus* cv. WK10039 genome localizes centromeres, uncovers variation of DNA methylation and resolves arrangement of the ancestral *Brassica* genome blocks in radish chromosomes. *Theor. Appl. Genet.* **135**, 1731–1750 (2022).
- Ahmed, H. et al. Einkorn genomics sheds light on history of the oldest domesticated wheat. *Nature* **620**, 830–838 (2023).
- Thakur, J., Packiaraj, J. & Henikoff, S. Sequence, chromatin and evolution of satellite DNA. *Int. J. Mol. Sci.* **22**, 4309 (2021).
- Hartley, G. & O'Neill, R. Centromere repeats: Hidden gems of the genome. *Genes (Basel)* **10**, 223 (2019).
- McFarlane, R. & Humphrey, T. A role for recombination in centromere function. *Trends Genet.* **26**, 209–213 (2010).
- Tanaka, Y., Sasaki, N. & Ohmiya, A. Biosynthesis of plant pigments: anthocyanins, betalains and carotenoids. *Plant J.* **54**, 733–749 (2008).
- Meier, S., Tzfadia, O., Vallabhaneni, R., Gehring, C. & Wurtzel, E. A transcriptional analysis of carotenoid, chlorophyll and plastidial

- isoprenoid biosynthesis genes during development and osmotic stress responses in *Arabidopsis thaliana*. *BMC Syst. Biol.* **5**, 77 (2011).
42. Zhou, X. et al. *Arabidopsis* OR proteins are the major posttranscriptional regulators of phytoene synthase in controlling carotenoid biosynthesis. *Proc. Natl. Acad. Sci. USA* **112**, 3558–3563 (2015).
 43. Weigel, D. & Nordborg, M. Population genomics for understanding adaptation in wild plant species. *Annu. Rev. Genet.* **49**, 315–338 (2015).
 44. Kim, J. et al. Effect of structural variation in the promoter region of *RsMYB1.1* on the skin color of radish taproot. *Front. Plant Sci.* **14**, 1327009 (2024).
 45. Yuan, Y., Bayer, P., Batley, J. & Edwards, D. Current status of structural variation studies in plants. *Plant Biotechnol. J.* **19**, 2153–2163 (2021).
 46. Barrett, S. The evolution of plant sexual diversity. *Nat. Rev. Genet.* **3**, 274–284 (2002).
 47. Li, J. et al. Genetic architecture and evolution of the S locus supergene in *Primula vulgaris*. *Nat. Plants* **2**, 1–7 (2016).
 48. Gill, R. et al. On the role of transposable elements in the regulation of gene expression and subgenomic interactions in crop genomes. *Crit. Rev. Plant Sci.* **40**, 157–189 (2021).
 49. Zhang, L. et al. Genome-wide patterns of large-size presence/absence variants in sorghum. *J. Integr. Plant Biol.* **56**, 24–37 (2014).
 50. Steinthorsdottir, M. et al. The Miocene: The Future of the Past. *Paleoceanogr. Paleoclimatol.* **36**, e2020PA004037 (2021).
 51. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W258 (2007).
 52. Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792–793 (2013).
 53. Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
 54. Zhang, Y., Chu, J., Cheng, H. & Li, H. De novo reconstruction of satellite repeat units from sequence data. *Genome Res.* **33**, 1994–2001 (2023).
 55. Genovese, L., Mosca, M., Pellegrini, M. & Geraci, F. Dot2dot: Accurate whole-genome tandem repeats discovery. *Bioinformatics* **35**, 914–922 (2019).
 56. Chang, S., Puryear, J. & Cairney, J. A simple and efficient method for isolating RNA from pine trees. *Plant Mol. Biol. Rep.* **11**, 113–116 (1993).
 57. Bolger, A., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2020 (2014).
 58. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
 59. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinform.* **10**, 232 (2009).
 60. Tarasov, A., Vilella, A., Cuppen, E., Nijman, I. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
 61. Zhang, M. et al. Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. *Nat. Protoc.* **7**, 467–478 (2012).
 62. Chin, C. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
 63. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
 64. Chen, Y. et al. Deciphering the exact breakpoints of structural variations using long sequencing reads with DeBreak. *Nat. Commun.* **14**, 283 (2023).
 65. Thorvaldsdóttir, H., Robinson, J. & Mesirov, J. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
 66. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
 67. Soderlund, C., Bomhoff, M. & Nelson, W. SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res.* **39**, e68 (2011).
 68. Emms, D. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
 69. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
 70. Larkin, M. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
 71. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
 72. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
 73. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, 2016).
 74. Drummond, A. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
 75. Baek, S. et al. Draft genome sequence of wild *Prunus yedoensis* reveals massive inter-specific hybridization between sympatric flowering cherries. *Genome Biol.* **19**, 127 (2018).
 76. Zhang, C. et al. Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications. *Mol. Biol. Evol.* **37**, 3188–3210 (2020).
 77. Han, M., Thomas, G., Lugo-Martinez, J. & Hahn, M. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol. Biol. Evol.* **30**, 1987–1997 (2013).

Acknowledgements

This work was supported by a grant from the Scientific Research Program (KNA1-1-13-14-1) of the Korea National Arboretum, Korea.

Author contributions

JHM planned the projects, designed the research, analyzed data, and wrote the manuscript. HJ performed the experiments, assembled the genomes, analyzed data, and participated in manuscript preparation. HK, SMH, and HJY performed the transcriptome, methylome, and repeat analyses. AC performed the comparative genome analysis. HJK, DKK, and JHK planned the projects and participated in manuscript preparation. JJ and JHK performed the molecular dating.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-025-07683-y>.

Correspondence and requests for materials should be addressed to Jeong-Hwan Mun.

Peer review information *Communications Biology* thanks Yong Li, Laxman Adhikari and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: David Favero.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025