Article

The water lily genome and the early evolution of flowering plants

https://doi.org/10.1038/s41586-019-1852-5

Received: 6 August 2019

Accepted: 31 October 2019

Published online: 18 December 2019

Open access

Liangsheng Zhang[1,23,24]*, Fei Chen[1,2,23,24], Xingtan Zhang[1,23], Zhen Li[3,4,23], Yiyong Zhao[5,6,23], Rolf Lohaus[3,4,23], Xiaojun Chang[1,7,23], Wei Dong[1], Simon Y. W. Ho[8], Xing Liu[1], Aixia Song[1], Junhao Chen[9], Wenlei Guo[9], Zhengjia Wang[9], Yingyu Zhuang[1], Haifeng Wang[1], Xuequn Chen[1], Juan Hu[1], Yanhui Liu[1], Yuan Qin[1], Kai Wang[1], Shanshan Dong[7], Yang Liu[7,10], Shouzhou Zhang[7], Xianxian Yu[11], Qian Wu[12,13], Liangsheng Wang[12,13], Xueqing Yan[13,14], Yuannian Jiao[13,14], Hongzhi Kong[13,14], Xiaofan Zhou[15], Cuiwei Yu[16], Yuchu Chen[16], Fan Li[17], Jihua Wang[17], Wei Chen[18], Xinlu Chen[19], Qidong Jia[20], Chi Zhang[19], Yifan Jiang[2], Wanbo Zhang[2], Guanhua Liu[21], Jianyu Fu[21], Feng Chen[2,19,20,24], Hong Ma[6,24], Yves Van de Peer[3,4,22,24] & Haibao Tang[1,24]

Water lilies belong to the angiosperm order Nymphaeales. Amborellales, Nymphaeales and Austrobaileyales together form the so-called ANA-grade of angiosperms, which are extant representatives of lineages that diverged the earliest from the lineage leading to the extant mesangiosperms[1–3]. Here we report the 409-megabase genome sequence of the blue-petal water lily (*Nymphaea colorata*). Our phylogenomic analyses support Amborellales and Nymphaeales as successive sister lineages to all other extant angiosperms. The *N. colorata* genome and 19 other water lily transcriptomes reveal a Nymphaealean whole-genome duplication event, which is shared by Nymphaeaceae and possibly Cabombaceae. Among the genes retained from this whole-genome duplication are homologues of genes that regulate flowering transition and flower development. The broad expression of homologues of floral ABCE genes in *N. colorata* might support a similarly broadly active ancestral ABCE model of floral organ determination in early angiosperms. Water lilies have evolved attractive floral scents and colours, which are features shared with mesangiosperms, and we identified their putative biosynthetic genes in *N. colorata*. The chemical compounds and biosynthetic genes behind floral scents suggest that they have evolved in parallel to those in mesangiosperms. Because of its unique phylogenetic position, the *N. colorata* genome sheds light on the early evolution of angiosperms.

Many water lily species, particularly from *Nymphaea* (Nymphaeaceae), have large and showy flowers and belong to the angiosperms (also called flowering plants). Their aesthetic beauty has captivated notable artists such as the French impressionist Claude Monet. Water lily flowers have limited differentiation in perianths (outer floral organs), but they possess both male and female organs and have diverse scents and colours, similar to many mesangiosperms (core angiosperms, including eudicots, monocots, and magnoliids) (Supplementary Note 1). In addition, some water lilies have short life cycles and enormous numbers of seeds[4], which increase their potential as a model plant to represent the ANA-grade of angiosperms and to study early evolutionary events within the angiosperms. In particular, *N. colorata* Peter has a relatively small genome size ($2n = 28$ and approximately 400 Mb) and blue petals that make it popular in breeding programs (Supplementary Note 1).

We report here the genome sequence of *N. colorata*, obtained using PacBio RSII single-molecule real-time (SMRT) sequencing technology. The genome was assembled into 1,429 contigs (with a contig N50 of 2.1 Mb) and total length of 409 Mb with 804 scaffolds, 770 of which were anchored onto 14 pseudo-chromosomes (Extended Data Fig. 1 and Extended Data Table 1). Genome completeness was estimated to be 94.4% (Supplementary Note 2). We annotated 31,580 protein-coding genes and predicted repetitive elements with a collective length of 160.4 Mb, accounting for 39.2% of the genome (Supplementary Note 3).

The *N. colorata* genome provides an opportunity to resolve the relationships between Amborellales, Nymphaeales and all other extant angiosperms (Fig. 1a). Using six eudicots, six monocots, *N. colorata* and *Amborella*[5], and each of three gymnosperm species (*Ginkgo biloba*, *Picea abies* and *Pinus taeda*) as an outgroup in turn, we identified 2,169, 1,535 and 1,515 orthologous low-copy nuclear (LCN) genes, respectively (Fig. 1b). Among the LCN gene trees inferred from nucleotide sequences using *G. biloba* as an outgroup, 62% (294 out of 475 trees) place *Amborella* as the sister lineage to all other extant angiosperms with bootstrap support greater than 80% (type II, Fig. 1c). Using *P. abies* or *P. taeda* as the outgroup, *Amborella* is placed as the sister lineage to the remaining angiosperms in 57% and 54% of the LCN gene trees, respectively. LCN gene trees inferred using amino acid sequences show similar phylogenetic patterns (Supplementary Note 4.1).

A list of affiliations appears at the end of the paper.

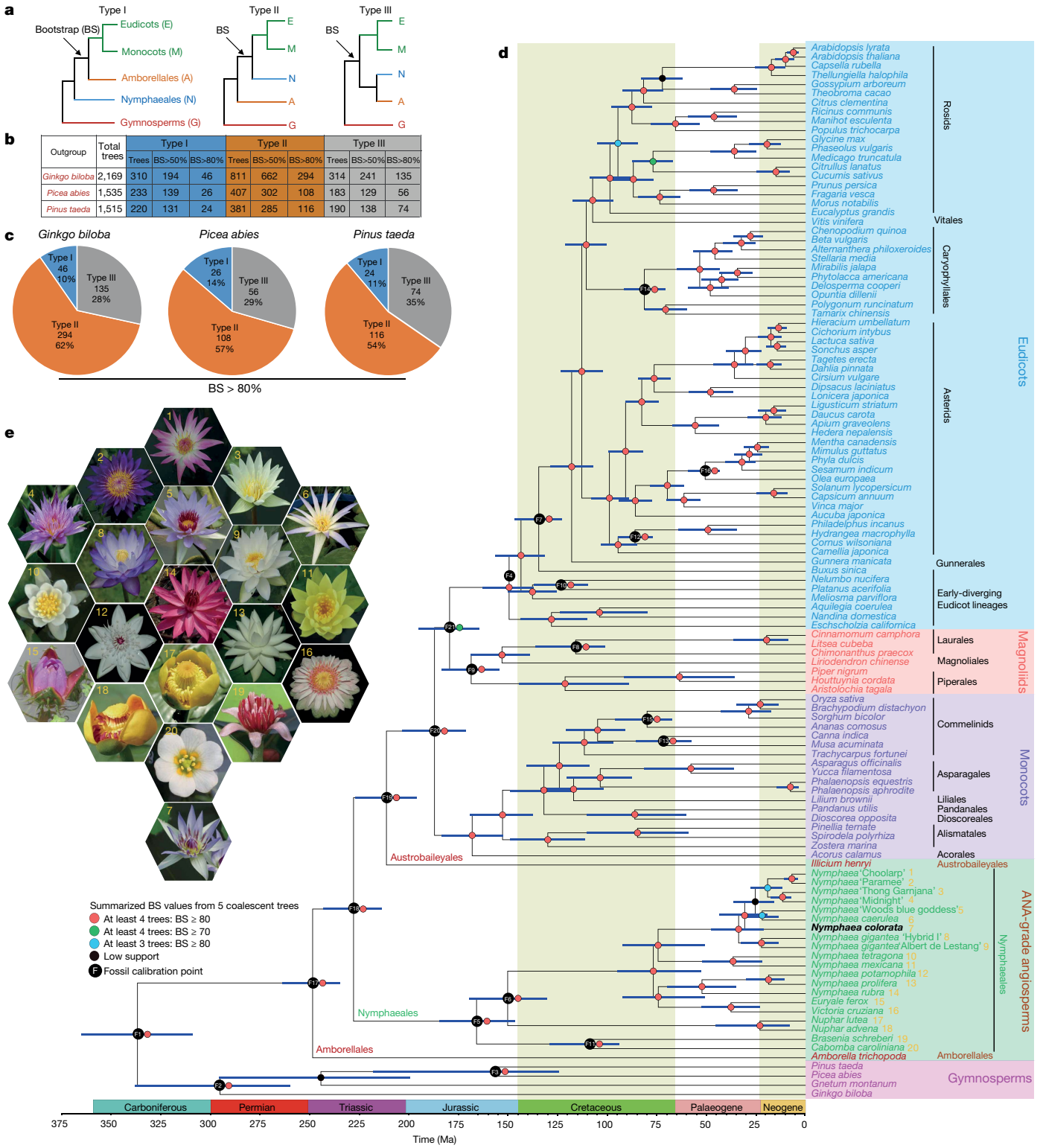


Fig. 1 | Phylogenomic relationships of angiosperms. a, Three different evolutionary relationships among major clades of angiosperms. **b**, Number of LCN gene trees with different bootstrap support (BS) values based on nucleotide sequences from six eudicots, six monocots, *N. colorata*, *Amborella* and three different gymnosperms. **c**, Comparison of gene trees supporting the three evolutionary relationships using each gymnosperm in turn as the outgroup. The percentage was calculated by dividing the number of type I, II or III trees (BS > 80%) by the total number of trees. **d**, Summary phylogeny and timescale of 115 plant species. Blue bars at nodes represent 95% credibility intervals of the estimated dates. **e**, The flowers of the 20 sampled water lilies in Nymphaeales used in **d**.

To minimize the potential shortcomings of sparse taxon sampling[6], we also inferred an angiosperm species tree using sequences from 44 genomes and 71 transcriptomes, including representatives of the ANA-grade, eudicots, magnoliids, monocots and a gymnosperm outgroup (*Gnetum montanum*, *G. biloba*, *P. abies* and *P. taeda*) (Methods). For further phylogenetic inference of these 115 species, we selected,

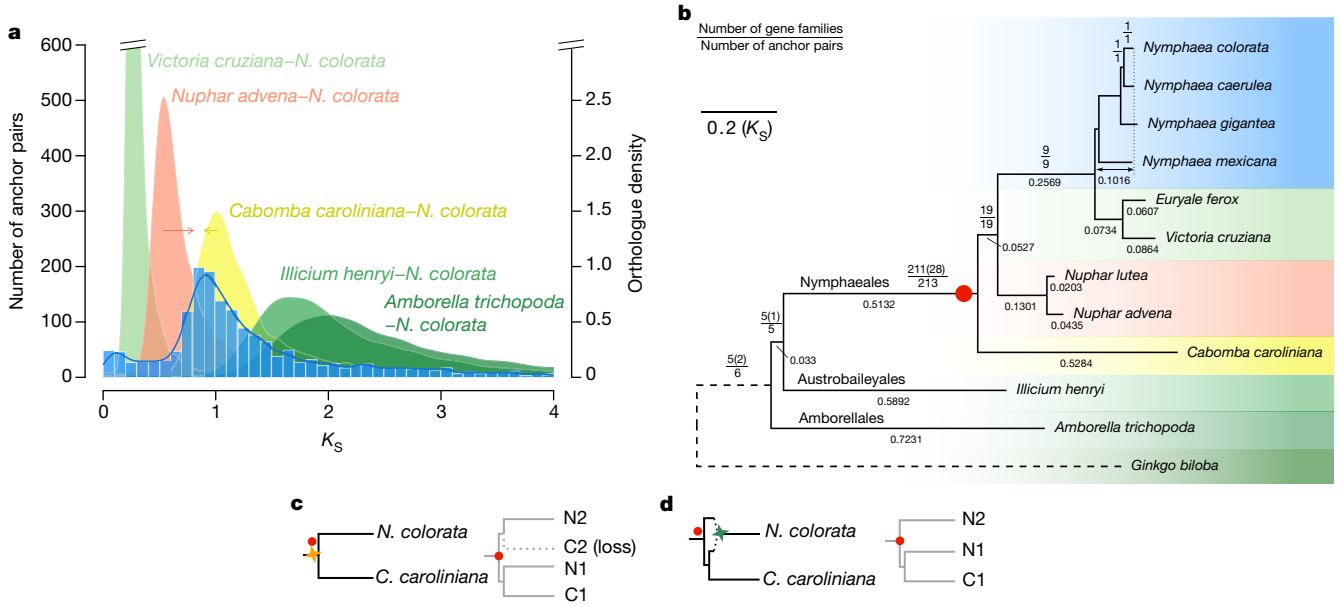


Fig. 2 | A Nymphaealean WGD shared by Nymphaeaceae and possibly Cabombaceae. a, K_S age distributions for paralogues found in collinear regions (anchor pairs) of *N. colorata* and for orthologues between *N. colorata* and selected Nymphaealean and angiosperm species. Red and yellow arrows indicate under- and overestimations of the *N. colorata*–*Nuphar advena* and *N. colorata*–*C. caroliniana* divergence, respectively. **b**, WGD phylogenomic analysis. Numbers in parentheses are the number of gene families with retained *C. caroliniana* duplicates supporting the duplication events. Numbers below branches show branch lengths in K_S units. The double-arrowed line denotes total K_S from the pointed node to *N. colorata*. We used *G. biloba* (dashed branch) as an outgroup. The red dot denotes the branch on which most of the anchor pairs in *N. colorata* coalesced. All mapped duplication events have BS ≥ 80% in the gene trees. **c**, Left, the scenario of a WGD (yellow four-pointed star) before the divergence between Nymphaeaceae and Cabombaceae. Right, a possible gene tree under this scenario, with loss of one duplicate in *C. caroliniana*. Two red dots show where the anchor pair of *N. colorata* would coalesce. **d**, Left, scenario of a WGD in the stem lineage of Nymphaeaceae involving an allotetraploid (green four-pointed star) that formed between two ancestral parents after the divergence of the lineages leading to *N. colorata* and *C. caroliniana*, with one of the parents being more closely related to *C. caroliniana*. Right, a gene tree under such a scenario. Red dots are as in **c**.

based on various criteria, five different LCN gene sets including 1,167, 834, 683, 602 and 445 genes. Analyses of these five datasets all yielded similar tree topologies with *Amborella* and Nymphaeales as successive sister lineages to all other extant angiosperms (Fig. 1d, e, Supplementary Note 4.2).

Molecular dating of angiosperm lineages, using a stringent set of 101 LCN genes and with age calibrations based on 21 fossils[7], inferred the crown age of angiosperms at 234–263 million years ago (Ma) (Fig. 1d). The split between monocots and eudicots was estimated at 171–203 Ma and that between Nymphaeaceae and Cabombaceae at 147–185 Ma.

Genomic collinearity unveiled evidence of a whole-genome duplication (WGD) event in *N. colorata* (Extended Data Figs. 1f, 2a and Supplementary Note 5.1). The number of synonymous substitutions per synonymous site (K_S) distributions for *N. colorata* paralogues further showed a signature peak at K_S of approximately 0.9 (Fig. 2a) and peaks at similar K_S values were identified in other Nymphaeaceae species (Supplementary Note 5.2), which suggests an ancient single WGD event that is probably shared among Nymphaeaceae members. Comparison of the *N. colorata* paralogue K_S distribution with K_S distributions of orthologues (representing speciation events) between *N. colorata* and other Nymphaeales lineages, *Illicium henryi*, and *Amborella* suggests that the WGD occurred just after the divergence between Nymphaeaceae and Cabombaceae (Fig. 2a). By contrast, phylogenomic analyses of gene families that contained at least one paralogue pair from collinear regions of *N. colorata* suggest that the WGD is shared between Nymphaeaceae and Cabombaceae (Fig. 2b, Supplementary Note 5.4). If true, *Cabomba caroliniana* seems to have retained few duplicates (Fig. 2b, c), which would explain the absence of a clear peak in the *C. caroliniana* paralogue K_S distribution (Supplementary Note 5.2). Absolute dating of the paralogues of *N. colorata* does suggest that the WGD could have occurred before or close to the divergence between Nymphaeaceae and Cabombaceae (Extended Data Fig. 2d, Supplementary Note 5.3), considering the variable substitution rates among Nymphaealean lineages (Fig. 2a, b, Extended Data Fig. 2c). An alternative interpretation of the above results could be that the WGD signatures were from an allopolyploidy event that occurred between ancestral Nymphaeaceae and Cabombaceae lineages shortly after their divergence and that gave rise to the Nymphaeaceae (but not Cabombaceae) stem lineage (Fig. 2d, Supplementary Note 5.4).

The water lily lineage descended from one of the early divergences among angiosperms, before the radiation of mesangiosperms. Thus, this group offers a unique window into the early evolution of angiosperms, particularly that of the flower. We identified 70 MADS-box genes, including homologues of the genes for the ABCE model of floral organ identities: *AP1* (and also *FUL*) and *AGL6* (A function for sepals and petals), *AP3* and *PI* (B function for petals and stamen), *AG* (C function for stamen and carpel), and *SEP1* (E function for interacting with ABC function proteins). Phylogenetic and collinearity analyses of the MADS-box genes and their genomic neighbourhood indicate that an ancient tandem duplication before the divergence of seed plants gave birth to the ancestors of A function (*FUL*) and E function genes (*SEP*) (Extended Data Fig. 3, Supplementary Note 6.1). Also, owing to the Nymphaealean WGD, *N. colorata* has two paralogues, *AGa* and *AGb* of the C-function gene *AG* (Extended Data Fig. 4). Similarly, the Nymphaealean WGD-derived duplicates are homologous to other genes associated with development of carpel and stamen[8], and to genes that regulate flowering time[9] and auxin-controlled circadian opening and closure of the flower[10] (Extended Data Figs. 4–6, Supplementary Note 6.2–6.4).

The expression profiles of *N. colorata* ABCE homologues largely agree with their putative ascribed roles in floral organ patterning (Fig. 3a).

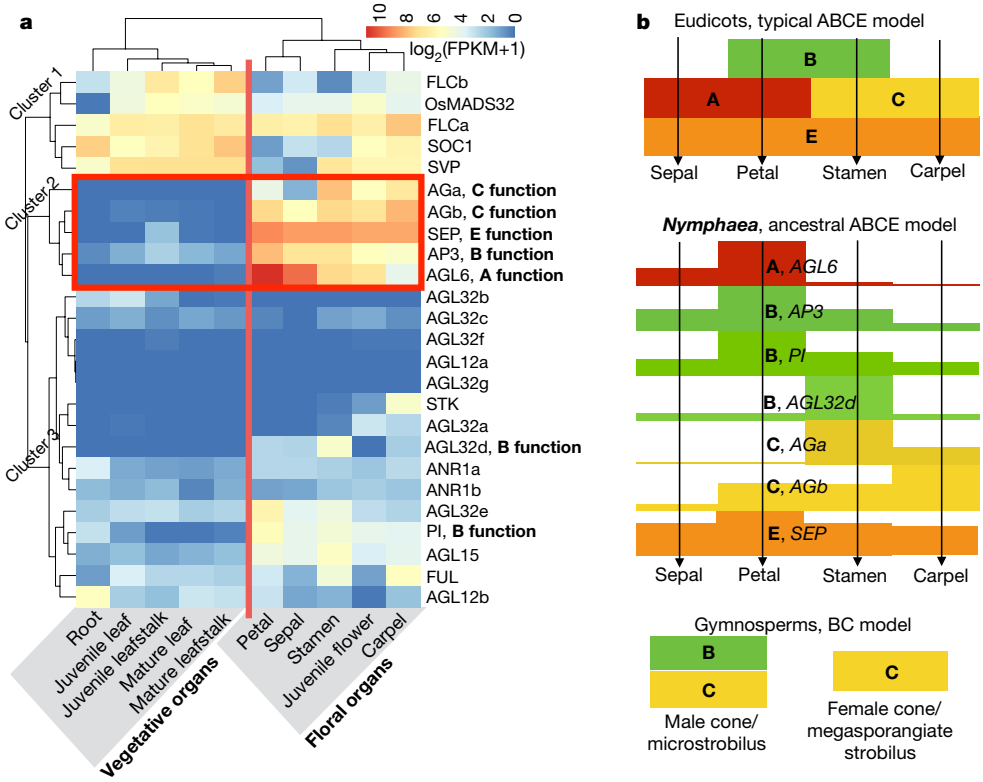


Fig. 3 | MADS-box genes in *N. colorata* and proposed floral ABCE model in early angiosperms. a, Gene expression patterns of MIKC^c from various organs of *N. colorata*. Three clusters of genes were classified according to the expression of type II MADS-box genes. The organ types (vegetative organs and floral organs) were matched to the expression patterns of type II MADS-box genes. Expression values were scaled by $\log_2$(FPKM + 1), in which FPKM is fragments per kilobase of exon per million mapped reads. **b**, The flowering ABCE model in *N. colorata* that specifies floral organs is proposed based on the gene expression values (bar heights) from **a**.

Notably, the *N. colorata AGL6* homologue is mainly expressed in sepals and petals, whereas the *FUL* homologue is mainly expressed in carpels, suggesting that *AGL6* acts as an A-function gene in *N. colorata*. The two C-function homologues *AGa* and *AGb* are highly expressed in stamens and carpels, respectively, whereas *AGb* is also expressed in sepals and petals, suggesting that they might have undergone subfunctionalization and possibly neofunctionalization for flower development after the Nymphaealean WGD. Furthermore, the ABCE homologues in *N. colorata* generally exhibit wider ranges of expression in floral organs than their counterparts in eudicot model systems (Fig. 3b). This wider expression pattern, in combination with broader expression of at least some ABCE genes in some eudicots representing an early-diverging lineage[11], some monocots[12] and magnoliids[13], suggest an ancient ABCE model for flower development, with subsequent canalization of gene expression and function regulated by the more specialized ABCE genes during the evolution of mesangiosperms, especially core eudicots[8]. This could also account for the limited differentiation between sepals and petals in Nymphaeales species, and is consistent with a single type of perianth organ proposed in an ancestral angiosperm flower[14].

Floral scent serves as olfactory cues for insect pollinators[15]. Whereas *Amborella* flowers are scentless[16], *N. colorata* flowers release 11 different volatile compounds, including terpenoids (sesquiterpenes), fatty-acid derivatives (methyl decanoate) and benzenoids (Fig. 4a). The *N. colorata* genome contains 92 putative terpene synthase (*TPS*) genes, which are ascribed to four previously recognized TPS subfamilies in angiosperms: TPS-b, TPS-c, TPS-e/f and TPS-g (Fig. 4b), but none was found for TPS-a, which is responsible for sesquiterpene biosynthesis in mesangiosperms[17]. Notably, TPS-b contains more than 80 genes in *N. colorata*; NC11G0123420 is highly expressed in flowers (Extended Data Fig. 7); this result suggests that it may be a candidate gene for sesquiterpene biosynthase in *N. colorata*. Also, methyl decanoate has not been detected as a volatile compound in monocots and eudicots[18] and is thought to be synthesized in *N. colorata* by the SABATH family of methyltransferases[19]. The *N. colorata* genome contains 13 *SABATH* homologues and 12 of them form a Nymphaeales-specific group (Supplementary Fig. 41). Among these 12 members, NC11G0120830 showed the highest expression in petals (Fig. 4c) and its corresponding recombinant protein was demonstrated to be a fatty acid methyltransferase that had the highest activity with decanoic acid as the substrate (Fig. 4d, Supplementary Note 7.1). These results suggest that the floral scent biosynthesis in *N. colorata* has been accomplished through enzymatic functions that have evolved independently from those in mesangiosperms (Fig. 4e).

Nymphaea colorata is valued for the aesthetically attractive blue colour of petals, which is a rare trait in ornamentals. To understand the molecular basis of the blue colour, we identified delphinidin 3′-*O*-(2″-*O*-galloyl-6″-*O*-acetyl-β-galactopyranoside) as the main blue anthocyanidin pigment (Extended Data Fig. 8a–c). By comparing the expression profiles between two *N. colorata* cultivars with white and blue petals for genes in a reconstructed anthocyanidin biosynthesis pathway, we found genes for an anthocyanidin synthase and a delphinidin-modification enzyme, the expression of which was significantly higher in blue petals than in white petals (Extended Data Fig. 8d, e). These two enzymes catalyse the last two steps of anthocyanidin biosynthesis and are therefore key enzymes specialized in blue pigment biosynthesis[20,21] (Supplementary Note 7.2).

Water lilies have a global distribution that includes cold regions (northern China and northern Canada), unlike the other ANA-grade angiosperms *Amborella* (Pacific Islands) and Austrobaileyales (temperate and tropical regions). We detected marked expansions of genes

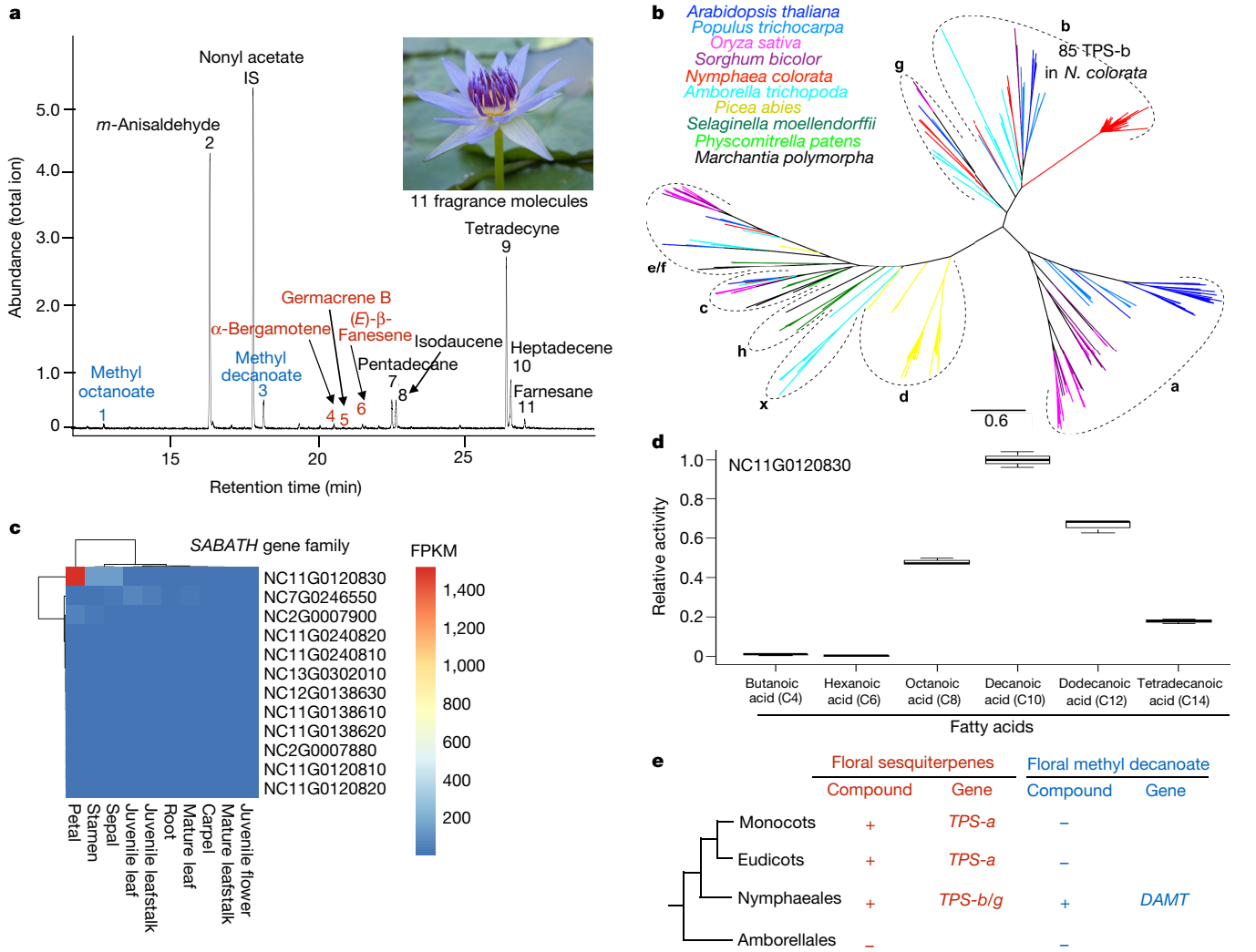


Fig. 4 | Floral scent and biosynthesis in *N. colorata*. **a**, Gas chromatogram of floral volatiles from the flower of *N. colorata*. The internal standard (IS) is nonyl acetate. Methyl esters are in blue; terpenes are in red. Floral scent was measured three times independently with similar results. **b**, Phylogenetic tree of terpene synthases from *N. colorata* and representative plants showing the subfamilies from a–h and x. **c**, Expression analysis of *SABATH* genes of *N. colorata* showed that NC11G0120830 had the highest expression level in petal. **d**, Relative activity of *Escherichia coli*-expressed NC11G0120830 with six fatty acids as substrates, with the activity on decanoic acid set at 1.0. Data are mean ± s.d. of three independent measurements. **e**, The presence (+) and absence (–) of sesquiterpenes and methyl decanoate as floral scent compounds and their respective biosynthetic genes in four major lineages of angiosperms when known. *DAMT*, decanoic acid methyltranferase.

related to immunity and stress responses in *N. colorata*, including genes encoding nucleotide-binding leucine-rich repeat (NLR) proteins, protein kinases and WRKY transcription factors, compared with those in *Amborella* and some mesangiosperms (Extended Data Fig. 9, Supplementary Note 8). It is possible that increased numbers of these genes enabled water lilies to adapt to various ecological habitats globally.

In conclusion, the *N. colorata* genome offers a reference for comparative genomics and for resolving the deep phylogenetic relationships among the ANA-grade and mesangiosperms. It has also revealed a WGD specific to Nymphaeales, and provides insights into the early evolution of angiosperms on key innovations such as flower development and floral scent and colour.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1852-5.

1. Byng, J. W. et al. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* **181**, 1–20 (2016).
2. Zeng, L. et al. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nat. Commun.* **5**, 4956 (2014).
3. Qiu, Y. L. et al. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* **402**, 404–407 (1999).
4. Chen, F. et al. Water lilies as emerging models for Darwin's abominable mystery. *Hortic. Res.* **4**, 17051 (2017).
5. Amborella Genome Project. The *Amborella* genome and the evolution of flowering plants. *Science* **342**, 1241089 (2013).
6. Wiens, J. J. Missing data, incomplete taxa, and phylogenetic accuracy. *Syst. Biol.* **52**, 528–538 (2003).
7. Coiro, M., Doyle, J. A. & Hilton, J. How deep is the conflict between molecular and fossil evidence on the age of angiosperms? *New Phytol.* **223**, 83–99 (2019).
8. Alvarez-Buylla, E. R. et al. Flower development. *Arabidopsis Book* **8**, e0127 (2010).
9. Zhao, N. et al. Identification of flowering regulatory genes in allopolyploid *Brassica juncea*. *Hortic. Plant J.* **5**, 109–119 (2019).
10. Ke, M. et al. Auxin controls circadian flower opening and closure in the waterlily. *BMC Plant Biol.* **18**, 143 (2018).
11. Sharma, B. & Kramer, E. M. *Aquilegia* B gene homologs promote petaloidy of the sepals and maintenance of the C domain boundary. *Evodevo* **8**, 22 (2017).
12. Dodsworth, S. Petal, sepal, or tepal? B-genes and monocot flowers. *Trends Plant Sci.* **22**, 8–10 (2017).
13. Chanderbali, A. S. et al. Conservation and canalization of gene expression during angiosperm diversification accompany the origin and evolution of the flower. *Proc. Natl Acad. Sci. USA* **107**, 22570–22575 (2010).

14. Sauquet, H. et al. The ancestral flower of angiosperms and its early diversification. *Nat. Commun.* **8**, 16047 (2017).
15. Kessler, D. et al. How scent and nectar influence floral antagonists and mutualists. *eLife* **4**, e07641 (2015).
16. Thien, L. B. et al. The population structure and floral biology of *Amborella trichopoda* (Amborellaceae). *Ann. Mo. Bot. Gard.* **90**, 466–490 (2003).
17. Chen, F., Tholl, D., Bohlmann, J. & Pichersky, E. The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* **66**, 212–229 (2011).
18. Knudsen, J. T., Tollsten, L. & Bergstrom, L. G. Floral scents-a checklist of volatile compounds isolated by head-space techniques. *Phytochemistry* **33**, 253–280 (1993).
19. Zhao, N. et al. Structural, biochemical, and phylogenetic analyses suggest that indole-3-acetic acid methyltransferase is an evolutionarily ancient member of the SABATH family. *Plant Physiol.* **146**, 455–467 (2008).
20. Chen, W. H. et al. Downregulation of putative UDP-glucose: flavonoid 3-*O*-glucosyltransferase gene alters flower coloring in *Phalaenopsis*. *Plant Cell Rep.* **30**, 1007–1017 (2011).
21. Wu, Q. et al. Transcriptome sequencing and metabolite analysis for revealing the blue flower formation in waterlily. *BMC Genomics* **17**, 897 (2016).

[1]Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Key Laboratory of Ministry of Education for Genetics, Breeding and Multiple Utilization of Crops, Key Laboratory of National Forestry and Grassland Administration for Orchid Conservation and Utilization, Fujian Agriculture and Forestry University, Fuzhou, China. [2]College of Horticulture, Nanjing Agricultural University, Nanjing, China. [3]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium. [4]VIB Center for Plant Systems Biology, Ghent, Belgium. [5]State Key Laboratory of Genetic Engineering, Ministry of Education Key Laboratory of Biodiversity Sciences and Ecological Engineering, School of Life Sciences, Fudan University, Shanghai, China. [6]Department of Biology, Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA, USA. [7]Fairy Lake Botanical Garden, Shenzhen and Chinese Academy of Sciences, Shenzhen, China. [8]School of Life and Environmental Sciences, University of Sydney, Sydney, New South Wales, Australia. [9]State Key Laboratory of Subtropical Silviculture, School of Forestry and Biotechnology, Zhejiang A&F University, Hangzhou, China. [10]BGI-Shenzhen, Shenzhen, China. [11]School of Urban-Rural Planning and Landscape Architecture, Xuchang University, Xuchang, China. [12]Key Laboratory of Plant Resources/Beijing Botanical Garden, Institute of Botany, Chinese Academy of Sciences, Beijing, China. [13]University of the Chinese Academy of Sciences, Beijing, China. [14]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China. [15]Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Integrative Microbiology Research Centre, South China Agricultural University, Guangzhou, China. [16]Hangzhou Tianjing Aquatic Botanical Garden, Zhejiang Humanities Landscape Co. Ltd., Hangzhou, China. [17]National Engineering Research Center for Ornamental Horticulture, Key Laboratory for Flower Breeding of Yunnan Province, Floriculture Research Institute, Yunnan Academy of Agricultural Sciences, Kunming, China. [18]Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu, China. [19]Department of Plant Sciences, University of Tennessee, Knoxville, TN, USA. [20]Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, TN, USA. [21]Key Laboratory of Tea Quality and Safety Control, Ministry of Agriculture and Rural Affairs, Tea Research Institute, Chinese Academy of Agricultural Sciences, Hangzhou, China. [22]Centre for Microbial Ecology and Genomics, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa. [23]These authors contributed equally: Liangsheng Zhang, Fei Chen, Xingtan Zhang, Zhen Li, Yiyong Zhao, Rolf Lohaus, Xiaojun Chang. [24]These authors jointly supervised this work: Liangsheng Zhang, Fei Chen, Feng Chen, Hong Ma, Yves Van de Peer, Haibao Tang. *e-mail: fafuzhang@163.com

Methods

Genome and transcriptome sequencing

Total DNA for genome sequencing was extracted from young leaves. Leaf RNA was extracted from 18 water lily species: *N. colorata*, *Euryale ferox*, *Brasenia schreberi*, *Victoria cruziana*, *Nymphaea mexicana*, *Nymphaea prolifera*, *Nymphaea tetragona*, *Nymphaea potamophila*, *Nymphaea caerulea*, *Nymphaea rubra*, *Nymphaea* 'midnight', *Nymphaea* 'Choolarp', *Nymphaea* 'Paramee', *Nymphaea* 'Woods blue goddess', *Nymphaea gigantea* 'Albert de Lestang', *N. gigantea* 'Hybrid I', *Nymphaea* 'Thong Garnjana' and *Nuphar lutea*. In addition, for transcriptome sequencing we sampled several organs and tissues from *N. colorata* including mature leaf, mature leafstalk, juvenile flower, juvenile leaf, juvenile leafstalk, carpel, stamen, sepal, petal and root.

For PacBio sequencing, we prepared approximately 20-kb SMRTbell libraries. A total of 34 SMRT cells and 49.8 Gb data composed of 5.5 million reads were sequenced on PacBio RSII system with P6-C4 chemistry. All transcriptome libraries were sequenced using the Illumina platform, generating paired-end reads. For the Hi-C sequencing and scaffolding, a Hi-C library was created from tender leaves of *N. colorata*. In brief, the leaves were fixed with formaldehyde and lysed, and the cross-linked DNA was then digested with MboI overnight. Sticky ends were biotinylated and proximity-ligated to form chimeric junctions, which were physically sheared to and enriched for sizes of 500–700 bp. Chimeric fragments representing the original cross-linked long-distance physical interactions were then processed into paired-end sequencing libraries and 346 million 150-bp paired-end reads, which were sequenced on the Illumina platform.

Sequence assembly and gene annotation

To assemble the 49.8 Gb data composed of 5.5 million reads, we filtered the reads to remove organellar DNA, reads of poor quality or short length, and chimaeras. The contig-level assembly was performed on full PacBio long reads using the Canu package[22]. Canu v.1.3 was used for self-correction and assembly. We then polished the draft assembly using Arrow (https://github.com/PacificBiosciences/GenomicConsensus). To increase the accuracy of the assembly, Illumina short reads were recruited for further polishing with the Pilon program (https://github.com/broadinstitute/pilon). The genome assembly quality was measured using BUSCO (Benchmarking Universal Single-Copy Orthologues)[23] v.3.0. The paired-end reads from Hi-C were uniquely mapped onto the draft assembly contigs, which were grouped into chromosomes and scaffolded using the software Lachesis (https://github.com/shendurelab/LACHESIS).

Genscan (http://genes.mit.edu/GENSCAN.html) and Augustus[24] were used to carry out de novo predictions with gene model parameters trained from *Arabidopsis thaliana*. Furthermore, gene models were de novo predicted using MAKER[25]. We then evaluated the genes by comparing MAKER results with the corresponding transcript evidence to select gene models that were the most consistent on the basis of an AED metric.

The evolutionary position of water lily and divergence-time estimation

LCN genes were identified based on OrthoFinder[26] results. The orthologues were obtained from six monocots (*Spirodela polyrhiza*, *Zostera marina*, *Musa acuminata*, *Ananas comosus*, *Sorghum bicolor* and *Oryza sativa*) and six eudicots (*Nelumbo nucifera*, *Vitis vinifera*, *Populus trichocarpa*, *A. thaliana*, *Solanum lycopersicum* and *Beta vulgaris*), *N. colorata*, *Amborella*, and the gymnosperms *G. biloba*, *P. abies* and *P. taeda*. LCN genes needed to meet the following requirements: strictly single-copy in *N. colorata*, *Amborella*, *G. biloba*, *P. abies* or *P. taeda*, and single-copy in at least five of the 12 eudicots or monocots. With *G. biloba*, *P. abies* or *P. taeda* as the outgroup, we identified 2,169, 1,535 and 1,515 orthologous LCN genes, respectively. Furthermore, we trimmed the sites with less than 90% coverage. LCN gene trees were estimated from the remaining sites using RAxML v.7.7.8 using the GTR+G+I model for nucleotide sequences (Fig. 1c) and the JTT+G+I model for amino acid sequences (Supplementary Note 4.1). To account for incomplete lineage sorting and different substitution rates, we applied the multispecies coalescent model and a supermatrix method, respectively, to the LCN genes and found further support for the sister relationship between *Amborella* and all other extant flowering plants (Supplementary Note 4.2).

We further carefully selected five LCN gene sets (1,167, 834, 683, 602 and 445) from 115 species and applied both a supermatrix method[27–29] and the multi-species coalescent model to infer the phylogeny of angiosperms (Supplementary Note 4.2). The phylogeny inferred from 1,167 LCN genes is shown in Fig. 1d, with different support values from the multi-species coalescent analyses of the other four LCN gene sets.

To estimate the evolutionary timescale of angiosperms, we calibrated a relaxed molecular clock using 21 fossil-based age constraints[7] throughout the tree, including the earliest fossil tricoplate pollen (approximately 125 Ma) associated with eudicots[30]. We concatenated 101 selected genes (205,185 sites) and fixed the tree topology to that inferred from our coalescent-based analysis of 1,167 genes from 115 taxa. We performed a Bayesian phylogenomic dating analysis of the 101 selected genes in MCMCtree, part of the PAML package[31,32], and used approximate likelihood calculation for the branch lengths[33]. Molecular dating was performed using an auto-correlated model of among-lineage rate variation, the GTR substitution model, and a uniform prior on the relative node times. Posterior distributions of node ages were estimated using Markov chain Monte Carlo sampling, with samples drawn every 250 steps over 10 million steps following a burn-in of 500,000 steps. We checked for convergence by running the analysis in duplicate and checked for sufficient sampling.

We also implemented the penalized likelihood method under a variable substitution rate using TreePL[34] and r8s[35], as a constant substitution rate across the phylogenetic tree was rejected ($P < 0.01$) for all cases by likelihood-ratio tests in PAUP[36]. Three fossil calibrations, corresponding to the crown groups of Lamiales, Cornales and Laurales, were implemented as minimum age constraints in our penalized likelihood dating analysis, except that the earliest appearance of tricolpate pollen grains (about 125 Ma)[30] was used to fix the age of crown eudicots. We determined the best smoothing parameter value of the concatenated 101 LCN genes as 0.32 by performing cross-validations of a range of smooth parameters from 0.01 to 10,000 (algorithm = TN; crossv = yes; cvstart = −2; cvinc = 0.5; cvnum = 15). We used 100 bootstrap trees with branch lengths generated by RAxML[37] to infer the 95% confidence intervals of age estimates (Supplementary Note 4.2).

Identification of WGD

The *N. colorata* genome was compared with each of the other genomes by pairwise alignment using Large-Scale Genome Alignment Tool (LAST; http://last.cbrc.jp/). We defined syntenic blocks using LAST hits with a distance cut-off of 20 genes apart from the two retained homologous pairs, in which at least four consecutive retained homologous pairs were required. We then obtained the one-to-one blocks to exclude ancient duplication blocks with QUOTA-ALIGN[38].

K_S-based paralogue age distributions were constructed as previously described[39]. In brief, the paranome was constructed by performing an all-against-all protein sequence similarity search using BLASTP with an *E*-value cut-off of 10^{-10}, after which gene families were built with the mclblastline pipeline (v.10-201) (micans.org/mcl). Each gene family was aligned using MUSCLE (v.3.8.31)[40], and K_S estimates for all pairwise comparisons within a gene family were obtained using maximum likelihood in the CODEML program[41] of the PAML package (v.4.4c)[31]. We then subdivided gene families into subfamilies for which K_S estimates between members did not exceed a value of 5.

To correct for the redundancy of K_S values (a gene family of n members produces $n(n-1)/2$ pairwise K_S estimates for $n-1$ retained duplication events), we inferred a phylogenetic tree for each subfamily using PhyML[42] with the default settings. For each duplication node in the resulting phylogenetic tree, all m K_S estimates between the two child clades were added to the K_S distribution with a weight of $1/m$ (in which m is the number of K_S estimates for a duplication event), so that the weights of all K_S estimates for a single duplication event summed to one. Paralogous gene pairs found in duplicated collinear segments (anchor pairs) from *N. colorata* were detected using i-ADHoRe (v.3.0) with 'level_2_only = TRUE'[43,44]. The identified anchor pairs are assumed to correspond to the most recent WGD event.

The K_S-based orthologue age distributions were constructed by identifying one-to-one orthologues between species using InParanoid[45] with default settings, followed by K_S estimation using the CODEML program as above. K_S distributions for one-to-one orthologues between *N. colorata* and each of *V. cruziana*, *N. advena*, *C. caroliniana*, *I. henryi* and *Amborella* were used to compare the relative timing of the WGD in *N. colorata* with speciation events within Nymphaeales. K_S distributions for one-to-one orthologues between the outgroup species *I. henryi* and each of *N. lutea*, *N. advena*, *N. mexicana*, *Nymphaea* 'Woods blue goddess', *N. colorata*, and *C. caroliniana* were used to estimate and compare relative substitution rates among these Nymphaealean species. Additional comparisons using *V. vinifera* and *Amborella* as outgroup species instead of *I. henryi* gave similar results (data not shown).

Absolute dating of the identified WGD event in *N. colorata* was performed as previously described[46]. Briefly, paralogous gene pairs located in duplicated segments (anchor pairs) and duplicated pairs lying under the WGD peak (peak-based duplicates) were collected for phylogenetic dating. We selected anchor pairs and peak-based duplicates present under the *N. colorata* WGD peak and with K_S values between 0.7 and 1.2 (grey-shaded area in Extended Data Fig. 2b) for absolute dating. For each WGD paralogous pair, an orthogroup was created that included the two paralogues plus several orthologues from other plant species as identified by InParanoid[45] using a broad taxonomic sampling: one representative orthologue from the order Cucurbitales, two from Rosales, two from Fabales, two from Malpighiales, two from Brassicales, one from Malvales, one from Solanales, two from Poaceae (Poales), one from *A. comosus*[47] (Bromeliaceae, Poales), one from either *M. acuminata*[48] (Zingiberales) or *Phoenix dactylifera*[49] (Arecales), one from the Asparagales (from *Asparagus officinalis*[50], *Apostasia shenzhenica*[46], or *Phalaenopsis equestris*[51]), one from the Alismatales (either from *S. polyrhiza*[52] or *Z. marina*[53]), one from *Amborella*, and one from *G. biloba*[54]. In total, 217 orthogroups based on anchor pairs and 142 orthogroups based on peak-based duplicates were collected.

The node joining the two WGD paralogues of *N. colorata* was then dated using the BEAST v1.7 package[55] under an uncorrelated relaxed-clock model and an LG+G model with four site-rate categories. A starting tree with branch lengths satisfying all fossil prior constraints was created according to the consensus APG IV phylogeny[1]. Fossil calibrations were implemented using log-normal calibration priors on the following nodes: the node uniting the Malvidae based on the fossil *Dressiantha bicarpellata*[56] with prior offset = 82.8, mean = 3.8528, and s.d. = 0.5[57]; the node uniting the Fabidae based on the fossil *Paleoclusia chevalieri*[58] with prior offset = 82.8, mean = 3.9314, and s.d. = 0.5[59]; the node uniting the non-Alismatalean monocots based on fossil *Liliacidites*[60] with prior offset = 93.0, mean = 3.5458, and s.d. = 0.5[61]; the node uniting the *N. colorata* WGD paralogues with the eudicots and monocots based on the sudden abundant appearance of eudicot tricolpate pollen in the fossil record with prior offset = 124, mean = 4.8143 and s.d. = 0.5[62]; and the root uniting the above clades with *Amborella* and then *G. biloba* with prior offset = 307, mean = 3.8876, and s.d. = 0.5[63]. The offsets of these calibrations represent hard minimum boundaries, and their means represent locations for their respective peak mass probabilities in accordance with previous dating studies of these specific clades[63] (see Supplementary Note 5.3 for an alternative setting of orthogroups).

A run without data was performed to ensure proper placements of the marginal calibration priors, which do not necessarily correspond to the calibration priors specified above, because they interact with each other and the tree prior[64]. Indeed, a run without data indicated that the distribution of the marginal calibration prior for the root did not correspond to the specified calibration density, so we reduced the mean in the calibration prior of the node combining the *N. colorata* WGD paralogues with the eudicots and monocots with offset = 124, mean = 4.4397, s.d. = 0.5 to locate the marginal calibration prior at 220 Ma[62].

Markov chain Monte Carlo sampling for each orthogroup was run for 10 million steps, with sampling every 1,000 steps to produce a sample size of 10,000. The resulting trace files were inspected using Tracer v.1.5[55], with a burn-in of 1,000 samples, to check for convergence and sufficient sampling (minimum effective sample size of 200 for all parameters). In total, 263 orthogroups were accepted, and absolute age estimates of the node uniting the WGD paralogous pairs based on both anchor pairs and peak-based duplicates were grouped into one absolute age distribution, for which kernel density estimation and a bootstrapping procedure were used to find the peak consensus WGD age estimate and its 90% confidence interval boundaries, respectively. More detailed methods have been previously described[39].

To identify the duplication events that resulted in the 2,648 anchor pairs detected in the genome of *N. colorata*, we performed phylogenomic analyses to determine the timing of the duplication events relative to the lineage divergences in Nymphaeales as described previously[46]. Protein-coding genes from 12 species were used, including eight species from Nymphaeaceae and one species from Cabombaceae in Nymphaeales, one species (*I. henryi*) from Austrobaileyales, plus *Amborella* and *G. biloba*. The phylogeny of the 12 species was obtained from Fig. 1d, and the branch lengths in K_S units were estimated from 23 LCN genes (selected from the 101 LCN genes used in Fig. 1d, because only 23 are shared across all of the species studied) using PAML[31] under the free-ratio model. OrthoMCL (v.2.0.9)[65] was used with default parameters to identify gene families. Then, we removed 907 of the 2,648 anchor pairs with K_S values greater than five. If the remaining anchor pairs fell into different gene families, thus indicating incorrect assignment of gene families by OrthoMCL, we merged the corresponding gene families and finally obtained 53,243 multi-gene gene families. Next, phylogenetic trees were constructed for a subset of 881 gene families with no more than 200 genes that had at least one pair of anchors and one gene from *G. biloba*. Multiple sequence alignments were produced by MUSCLE (v3.8.31)[40] and were trimmed by trimAl (v.1.4)[66] to remove low-quality regions based on a heuristic approach (-automated1).

We then used RAxML (v.8.2.0)[67] with the GTR+G model to estimate a maximum-likelihood tree, starting with 200 rapid bootstraps followed by maximum-likelihood optimizations on every fifth bootstrap tree. Gene trees were rooted based on genes from *G. biloba* if these formed a monophyletic group in the tree; otherwise, mid-point rooting was applied. The timing of the duplication event for each anchor pair relative to the lineage divergence events was then inferred. In brief, internodes from a gene tree were first mapped to the species phylogeny according to the common ancestor of the genes in the gene tree. Each internode was then classified as a duplication node, a speciation node, or a node that has no paralogues and is inconsistent with divergence in the species phylogeny. The parental node(s) of a duplication node supported by an anchor pair were traced towards the root until reaching a speciation node in the gene tree. The duplication event that resulted in the anchor pair was hence circumscribed between the duplication node as the lower bound and the speciation node as the upper bound on the species tree. If the two nodes were directly connected by a single branch on the species tree, the duplication was thus considered to have occurred on the branch. To reduce biased estimations, we used

the bootstrap value on the branch leading to a duplication node as support for a duplication event. In total, 497 anchor pairs in 473 gene families coalesced as duplication events on the species phylogeny, and duplication events from 254 anchor pairs in 246 gene families (or from 380 anchor pairs in 364 gene families) had bootstrap values greater than or equal to 80% (or 50%).

Floral scent measurement, gene identification, and functional characterization

We collected floral volatiles of *N. colorata* using a dynamic headspace sampling system and analysed them using gas chromatography–mass spectrometry (GC–MS) as previously described[68]. After 2 h of collection from the headspace of detached open flowers of *N. colorata* in a glass chamber (10 cm diameter, 30 cm height), volatiles were eluted from the SuperQ volatile collection trap using 100 µl of methylene chloride containing nonyl acetate as an internal standard. We then analysed samples using an Agilent Intuvo 9000 GC system coupled with an Agilent 7000D Triple Quadrupole mass detector. Separation was performed on an Agilent HP 5 MS capillary column (30 m × 0.25 mm) with helium as carrier gas (flow rate of 1 ml min^{-1}). We applied splitless injections of 1 µl samples, injection temperature of 250 °C, an initial oven temperature of 40 °C (3-min hold) and a temperature gradient of 5 °C per min increase from 40 °C to 250 °C. Products were identified using the National Institute of Standards and Technology mass spectral database (https://chemdata.nist.gov).

A full-length cDNA of NC11G0120830 was amplified from the open flowers of *N. colorata* using reverse transcription PCR (RT–PCR), and cloned into pET-32a (MilliporeSigma). After confirmation by sequencing, NC11G0120830 was expressed in *E. coli* strain BL21 (DE3) (Stratagene) and the recombinant protein produced was purified using a modified nickel-nitrilotriacetic acid agarose (Invitrogen) protocol as previously reported[69]. For methyltransferase enzyme assays, we used both radiochemical and non-radiochemical reaction systems. The radiochemical reaction system (50 µl) was composed of 50 mM Tris-HCl, pH 7.8, 1 mM substrate, 1 µl ^{14}C-*S*-adenosyl-L-methionine, and 1 µl of purified NC11G0120830. After 30 min of incubation at room temperature, 150 µl of ethyl acetate was added to extract the ^{14}C-labelled reaction products. The extracts were counted using a scintillation counter (Beckman Coulter) to measure the activity of NC11G0120830. To determine the chemical identity of the reaction product, we performed non-radiochemical assays in which nonradioactive *S*-adenosyl-L-methionine was used as the methyl donor. The reaction product was collected by headspace solid-phase microextraction and analysed by GC–MS as previously described[70].

Reporting summary

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

Data availability

PacBio whole-genome sequencing data, Illumina data and genome assembly sequences have been deposited to the NCBI Sequence Read Archive (SRA) as Bioproject PRJNA565347, and were also deposited in the BIG Data Center (http://bigd.big.ac.cn) under project number PRJCA001283. The genome assembly sequences and gene annotations have been deposited in the Genome Warehouse in BIG Data Center under accession number GWHAAYW00000000. The genome assembly sequences, gene annotations, and the LCN genes used in this study, have been also deposited in the Waterlily Pond (http://waterlily.eplant.org). All other data are available from the corresponding author upon reasonable request.

22. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
23. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
24. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).
25. Cantarel, B. L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
26. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
27. Zeng, L. et al. Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. *New Phytol.* **214**, 1338–1354 (2017).
28. Xiang, Y. et al. Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Mol. Biol. Evol.* **34**, 262–281 (2017).
29. Huang, C. H. et al. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Mol. Biol. Evol.* **33**, 394–412 (2016).
30. Hickey, L. J. & Doyle, J. A. Early cretaceous fossil evidence for angisperm evolution. *Bot. Rev.* **43**, 3–104 (1977).
31. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
32. Morris, J. L. et al. The timescale of early land plant evolution. *Proc. Natl Acad. Sci. USA* **115**, E2274–E2283 (2018).
33. dos Reis, M. & Yang, Z. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Mol. Biol. Evol.* **28**, 2161–2172 (2011).
34. Smith, S. A. & O'Meara, B. C. treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* **28**, 2689–2690 (2012).
35. Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301–302 (2003).
36. Wilgenbusch, J. C. & Swofford, D. Inferring evolutionary trees with PAUP*. *Curr. Prot. Bioinformatics* **6**, 6.4.1–6.4.28 (2003).
37. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
38. Tang, H. et al. Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics* **12**, 102 (2011).
39. Vanneste, K., Van de Peer, Y. & Maere, S. Inference of genome duplications from age distributions revisited. *Mol. Biol. Evol.* **30**, 177–190 (2013).
40. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
41. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736 (1994).
42. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
43. Proost, S. et al. i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.* **40**, e11 (2012).
44. Fostier, J. et al. A greedy, graph-based algorithm for the alignment of multiple homologous gene lists. *Bioinformatics* **27**, 749–756 (2011).
45. Ostlund, G. et al. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* **38**, D196–D203 (2010).
46. Zhang, G. Q. et al. The *Apostasia* genome and the evolution of orchids. *Nature* **549**, 379–383 (2017).
47. Ming, R. et al. The pineapple genome and the evolution of CAM photosynthesis. *Nat. Genet.* **47**, 1435–1442 (2015).
48. D'Hont, A. et al. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488**, 213–217 (2012).
49. Al-Dous, E. K. et al. De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nat. Biotechnol.* **29**, 521–527 (2011).
50. Harkess, A. et al. The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nat. Commun.* **8**, 1279 (2017).
51. Cai, J. et al. The genome sequence of the orchid *Phalaenopsis equestris*. *Nat. Genet.* **47**, 65–72 (2015).
52. Wang, W. et al. The *Spirodela polyrhiza* genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nat. Commun.* **5**, 3311 (2014).
53. Olsen, J. L. et al. The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. *Nature* **530**, 331–335 (2016).
54. Guan, R. et al. Draft genome of the living fossil *Ginkgo biloba*. *Gigascience* **5**, 49 (2016).
55. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
56. Gandolfo, M., Nixon, K. & Crepet, W. A new fossil flower from the Turonian of New Jersey: *Dressiantha bicarpellata* gen. et sp. nov. (Capparales). *Am. J. Bot.* **85**, 964 (1998).
57. Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R. & Mathews, S. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **107**, 18724–18728 (2010).
58. Crepet, W. & Nixon, K. Fossil Clusiaceae from the late Cretaceous (Turonian) of New Jersey and implications regarding the history of bee pollination. *Am. J. Bot.* **85**, 1122–1133 (1998).
59. Xi, Z. et al. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proc. Natl Acad. Sci. USA* **109**, 17519–17524 (2012).
60. Ramírez, S. R., Gravendeel, B., Singer, R. B., Marshall, C. R. & Pierce, N. E. Dating the origin of the Orchidaceae from a fossil orchid with its pollinator. *Nature* **448**, 1042–1045 (2007).
61. Janssen, T. & Bremer, K. The age of major monocot groups inferred from 800+rbcL sequences. *Bot. J. Linn. Soc.* **146**, 385–398 (2004).
62. Smith, S. A., Beaulieu, J. M. & Donoghue, M. J. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proc. Natl Acad. Sci. USA* **107**, 5897–5902 (2010).

63. Clarke, J. T., Warnock, R. C. M. & Donoghue, P. C. J. Establishing a time-scale for plant evolution. *New Phytol*. **192**, 266–301 (2011).
64. Heled, J. & Drummond, A. J. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Syst. Biol*. **61**, 138–149 (2012).
65. Li, L., Stoeckert, C. J., Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. **13**, 2178–2189 (2003).
66. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
67. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
68. Li, G. et al. Nonseed plant *Selaginella moellendorffi* has both seed plant and microbial types of terpene synthases. *Proc. Natl Acad. Sci. USA* **109**, 14711–14715 (2012).
69. Zhao, N., Guan, J., Lin, H. & Chen, F. Molecular cloning and biochemical characterization of indole-3-acetic acid methyltransferase from poplar. *Phytochemistry* **68**, 1537–1544 (2007).
70. Zhao, N. et al. Molecular and biochemical characterization of the jasmonic acid methyltransferase gene from black cottonwood (*Populus trichocarpa*). *Phytochemistry* **94**, 74–81 (2013).

Acknowledgements Fei Chen acknowledges funding from National Natural Science Foundation of China (31801898). L.Z. is partly supported by the open funds of the State Key Laboratory of Crop Genetics and Germplasm Enhancement (ZW201909) and State Key Laboratory of Tree Genetics and Breeding (TGB2018004). H.T. thanks the Fujian provincial government in China for a Fujian "100 Talent Plan". Y.V.d.P. acknowledges funding from the European Union Seventh Framework Programme (FP7/2007-2013) under European Research Council Advanced Grant Agreement 322739 – DOUBLEUP. Z.L. is funded by a postdoctoral fellowship from the Special Research Fund of Ghent University (BOFPDO2018001701).

Author contributions L.Z. led and managed the project. L.Z., Fei Chen and H.M. conceived the study. Fei Chen, L.Z., H.M., Feng Chen, Z.L., R.L. and Y.V.d.P. wrote the manuscript; L.Z., Fei Chen, X. Yu, X. Chang, C.Y., Y.C., Q.W., L.W. and H.K. collected and sequenced the plant material; L.Z., Y. Zhao, Fei Chen, S.Y.W.H., J.C., H.W., Xuequn Chen, J.H., A.S., X. Chang, W.D., X.L., Y. Zhuang, Y. Jiao, W.C., X. Yan, Y.Q., K.W. and H.M. performed gene family clustering and comparative phylogenomics. X. Zhang, L.Z., X. Zhou and Fei Chen assembled and annotated the genome. S.D., S.Z. and Yang Liu annotated the mitochondrial and chloroplast genomes. Yanhui Liu, L.Z. and Fei Chen conducted transcriptome sequencing and analysis. Z.L., R.L., Y.V.d.P., W.G., Fei Chen, H.T. and L.Z. conducted WGD analysis. Feng Chen, Q.J., Xinlu Chen, C.Z., Y. Jiang, W.Z., G.L., J.F. and Fei Chen conducted floral scent analysis. L.Z., Fei Chen, Q.W., L.W., Z.W., F.L. and J.W. conducted floral colour analysis. All authors read and approved the manuscript.

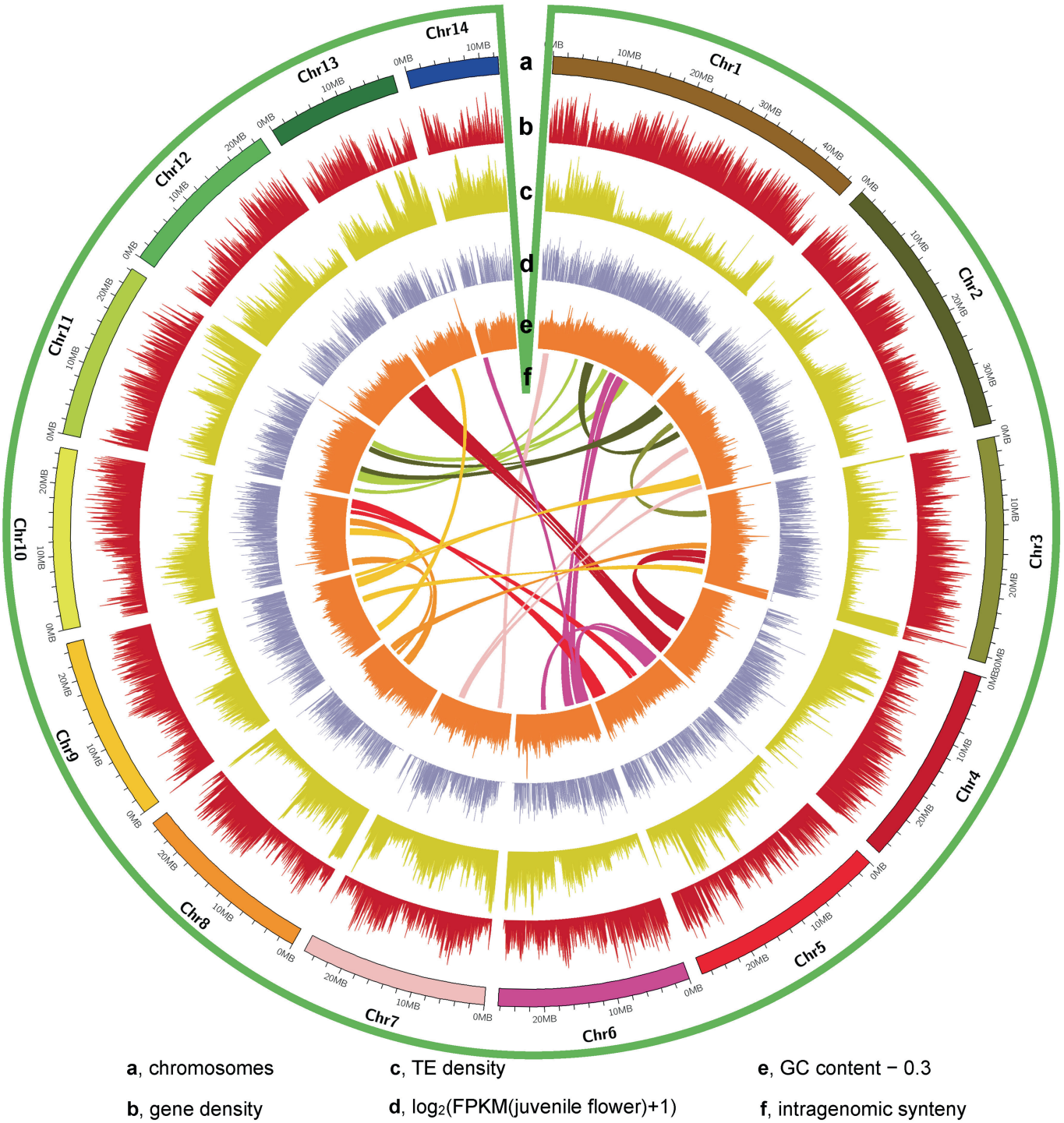
Competing interests The authors declare no competing interests.

Additional information

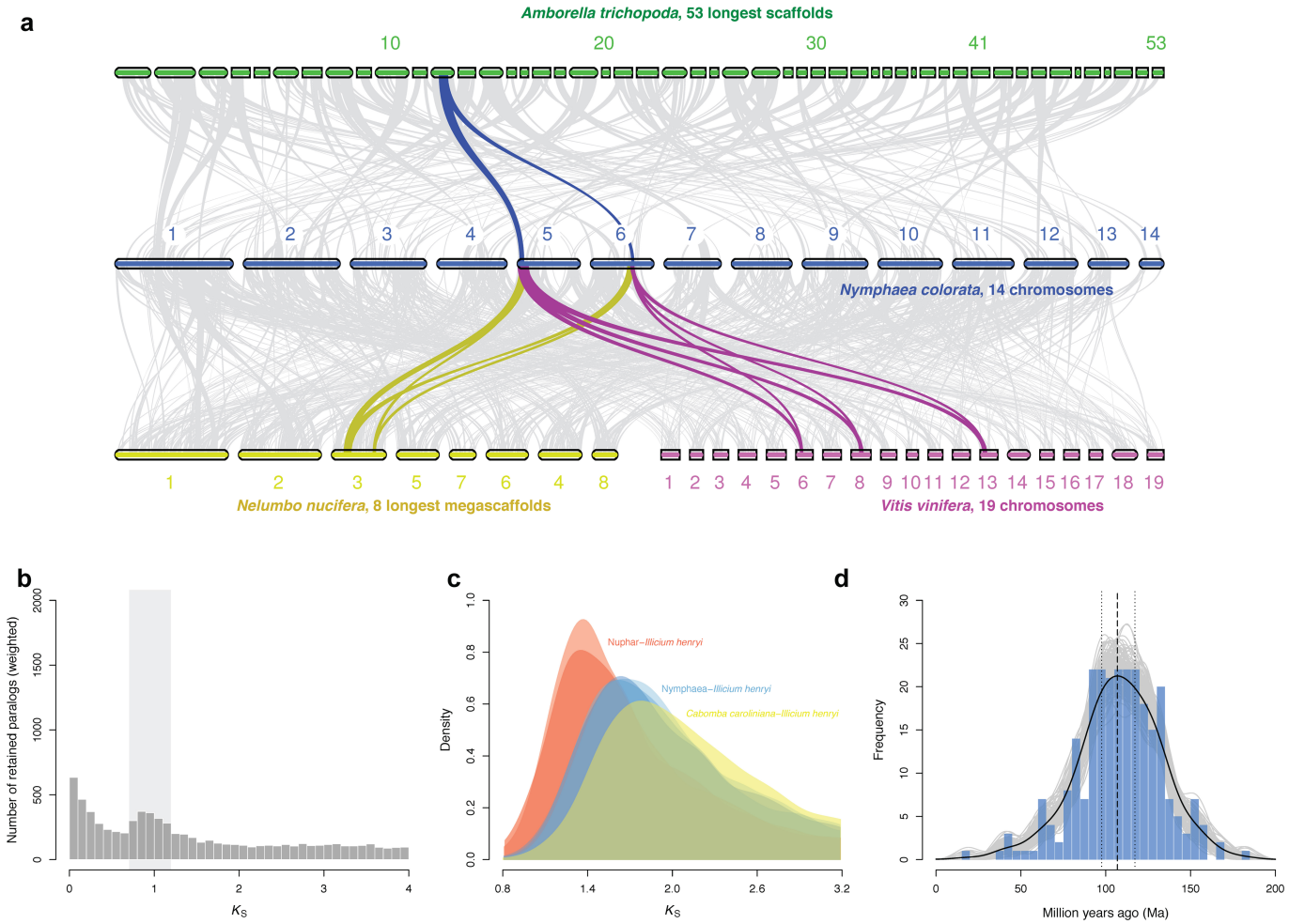
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1852-5.

Correspondence and requests for materials should be addressed to L.Z.

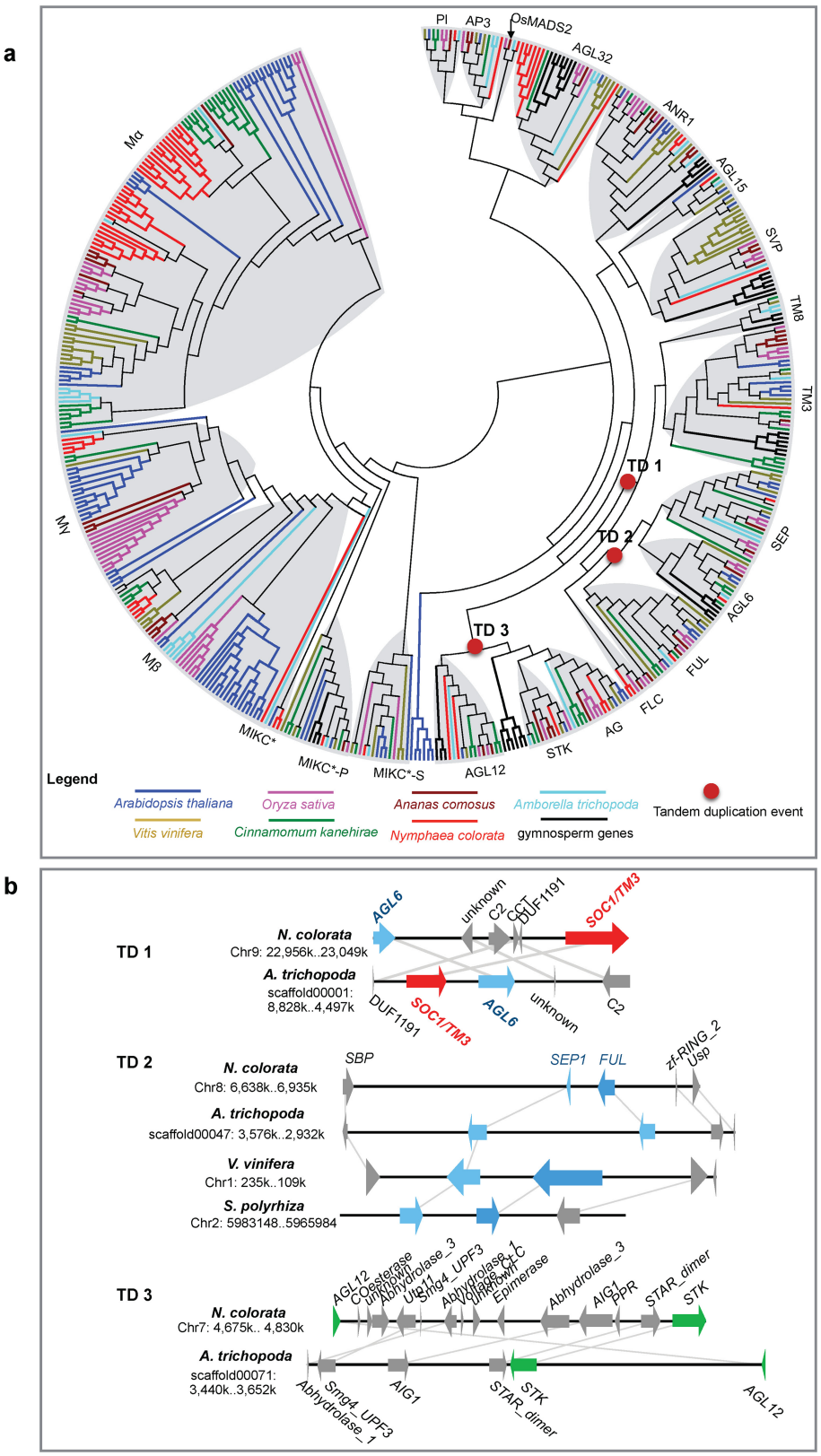
Peer review information *Nature* thanks Patrick Wincker and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



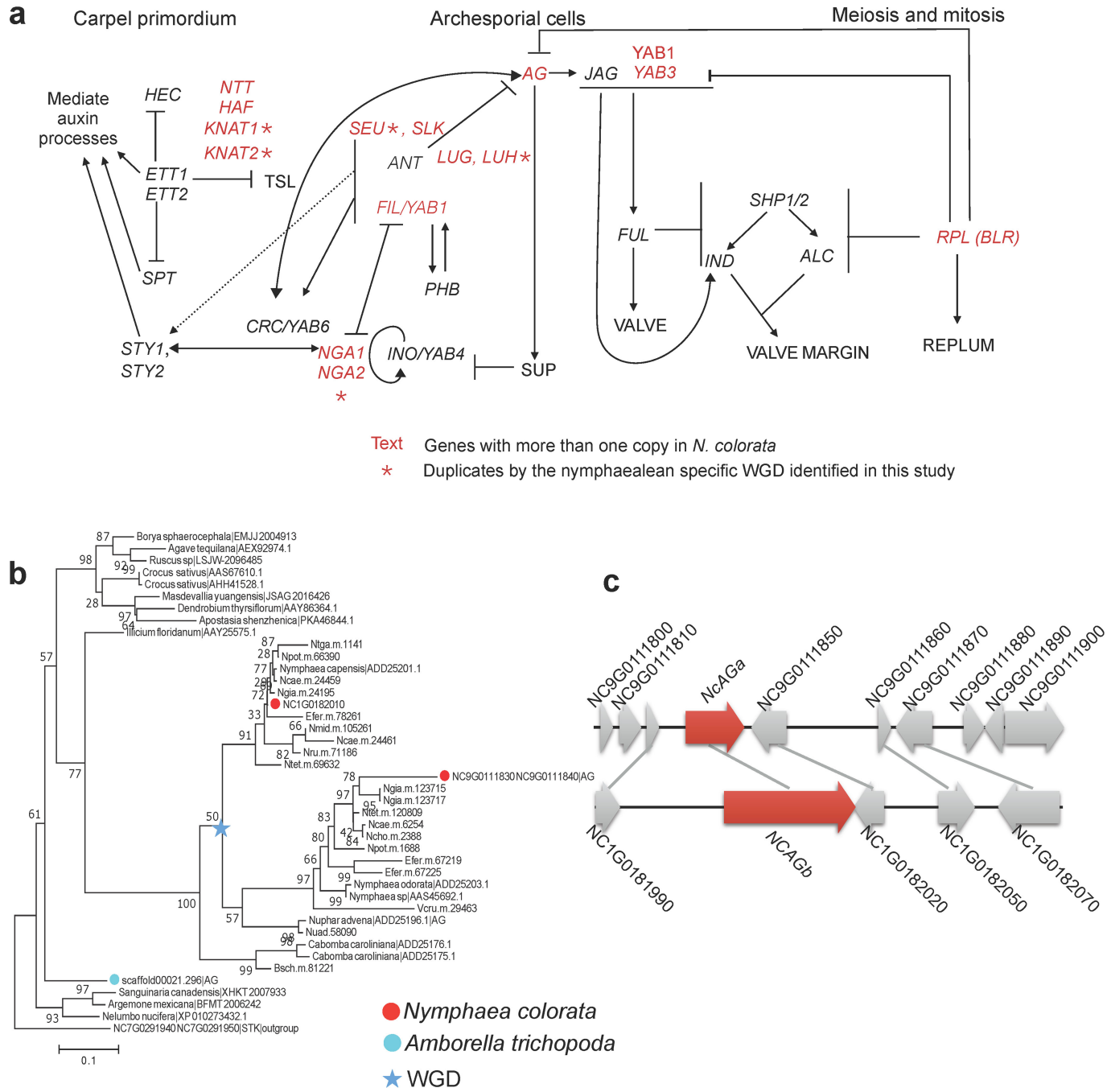
Extended Data Fig. 1 | High-quality genome of *N. colorata* allows integration of genetic and expression data. a, The assembled 14 chromosomes. **b**, Gene density plotted in a 100-kb sliding window. **c**, Transposable element (TE) density plotted in a 100-kb sliding window. **d**, Gene expression atlas of the juvenile flower, expression values were transformed with $\log_2$(FPKM + 1). **e**, GC content plotted in a 100-kb sliding window. **f**, Intragenomic syntenic regions denoted by a single line represent a genomic syntenic region covering at least 20 paralogues.



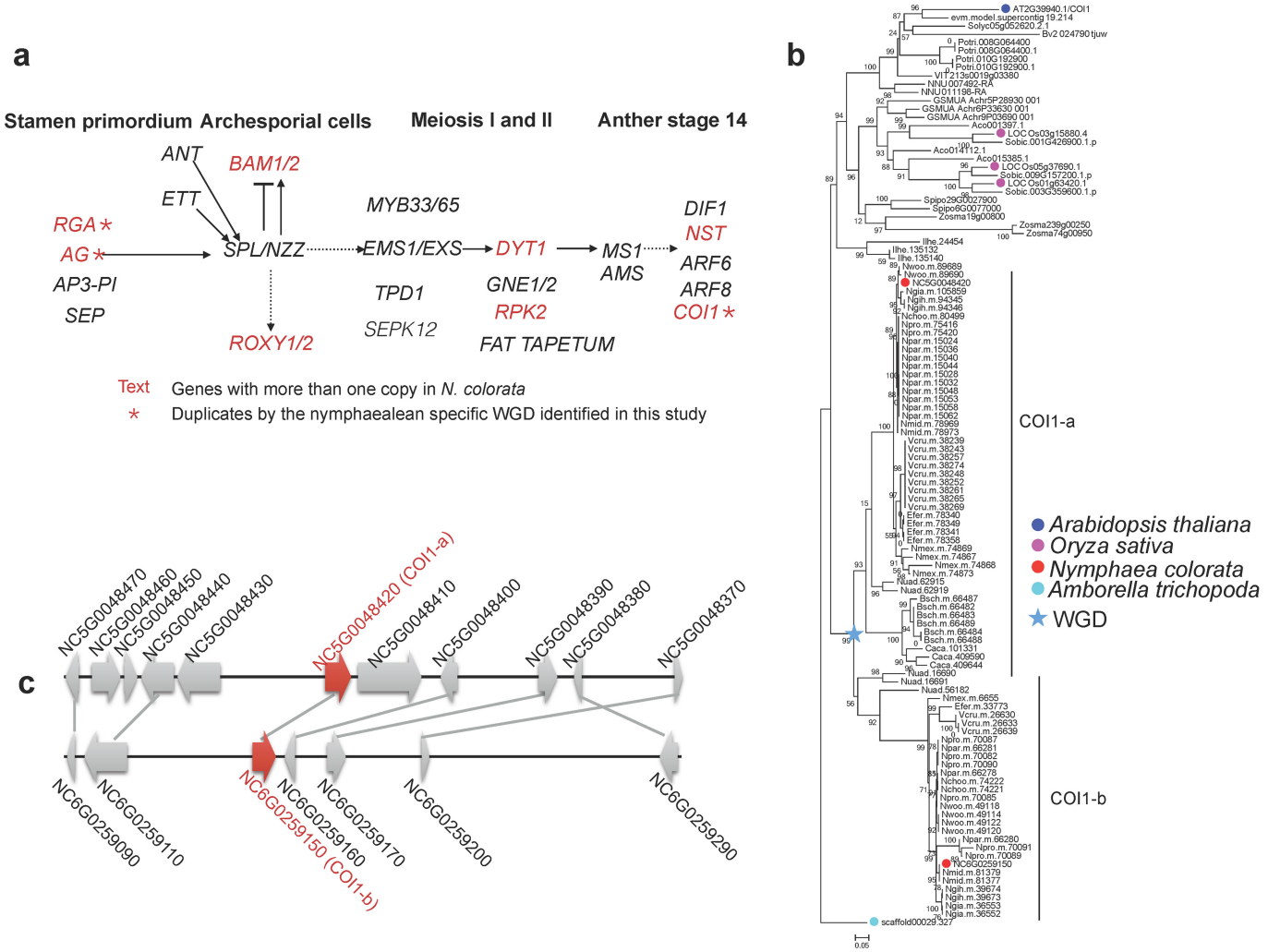
Extended Data Fig. 2 | WGD in Nymphaeales. a, Intergenomic synteny between *N. colorata* (14 chromosomes), *Amborella* (53 longest scaffolds), and the eudicots *N. nucifera* (8 longest megascaffolds) and *V. vinifera* (19 chromosomes). Five adjacent anchor pairs were plotted as one syntenic line. Coloured lines represent one example of syntenic genes found in other species that correspond to one copy in *Amborella*, two in *N. colorata*, two in *N. nucifera*, and three in *V. vinifera*. **b**, K_S distribution for the whole paranome of *N. colorata*. The light grey rectangle in the background indicates the K_S boundaries used to extract duplicate pairs for absolute phylogenomic dating of the WGD event, and also highlights the range in which WGD peaks can be identified in other species of Nymphaeaceae (Supplementary Note 5.2). **c**, Kernel-density estimates of K_S distributions for one-to-one orthologues between the outgroup species *I. henryi* and each of *N. lutea* and *N. advena* (red), *N. colorata*, *N. mexicana* and *Nymphaea* 'Woods blue goddess' (blue) and *C. caroliniana* (yellow). As each peak represents the same divergence event in the angiosperm phylogeny, the differences observed among the K_S values of the peaks indicate substantial substitution rate variation among these Nymphaealean lineages (see also Fig. 2b). **d**, Absolute age distribution obtained from phylogenomic dating of *N. colorata* WGD paralogues based on orthogroups with orthologues from *Amborella* and *G. biloba*. The solid black line represents the kernel density estimate of paralogue date estimates, and the vertical dashed black line represents its peak at 107 Ma. The grey lines represent density estimates from 2,500 bootstrap replicates and the vertical black dotted lines represent the corresponding 90% confidence interval for the WGD age estimate, 117–98 Ma (see Methods). The blue histogram shows the raw distribution of divergence date estimates for paralogues.



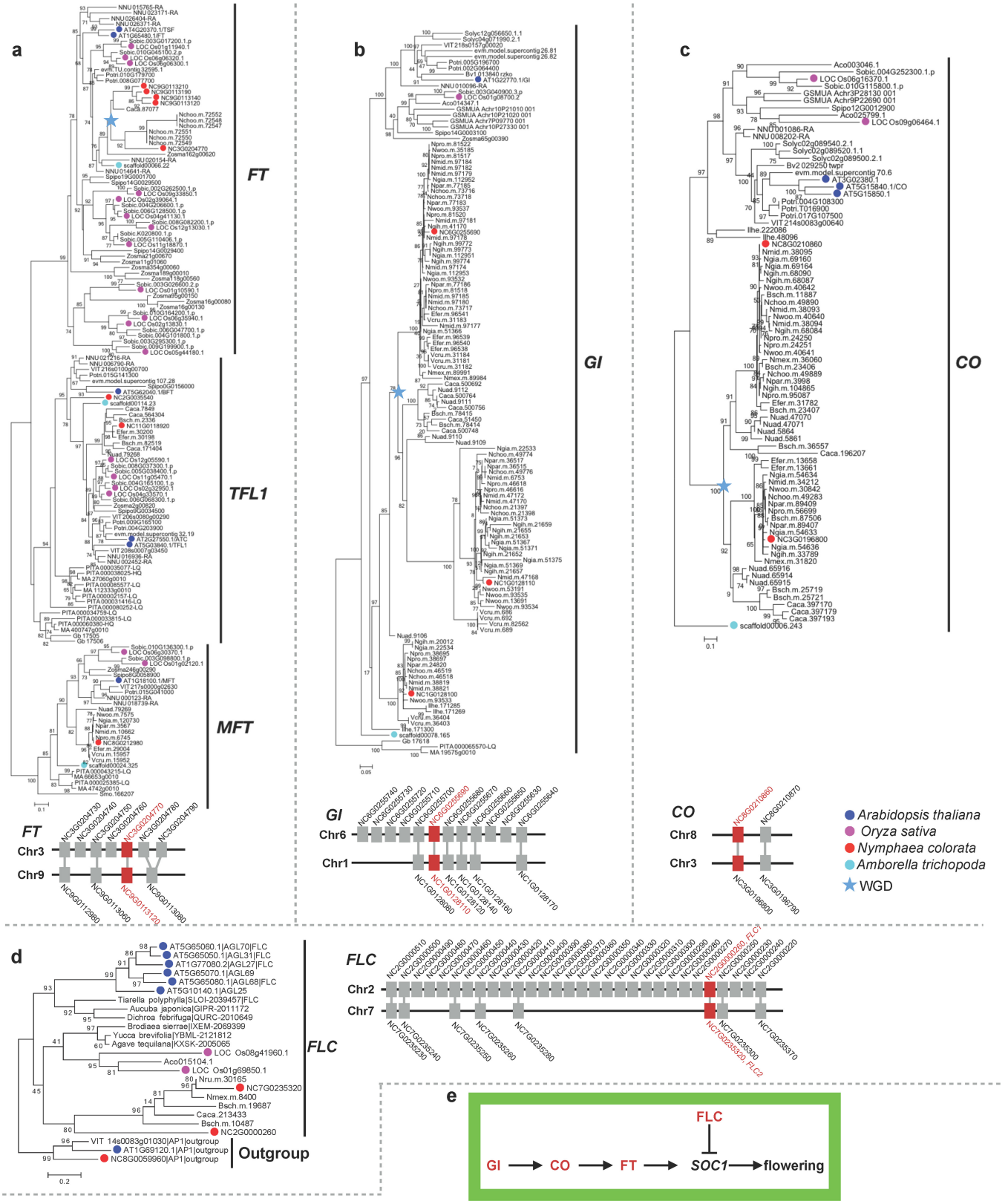
Extended Data Fig. 3 | The phylogenetic tree of MADS-box genes of *N. colorata*. **a**, The MADS-box genes are divided into type I and type II, and the latter was subdivided into MIKC^c and MIKC*. Branches of various species are shown in different colours, with the colour code below the tree. The nodes representing three tandem duplication events (TD1, TD2, and TD3 in **b**) are marked with red circles. **b**, Genomic regions with the duplicated genes derived from the three tandem duplication events (TD1, TD2 and TD3).



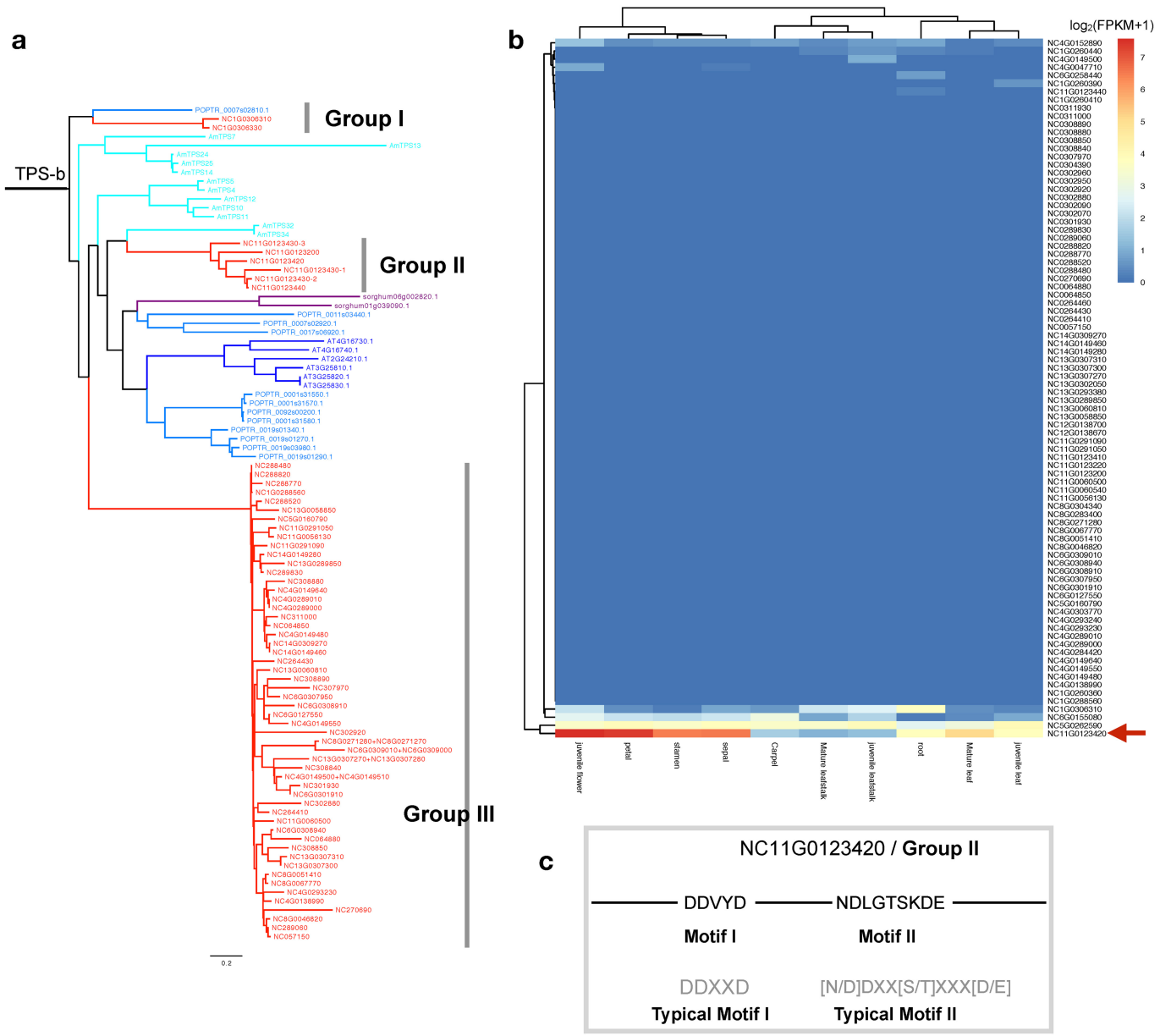
Extended Data Fig. 4 | Expansion of key genes regulating the carpel development by the Nymphaealean WGD. a, The reported pathway and genes that regulate carpel development. The red-labelled gene has two copies in *N. colorata*. The asterisk indicates that there is collinear support and is retained by the nymphaealean-specific WGD. **b**, Phylogenetic tree of *AG* genes, which specify floral meristem to determine the carpel and stamen identity. The star indicates the WGD specific to the water lily, as detected in this study. The duplicated *AG* genes in *N. colorata* are highlighted in red. **c**, Nc*AG* gene duplicates, Nc*AGa* and Nc*AGb*, are the result of the nymphaealean-specific WGD.



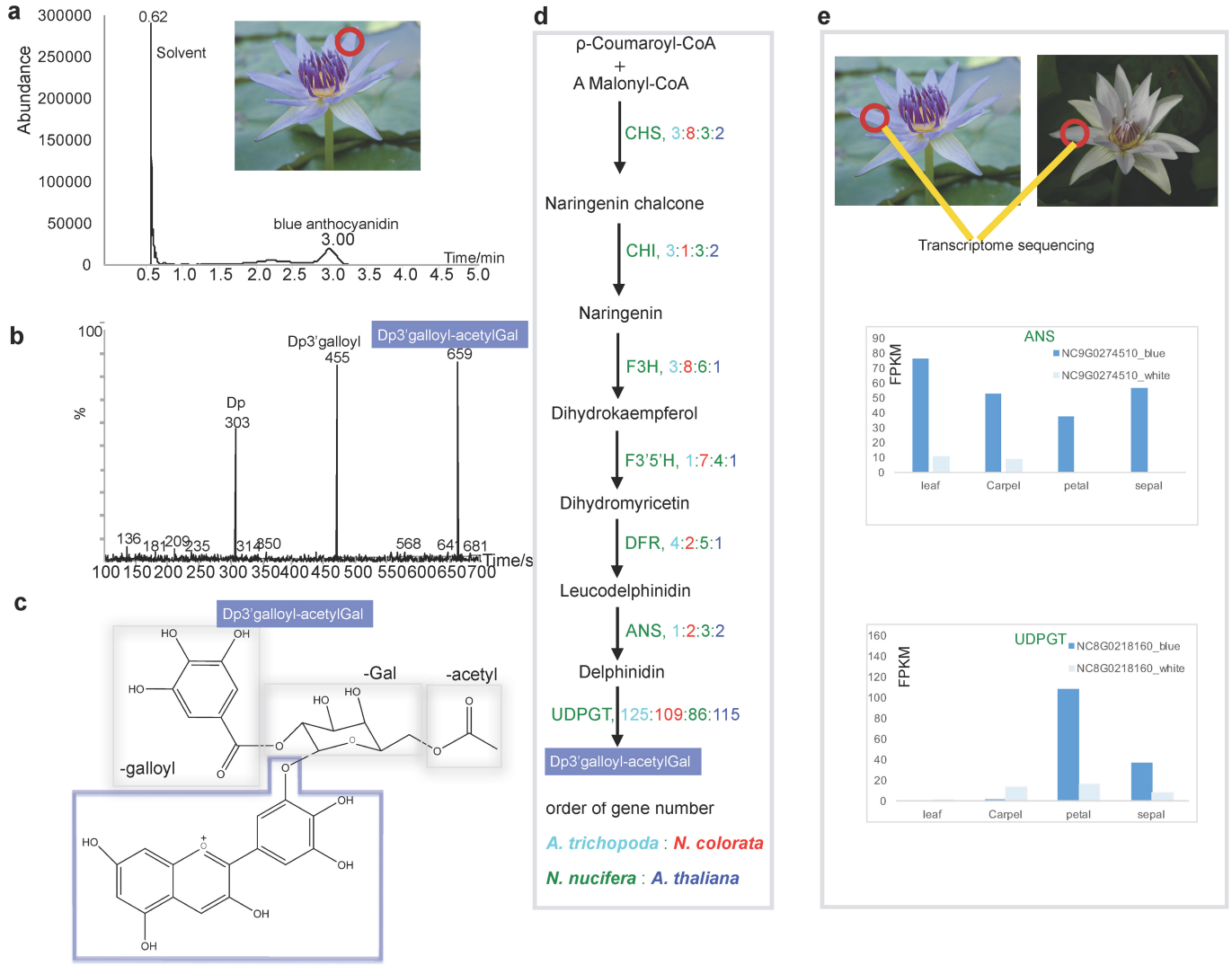
Extended Data Fig. 5 | Expansion of key genes regulating the development of the stamen by Nymphaealean WGD. a, The reported pathway and genes that regulate the stamen development. The red-labelled gene has two copies in *N. colorata*. The asterisk indicates that there is collinear support and is retained by the nymphaealean-specific WGD. **b**, Phylogenetic tree of *CORONATINE INSENSITIVE 1* (*COI1*), which recruits regulators of pollen development for modification by ubiquitination, needed in the JA response and regulating pollen fertility. **c**, Nc*COI1* gene duplicates have evolved through the WGD in Nymphaeales. The star indicates the nymphaealean-specific WGD.



Extended Data Fig. 6 | Nymphaealean-specific duplication of the genes that control the initiation of flowering. a, Phylogenetic tree of the PEBP-domain containing gene family, including *FT*, *TFL1* and *MFT* subfamilies across various water lily species and other representative seed plants. **b**–**d**, Phylogenetic tree of the *GI* (**b**), *CO* (**c**) and *FLC* (**d**) gene family across various water lily species and other representative seed plants. **e**, The regulatory pathway for the flowering time control. The red-labelled gene has two copies in *N. colorata* and is retained by nymphaealean-specific WGD.



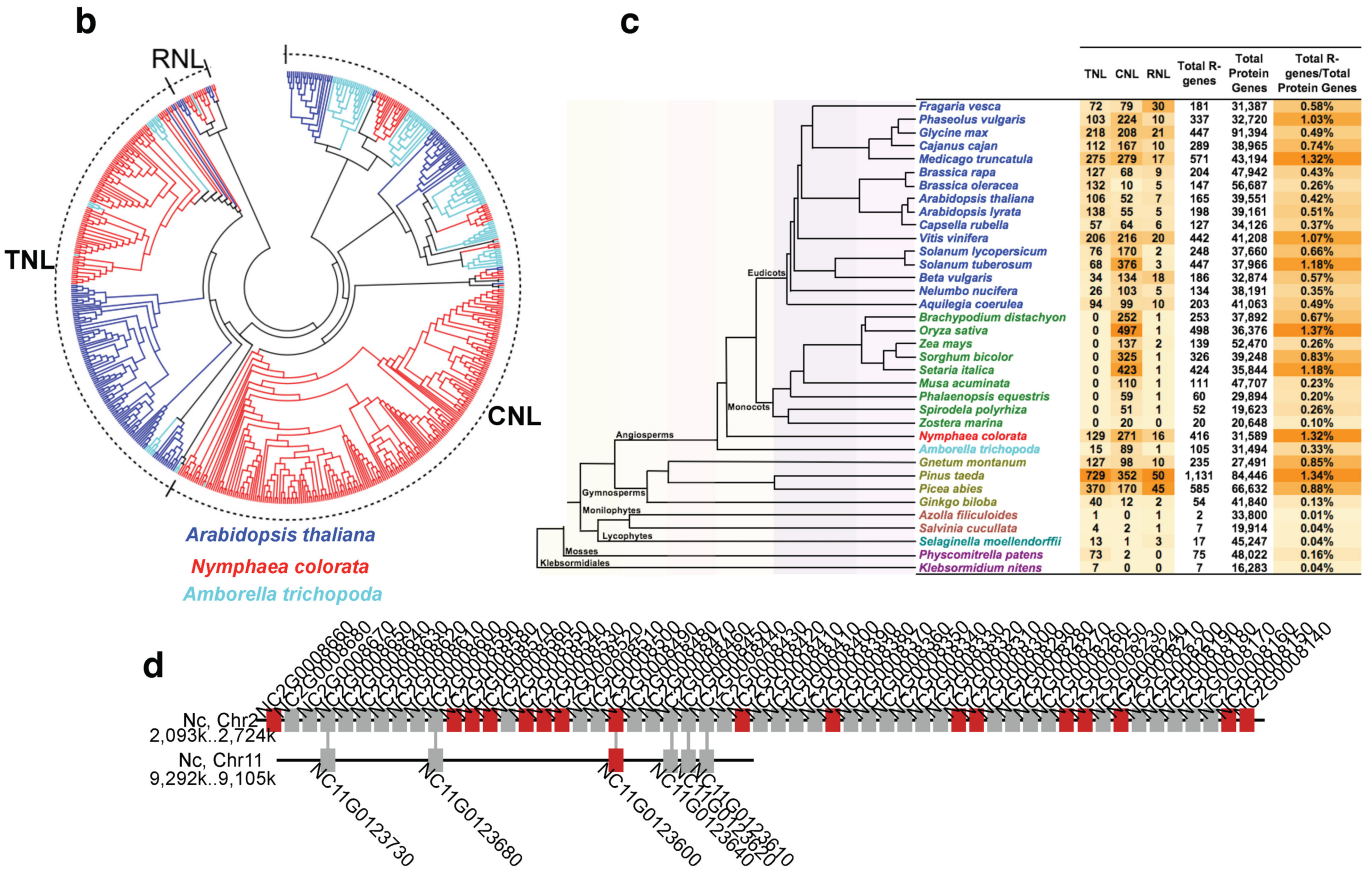
Extended Data Fig. 7 | Explosive expansion of the TPS-b subfamily and its implications. a, The phylogenetic classification of TPS-b subfamily into three groups. **b**, The group II member NC11G0123420 is the sole gene with high expression in the petal. **c**, Whereas most TPS-b members lack the two typical catalytic motifs, the NC11G0123420 retained both motifs, suggesting its potential role in producing sesquiterpene in *N. colorata*.



Extended Data Fig. 8 | The blue anthocyanidin and its potential biosynthesis pathway in *N. colorata*. **a**, The peak of the blue anthocyanidin appears at 3 min of the high-performance liquid chromatography (HPLC) detection. **b**, The three fragments of the blue anthocyanidin and their molecule mass. **c**, The molecule of the anthocyanidin was identified as delphinidin 3′-*O*-(2″-*O*-galloyl-6″-*O*-acetyl-β-galactopyranoside), abbreviated as Dp3′galloyl-acetylGal. **d**, The postulated pathway for the biosynthesis of Dp3′galloyl-acetylGal. Gene copy numbers are listed next to the enzymes. 3GGT, anthocyanidin 3-*O*-glucoside-2″-*O*-glucosyltransferase; 3′GT, 3′-*O*-beta-glucosyltranferase; 5AT, anthocyanin-5-aromatic acyltransferase; ANS, anthocyanidin synthase; CHI, chalcone isomerase; CHS, chalcone synthase; DFR, dihydroflavonol-4-reductase; F3H, flavanone-3-hydroxylase; F3′5′H, flavonoid-3′,5′-hydroxylase. **e**, Comparative transcriptomic analyses between the blue- and white-petal cultivars of *N. colorata* identified two genes, *ANS* and *UDPGT*, that are highly differentially expressed and might be potential regulators for blue coloration of the petals.

a

	Gene families	*Amborella trichopoda*	*Nymphaea colorata*	*Spirodela polyrhiza*	*Sorghum bicolor*	*Oryza sativa*	*Nelumbo nucifera*	*Vitis vinifera*	*Solanum lycopersicum*	*Arabidopsis thaliana*
Stress related	Gretchen Hagen3 (GH3)	6	36	13	13	14	16	9	31	20
	Auxin_inducible	30	62	55	76	62	88	93	106	78
	Salt stress response/antifungal (stress_antifung)	31	102	35	60	77	45	56	37	96
	Xylanase inhibitor (TAXi)	63	202	48	101	113	79	63	92	68
	NB-ARC (NLR)	105	416	52	326	498	134	442	248	165
	Terpene synthase	28	92	17	43	52	20	117	50	34
	Protein kinase	654	1176	784	1254	1468	1148	1320	1119	1022
	Cu oxidase	24	43	22	46	42	54	98	52	41
Transcription factors	WRKY	32	69	43	97	94	64	62	81	73
	MADS-box	34	70	44	78	75	39	81	105	108
	B3 DNA binding domain	32	57	39	89	87	68	73	102	93
	HSF_DNA-bind (Heat shock factor-type)	12	21	13	24	25	28	19	26	24
	NAC	44	68	53	128	136	86	83	96	111
	MYB	138	199	174	271	247	269	262	283	264
	AP2	73	102	81	173	157	124	140	165	141



Extended Data Fig. 9 | Expanded stress-related and transcription factor gene families in the genome of *N. colorata*. **a**, Markedly expanded gene families for stress response and transcriptional regulation. NLR genes contain NB-ARC domains. Notably, *N. colorata* encodes the highest proportion of kinase genes compared with gymnosperms or other land plants. **b**, NLR genes expanded in all of its three subfamilies (*RNL*, *TNL* and *CNL*). **c**, Distribution of NLR genes across the representative algae and land plants. The background colours indicate the number variation in each species. **d**, An example showing how tandem duplication and WGD contributed to the expansion of R genes in *N. colorata*.

Extended Data Table 1 | Statistics of the sequenced and assembled genome of *N. colorata*

Statistic	Reads*	Contigs	Scaffolds	Chromosomes
Number	5,521,269	1,429	804	14
Longest Length	78 Kb	12.79 Mb	44.61 Mb	44.61 Mb
Total size	49.76 Gb	409.09 Mb	409.15 Mb	378.81 Mb
N50	12.59 Kb	2.14 Mb	25.52 Mb	27.06 Mb

*The reads only include sequencing by PacBio RS II SMRT sequencing technology.

nature research

Corresponding author(s): Liangsheng Zhang

Last updated by author(s): Sep 23, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software had been used for data collection
Data analysis	CANU (v1.3), Quiver, Arrow, Pilon (v1.20), bowtie2, Lachesis, HiC-Pro, Genscan, Augustus, Tandem Repeats Finder (v4.09), LTR_FINDER (v1.0.6), RepeatModeler (v1.0.5), RepeatMasker (v4.0.6), MAKER (v2.31.8), BUSCO software (v3.0.1), Tophat (v2.1.1), TRIMMOMATIC, Cufflink, TransDecoder, CD-HIT, Orthofinder (v1.1.8), Pfam (V31.0),MUSCLE(v3.8.31), TreePL, r8s, MCMCtree, ModelFinder, PAL2NAL (v14), trimAl (v1.4), RaxML (v8.0.19), PAML package (v4.4c), PhyML, ASTRAL(v5.5.12), PAUP(v4.0), InParanoid, LAST, BLAST, QUOTA-ALIGN, BEAST(v1.7), OrthoMCL (v2.0.9), Geneious (V.10.0.2), R, Python. Perl. Specific parameters used during run-time are provided in the methods. All softwares or scripts are available from official websites or GitHub as indicated in the methods.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

PacBio whole-genome sequencing data and Illumina data were deposited to the SRA at the NCBI under the BioProject ID PRJNA565347.
PacBio whole-genome sequencing data and Illumina data also were deposited in the BIG Data Center (http://bigd.big.ac.cn) under project number PRJCA001283.
The genome assembly sequences and gene annotations have been deposited in the Genome Warehouse in BIG Data Center under accession number GWHAAYW00000000 and in ENA BioProject (PRJEB34452). The genome assembly sequences and gene annotations have been also deposited in the Waterlily Pond (http://waterlily.eplant.org). All these data are freely available to the public.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. Our samples were all from wild type and did not use processed samples and groups.
Data exclusions	No data were excluded.
Replication	The genome sequence was taken and sequenced with more than 120 fold coverage. No replication is needed our genome reports.
Randomization	No random sampling is required for genome sequencing, because the genome differences are very small within the wild population, thus any wild plant is allowed for genome sequencing.
Blinding	Blinding is not applicable in our study because it does not involve subjects which receive different treatments.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Nuclei were isolated from young leaves in spring ,using PI staining for 15 minutes.
Instrument	Beckman Coulter COULTER EPICS XL™
Software	FACS data analyses were performed using CXP v2.2 Software
Cell population abundance	abundance >8000 cells were collected for each sample. Total nuclei populations were gated using relative fluorescence intensity: the proportions of nuclei with different ploidy levels were determined based on their relative fluorescence intensity: Pear is a diploid (2N) as a reference, according to the peak position (Supplementary Figure 5).

Gating strategy

Total nuclei populations were gated using PI intensity. In PI+ singles cells, the proportions of nuclei with different ploidy levels were determined based on their PI intensity (Supplementary Figure 5).

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.