The slow-evolving Acorus tatarinowii genome sheds light on ancestral monocot evolution

Tao Shi1,2, Cécile Huneau3, Yue Zhang1,2,4, Yan Li1,2,4, Jinming Chen1,2 ⋆, Jérôme Salse3,5 and Qingfeng Wang1,2,5 ⋆

Monocots are one of the most diverse groups of flowering plants, and tracing the evolution of their ancestral genome into modern species is essential for understanding their evolutionary success. Here, we report a high-quality assembly of the Acorus tatarinowii genome, a species that diverged early from all the other monocots. Genome-wide comparisons with a range of representative monocots characterized Acorus as a slowly evolved genome with one whole-genome duplication. Our inference of the ancestral monocot karyotypes provides new insights into the chromosomal evolutionary history assigned to modern species and reveals the probable molecular functions and processes related to the early adaptation of monocots to wetland or aquatic habitats (that is, low levels of inorganic phosphate, parallel leaf venation and ephemeral primary roots). The evolution of ancestral gene order in monocots is constrained by gene structural and functional features. The newly obtained Acorus genome offers crucial evidence for delineating the origin and diversification of monocots, including grasses.

Monocots are one of the most diverse and dominant clades of flowering plants, accounting for approximately 21% of angiosperm species diversity. This clade not only includes commonly consumed horticultural products, such as banana, garlic, asparagus and coconut, but more importantly also contains the grass/cereal family (Poaceae), which comprises almost half of monocots, with economically important species such as rice, wheat, oat, sorghum and maize. The earliest fossil record of monocots, such as Cratolirion bognerianum, dates back to the Early Cretaceous, and molecular dating using fossil-calibrated phylogenetic trees suggests that the crown group of monocots can be traced back to approximately 132.4–149.1 million years ago (Ma) during the Early Cretaceous. This crown group diversified almost at the same time as the magnolids and eudicots. The ancestral monocot has been proposed to have an aquatic origin because the fossil record of Alismatales has been dated back to at least the Upper Cretaceous. In addition, fossils of some of the early-branching monocots morphologically resemble some extant members of those lineages and may, therefore, have shared similar habitats with typical submerged and amphibious aquatic species (Acorales, Alismatales and Hydatellaceae). However, this origin remains ambiguous because of a lack of compelling proof from either palaeontology or genetics.

Exploring genomic conservation and changes during monocot evolution in a considerable sampling of taxa can help to understand the driving factors that influenced the evolutionary trajectory of monocots in terms of gene order change during monocot diversification. Whole-genome duplications (WGDs) or polyploidizations are rampant during monocot diversification and have been proposed as a key mechanism driving species diversification and adaptation. To what extent polyploidization and derived genome reshuffling may have driven monocot diversification among the flowering plants is an open question that requires sampling from early-branching lineages and in-depth surveying. Moreover, at the chromosomal level (karyotype), uncovering patterns of chromosomal fusion, fission, duplication and loss during species radiation is important for our understanding of the evolutionary processes underlying monocot species diversity. By reconstructing ancestral monocot karyotypes (AMK) and gene family history, we can further uncover some key genomic changes underlying the evolutionary success of monocots.

According to phylogenetic evidence obtained by large-scale taxonomic sampling, Acorus is sister to other orders in monocots. Thus, similar to Amborellales for angiosperms and Ranunculales for eudicots, Acorales species are phylogenetically critical for understanding the evolutionary history of monocots. Therefore, to better track genome evolution during the emergence of monocots, we sequenced and assembled at the chromosomal level the genome of Acorus tatarinowii Schott (also known as Acorus gramineus), a medicinal plant from wetlands and creeks in East Asia with an essential oil that has antidepressant-like effects. Considerable comparative analysis between Acorus and the genomes of grasses (Poaceae) and other monocot orders (such as oil palm and asparagus) allowed us to reconstruct the karyotype of the most recent common ancestor (MRCA) of all extant monocots (AMK) and further uncover key genomic events associated with the important traits and aquatic or wetland origin of ancestral monocots.

Results

Genome assembly and ancient tetraploidization of Acorus tatarinowii. The Acorus tatarinowii Schott (Acorus) genome sequenced in this study is diploid (2n = 24, see http://ccdb.tau.ac.il/), with a size estimate of 470.3 Mb, an estimated heterozygosity of 0.88% and a repetitive content of 54.82%, as revealed by genomic character estimator analysis based on Illumina short reads (Supplementary Fig. 1). Based on PacBio, high-throughput chromosome conformation capture (Hi-C) and RNA-sequencing (RNA-seq) data, we delivered a chromosomal-level assembly and annotation of the Acorus

1CAS Key Laboratory of Aquatic Botany and Watershed Ecology, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, China. 2Center of Conservation Biology, Core Botanical Gardens, Chinese Academy of Sciences, Wuhan, China. 3UCA, INRAE, UMR 1095 GDEC (Genetics, Diversity & Ecophysiology of Cereals), Clermont-Ferrand, France. 4University of Chinese Academy of Sciences, Beijing, China. 5Sino-African Joint Research Center, Chinese Academy of Sciences, Wuhan, China. E-mail: jmchen@wbgcas.cn; jerome.salse@inrae.fr; qfwang@wbgcas.cn
genome. De novo assembly was based on 5,012,373 PacBio Sequel subreads with a total length of 110.07 Gb, a mean length of 21.96 kb and an N50 length (a metric for sequence or assembly) of 36.72 kb (Supplementary Table 1). The final 1,076 contigs covered approximately 415.18 Mb with an N50 length of 961.57 kb. Using 43 Gb of genome-wide Hi-C reads, 1,108 contigs (379.11 Mb) were anchored and ordered into 12 different pseudomolecules (Extended Data Fig. 1 and Supplementary Table 2). Among 1,614 conserved single-copy genes in BUSCO (version: embryophyta_odb10), 92.40% (1,491) and an N50 length (a metric for sequence or assembly) of 36.72 kb. For example, chro-

| VOL 8 | JULY 2022 | 764–777 | www.nature.com/natureplants | 765 |

**Fig. 1** | Genome assembly and a WGD of *Acorus*. a, Circos plot of *Acorus* genome assembly. From outer to inner circles: gene density, TE density, pseudochromosome length and *Acorus* in an aquatic habitat. b, Scatter plot of *Acorus* intraspecific synteny. c, Density distributions of syntenic paralogues and syntenic orthologues between *Acorus* and other monocots according to their 4dTV divergence. d, An illustration of biased subgenome fractionation between two homologous *Acorus* regions when compared with *Aristolochia*. e, Violin plot showing significantly higher gene expression in the LF subgenome (shown in red) than in the MF subgenome (shown in blue) for five tissues. Exact P values shown on the top of each violin plot are from one-sided paired t tests. f, Differences in average TE density along genes and flanking regions between duplicates residing in LF blocks and MF blocks. |
Phylogenetic positioning and genomic conservation of *Acorus*. Because *Acorus* shows genomic evidence of one single WGD, we suspect that it has a relatively conserved genome architecture within monocots. Thus, the interspecific syntenies are expected to be longer and less fragmented for *Acorus*, which is also supposed to share more collinear genes than other monocots when compared with non-monocot genome(s). Alignment of monocot genomes to outgroup taxa with an available chromosomal-level assembly, including *Amborella trichopoda* (the earliest branching angiosperm)\(^1\), *Nymphaea colorata* (closely related to the Nymphaeaceae ancestral genome\(^2\)), *Aristolochia fimbriata* (a Magnoliidae species without a WGD\(^3\)), *Cinnamomum kanehirae* (closely related to the Magnoliidae ancestral genome\(^4\)) and *Nelumbo nucifera* (closely related to the eudicots as major clades of the early-branching flowering plants). We found that *Acorus* shared more collinear orthologues (anchor genes) with all sequenced monocots compared with non-monocot references (representing the Amborellales, Nymphaeales, Magnoliidae and eudicots as major clades of the early-branching flowering plants).

Finally, we investigated what factors (such as substitution rate or ancient WGD) are associated with the synteny decay rate among monocots. Based on multiple sequence alignments of 104 single-copy orthologues, the maximum likelihood tree of monocots and outgroup taxa confirmed *Acorus* at the earliest branching position within all sequenced monocots (Fig. 2a and Supplementary Fig. 6). Notably, *Acorus* also showed the shortest sum of branch lengths from the MRCA of extant monocots, suggesting that *Acorus* is not only the earliest branching taxon (Fig. 2a), but also has the slowest sequence substitution rate among the surveyed monocot species (Supplementary Fig. 6). Furthermore, we reported that the synteny retention rates of monocots were strongly and negatively correlated with the relative sequence substitution rates and all had *P* values <0.01, indicating that rapid genome architecture change was associated with rapid sequence substitution (Fig. 2d and Supplementary Fig. 6).
Fig. 3 | Factors associated with the distinct patterns of synteny loss in 42 monocot species based on different genes in the outgroup Nelumbo. a. The number of monocot species syntenic to Nelumbo (red bar) across the eight Nelumbo chromosomes. b. Violin plot of the number of monocot species syntenic to Nelumbo regarding gene groups of different duplication origins. c–f. Violin plots showing incremental changes in upstream gene methylation (c), downstream gene methylation (d), tissue specificity of expression (e) and exon number (f) for Nelumbo genes from group I to group IV. One-way Kruskal-Wallis test significance is shown on the top of each violin plot (adjusted P values). w/ syntenic, with syntenic homologue.

Supplementary Fig. 7a–d). We also showed that the synteny retention rates were negatively correlated with the number of ancient WGDs (paleopolyploids), with P values of 0.0032, 0.011, 0.064, 0.016 and 0.012 for Amborella, Nymphaea, Cinnamomum, Aristolochia and Nelumbo, respectively, which were considered outgroups (Fig. 2d and Extended Data Fig. 4a–e). These results are in line with previous case studies that show extensive chromosomal rearrangements (synteny loss) after a single WGD25–27, as well as accelerated synteny loss with a series of WGDs. Nevertheless, we showed that there was no significant correlation between synteny loss rate and genome size, suggesting that the repetitive fraction of the genome does not significantly affect genome architecture or gene order conservation between monocots (Supplementary Fig. 8a–c).

Biased synteny retention among different genes during monocot evolution. To further explore the factors related to synteny retention or loss among different genes during monocot evolution, we aligned the genome of the closest outgroup (Extended Data Fig. 3 and Supplementary Fig. 4), Nelumbo⁷, to monocot genomes. This is because unlike other early-branching outgroups with limitations in functional and population data, Nelumbo offers abundant public data on gene expression from diverse organs and tissues, whole-genome methylation and population resequencing⁸. Examining this horticultural crop allowed us to gauge the variation in synteny retention rate during monocot radiation among different functional gene categories. We illustrated the rate of synteny conservation along the Nelumbo chromosomes and observed that the synteny retention rate was low for genes near centromeres that were enriched in TEs (Fig. 3a), putatively due to fewer genes being located near centromeres and the presence of rapid structural changes mediated by repeated sequences in these regions. We reported a difference in synteny retention depending on the status of a gene: whether it had been duplicated or not during the course of evolution⁹. We found that WGD-derived genes showed the highest retention rates, followed by ‘WGD&tandem’ genes, single-copy genes, tandem duplicates, proximal duplicates and dispersed duplicates (Fig. 3b). This result suggests that WGD, WGD&tandem genes and single-copy genes are older than those in other categories, which may reflect stronger functional constraints on these gene categories, whereas local duplicates (tandem and proximal) and dispersed duplicates are younger and under fewer structural and possibly functional constraints⁴. Despite the structural fate of syntenic genes, we also investigated their regulation, such as expression and epigenetic marks⁵. Based on the coefficient of determination R² that measures the strength of correlation, we found that the synteny retention rate of Nelumbo genes in monocots is significantly correlated with gene-related traits such as the methylation level of flanking regions around genes (−1 kb and +1 kb), tissue specificity of gene expression (τ index), number of exons, coding sequence (CDS) length, average expression level (fragments per kilobase of exon per million reads) and Extended Data Fig. 4a–e). These results are in line with previous case studies that show extensive chromosomal rearrangements (synteny loss) after a single WGD25–27, as well as accelerated synteny loss with a series of WGDs. Nevertheless, we showed that there was no significant correlation between synteny loss rate and genome size, suggesting that the repetitive fraction of the genome does not significantly affect genome architecture or gene order conservation between monocots (Supplementary Fig. 8a–c).
## Table 1 | Linear regressions between the number of monocot species with a syntenic anchor to the *Nelumbo* gene (x) and different gene-related traits (y) for all *Nelumbo* genes

| Gene traits (y) | Linear regression | r     | R²                        | P value          |
|----------------|-------------------|-------|---------------------------|-----------------|
| Methylation level (−1kb) | y = 0.0581 – 0.00111x | −0.22 | 0.0484                   | 2.20 x 10⁻¹⁶    |
| Methylation level (+1kb) | y = 0.0571 – 0.000983x | −0.21 | 0.0441                   | 2.20 x 10⁻¹⁶    |
| Tissue specificity (r index) | y = 0.571 – 0.00426x | −0.21 | 0.0441                   | 2.20 x 10⁻¹⁶    |
| No. of exons | y = 4.03 + 0.0657x | 0.18  | 0.0324                   | 2.20 x 10⁻¹⁶    |
| CDS length | y = 1020 + 12.1x | 0.17  | 0.0289                   | 2.20 x 10⁻¹⁶    |
| Average expression level (FPKM) | y = 2.07 + 0.0242x | 0.15  | 0.0225                   | 2.20 x 10⁻¹⁶    |
| Methylation level (gene) | y = 0.0547 – 0.000709x | −0.15 | 0.0225                   | 2.20 x 10⁻¹⁶    |
| Proportion of TEs (+3kb) | y = 0.392 – 0.00208x | −0.12 | 0.0144                   | 2.20 x 10⁻¹⁶    |
| Nucleotide diversity (π) | y = 0.000677 – (1.01 x 10⁻⁵)x | −0.11 | 0.0121                   | 2.20 x 10⁻¹⁶    |
| Proportion of TEs (−3kb) | y = 0.423 – 0.0017x | −0.094 | 0.00836                   | 2.20 x 10⁻¹⁶    |
| Gene length | y = 7460 + 83.6x | 0.069 | 0.004761                   | 2.20 x 10⁻¹⁶    |
| PPI | y = 0.871 + 0.0108x | 0.034 | 0.001156                   | 1.40 x 10⁻⁶      |
| Proportion of TEs (gene) | y = 0.187 – 0.000149x | 0.0081 | 0.00006561                   | 0.16             |

r, correlation coefficient; R², coefficient of determination; −1kb or +3kb, upstream gene regions; +1kb or +3kb, downstream gene regions.
AMK clearly defines the transition between the 12 CARs that were previously defined and the $n=5$ pre-$\tau$ AMK, introducing six ancestral chromosome fusions to reach an $n=6$ AMK intermediate (represented by six colours, namely orange, dark blue, pink, light green, light blue and dark green in Fig. 4b) followed by one fission (dark green) and two fusions (dark green–orange, dark green–light blue in Fig. 4b) explaining the transition between the $n=6$ AMK and the previously reported $n=5$ pre-$\tau$ AMK at the MRCA of Ananas, palm and grasses (Extended Data Fig. 6). From the $n=6$ AMK, Colocasia and Spirodela experienced two duplications to reach an $n=24$ intermediate followed by 14 and 20 chromosomes fusions to reach their modern genome structure of 14 and 20 chromosomes,
respectively. *Dioscorea* (with 20 chromosomes) is inherited directly from the \( n = 5 \) pre-\( \tau \) AMK with seven fissions and eight fusions (Extended Data Fig. 7). The dotplot-based deconvolution of the synteny between the \( n = 6 \) AMK and the extant genomes validates the number of rounds of WGDs (method step 5) with one event reported in *Acorus* (Fig. 1b), and two events reported in *Spirodela, Colocasia* and *Dioscorea* (Fig. 4c and Extended Data Fig. 8).

Recently, Xu et al. suggested an \( n = 7 \) AMK before and after the ancestral \( \tau \)-WGD event from the comparison of *Acorus (A. americana)*, *Spirodela, Colocasia, Ananas* (pineapple) and *Elaeis* (palm)\(^{36}\). We then compared our proposed \( n = 6 \) AMK structure with that of the seven chromosomes from Xu et al. (Supplementary Fig. 9). The two proposed AMK ancestors show a perfect chromosome-to-chromosome relationship for chromosomes 1-5, 3-4 and 5-6 between, respectively, the current \( n = 6 \) AMK and the \( n = 7 \) AMK from Xu et al.\(^{36}\). Differences are observed between the proposed AMK ancestors for chromosomes 2-(2-6), 4-(3-4) and 6-(5-7) between, respectively, the current \( n = 6 \) AMK and the \( n = 7 \) AMK from Xu et al.\(^{36}\), corresponding to different alternative scenarios proposed to explain the transition between the proposed AMKs and the modern genomes (Extended Data Fig. 7). From the proposed \( n = 6 \) AMK in the current study, an evolutionary scenario (method step 6) can then be inferred by taking into account the fewest number of genomic rearrangements (including inversions, deletions, fusions, fissions and translocations) that may have occurred between the AMK and modern monocot genomes (Extended Data Fig. 7). Figure 4b summarizes the number of rearrangements as well as the intermediate number of chromosomes from the AMK to the modern species investigated; in particular, when comparing *Acorus* with AMK, 12 CARs following a lineage-specific duplication occurred to create the 12 modern chromosomes. Overall, all the early-branching monocots showed far fewer mosaic fragments originating from the AMK than from the AKG, which is probably due to extensive chromosomal rearrangement (synteny loss) after multiple grass WGDs (\( \tau, \sigma \) and \( \rho \)). Finally, our comparative genomics-based evolutionary scenario reveals the monocot palaeohistory from the AMK, with *Acorus*, sister to other extant monocots, having a karyotype most strongly resembling the AMK. Our analysis also delivers a complete catalogue of orthologues (Supplementary Table 6) between monocot genomes, which can now be used as a guide to perform translational research between the investigated species to accelerate the dissection of conserved agronomic traits.

**Biological functions at the emergence of monocots.** Monocots, as a monophyletic group, process distinctive phenotypes such as parallel leaf venation, ephemeral primary roots and scattered vascular bundles in the stem; these phenotypes are similar to those of Nymphaeales but quite different from those of Amborellales, Austrobaileyales, magnoliids and eudicots\(^{37}\). To infer the functions of genes driving early monocot evolution, we built a chronogram based on 28 representative angiosperm species with fossil constraints, which predicted that the MRCA of monocots dates back to approximately 169.76 Ma, consistent with the TimeTree database (Supplementary Table 7). Applying the Dollo-Parsimony approach, we found that 77 and 964 orthologous groups (OGs) were acquired and lost in the AMK, respectively (Supplementary Fig. 10a and Supplementary Table 7). The 77 OGs acquired in the ancestral monocot were enriched in Gene Ontology (GO) terms such as transporter activity, plasma membrane vacuole, membrane, cell communication, transport and response to external stimulus (Supplementary Fig. 11 and Supplementary Table 8), whereas the 964 OGs lost in the ancestral monocots were enriched in GO terms such as intracellular, Golgi apparatus, mitochondrion and cytoplasm (Supplementary Fig. 12 and Supplementary Table 9). For example, OG0010560 which contains WOX1 involved in cotyledonary primordia development, was completely lost in monocots (Supplementary Table 9). In addition, by setting a \( P \) value threshold of 0.05 for gene family expansion and contraction in CAFE software analysis, we extracted 41 OGs with significant expansion and 1,278 OGs with significant contraction in monocots (Supplementary Fig. 10b and Supplementary Table 7). The 41 OGs that were expanded in the ancestral monocot were enriched in GO terms such as metabolic process, cellular process and response to stress (Supplementary Fig. 13), whereas the 1,278 OGs contracted in the ancestral monocot were enriched in GO terms such as signal transduction and cell communication (Supplementary Fig. 14). For example, OG0000057, a disease resistance protein (TIR-NBS-LRR class) family, was contracted in the ancestral monocot (Supplementary Table 10), whereas OG0000047, which belongs to leucine-rich repeat protein kinases containing bacterium defence-related members, including IOS1 and FRK1, was significantly expanded in the ancestral monocot (Supplementary Table 11). However, by comparing the frequency distributions of (significantly) rapidly evolving OGs detected through CAFE analysis with the OG member size (average gene copy number per species), we observed that CAFE may be insensitive to detecting significant evolutionary changes for small gene families or OGs (Supplementary Fig. 15).

To circumvent this limit in detecting rapidly evolving OGs of smaller gene family sizes between monocots and other lineages of angiosperms, we further assigned changes based on a significant copy number difference with a \( P \) value threshold of \(<0.01\) (two-sided Mann–Whitney U-test) and a fold change of \( \geq 2\) in the average copy number between monocots and non-monocot angiosperms (Supplementary Table 7). Among the 429 OGs with significant copy number differences between monocots and non-monocot angiosperms, 247 OGs included 607 *Arabidopsis* genes, which could be used for a deep inference of functional categories according to The Arabidopsis Information Resource annotations (Supplementary Table 12). Intriguingly, by investigating these copy number-shifting OGs based on *Arabidopsis* GO annotations related to roots, cotyledons and leaves, we found that OG0011748, containing *Arabidopsis* *DOT3* (DEFECTIVELY ORGANIZED TRIBUTARIES 3), involved in vascular tissue and primary root development\(^{38}\), showed a significant reduction in gene copy number in monocots (Fig. 5a–c).

Through a detailed phylogenetic analysis of OG0011748, we found that *DOT3* was completely lost in waterlilies (*Nymphaea* and *Euryale*) and monocots, which both coincidentally showed ephemeral primary roots and palminate/parallel venation\(^{37}\) (Fig. 5c). The single-copy gene *dot3* (loss-of-function) mutants exhibited severely stunted primary roots, fusion of rosette leaves, freely ending vein loops in the cotyledons and parallel veins in *Arabidopsis* (Fig. 5b)\(^{39}\), which seem to be similar to phenotypes observed in monocots and waterlilies (*Nymphaeales*); therefore, their losses probably contribute to the unique leaf venation and root phenotypes in these two groups.

Because early-branching monocots, including Acorales and Alismatales, are mostly aquatic or wetland plants and show convergent evolution of many diagnostic traits in the aquatic family Hydrocharitaceae (*Nymphaeales*), it is believed that ancestral monocots had an aquatic or wetland origin\(^{31}\). Intriguingly, expansion of COG2132 (LOW PHOSPHATE ROOT1 (LPR1) and LOW PHOSPHATE ROOT2 (LPR2)), a group of multicomponent oxidases that play a key role in the redox signalling of *Arabidopsis* primary root growth regulated by antagonistic interactions of inorganic phosphate (Pi) and Fe availability\(^{40}\) (Fig. 6a), may have played a seminal role in the adaptation of monocots to aquatic habitats with low Pi availability, which is similar to the expansion of COG2132 in the aquatic eudicot *Nelumbo*\(^{40}\) (Fig. 6). We found that aquatic- or wetland-related lineages (*Nymphaea, Euryale, Acorus, Colocasia, Nelumbo, Oryza, Spirodela and Zosteria*) had higher copy numbers of this gene family than terrestrial plant lineages (two-sided
Mann–Whitney U-test, P < 0.01), which supported the hypothesis that the expansion of LPR1/LPR2 may have played a seminal role in the aquatic lifestyles of early monocots (Fig. 6b). Whereas five duplication events yielded six copies of LPR1/LPR2 in Acorus, two events occurred before monocot diversification and produced three ancient duplicates, all of which were retained in early-diverged aquatic/wetland monocots, including Acorus, seagrass and duckweed (Fig. 6c). These results support the hypothesis that the acquisition of functions drove the aquatic or wetland origin of the monocot ancestor.

Discussion

Early phylogenetic studies strongly supported Acorus as the earliest branching monocot, being sister to all the other extant monocots14,15. Our comparative analysis of the Acorus genome together with those of other monocots provided further insight into monocot evolution. By identifying only one single palaeopolyploid event during Acorus evolution, together with its extremely slow rates of sequence substitution and synteny loss, Acorus could be considered a pivotal genome for comparative genomics investigation among monocots (including grasses). Based on this reference, we showed a positive correlation between the synteny loss rate and genome duplication events in monocots and a particularly accelerated evolution rate of genes in the grass family. Polyploidization events are often associated with accelerated rates of species diversification and rapid gene turnover42 and adaptation during stressful periods in plants43. Within monocots, WGDs are more frequent in cereals (grass family)31. This agrees with our result of a positive correlation between the synteny loss rate and genome duplication events in monocots. Because a rapid substitution rate is often a signature of adaptation, whereas rapidly evolving genes also show neofunctionalization, this rapid rate probably facilitated adaptive radiation of grasses45. In addition, in terms of reproductive isolation, according to the reinforcement model of evolution, differentiation of karyotypes often enhances prezygotic isolation and
facilitates speciation\(^{46,47}\), as we observed in the grass family when compared with other monocots.

With the signature of the slowest evolving lineage, \textit{Acorus} is a good candidate for ancestral monocot genome reconstruction, similar to wax gourd (\textit{Benincasa hispida}) for Cucurbitaceae\(^{48}\) and \textit{Amborella trichopoda} for angiosperms\(^{11,15}\). This idea was supported by alignments with five representative outgroup taxa which indicated many more ancestral angiosperm genomic regions preserved in \textit{Acorus} than in all the other monocots surveyed. The ancestral genome is often assigned to a hypothesized 'median' genome that minimizes the genomic distance between two groups under the DCJ model\(^{49}\), such as the ancestral \textit{Brassica} genome\(^{50}\) and ancestral legume genome\(^{51}\). By including \textit{Acorus} and other early-diverging monocots, we successfully updated the AMK, which further evolved into the five protochromosomes of our previously predicted AMK by two fusions and one fission\(^{11}\). Given the lowest rate of synteny loss in \textit{Acorus} among the sequenced monocots when compared with five representative outgroup taxa, these results confirmed the hypothesis that \textit{Acorus} contains the most ancestral genome architecture/karyotype among all the sequenced monocots.

The rhythm of synteny (ancestral gene order) loss via gene deletion and chromosome reshuffling was highly heterogeneous among species and among different functional genes. In high-resolution analyses of genome-wide alignments among monocots and outgroups, we illustrated the complex genome evolutionary patterns during lineage diversification associated with gene-related traits. For example, we observed a negative correlation between higher TE density in syntenic gene-flanking regions and syntenic retention

---

**Fig. 6** | Duplications of the \textit{LPR1}/\textit{LPR2} family in the ancestral monocot associated with adaptation to an aquatic lifestyle. \textbf{a}, Illustration of the functional role of \textit{LPR1}/\textit{LPR2} in \textit{Arabidopsis} root growth under low Pi conditions according to previous studies. \textbf{b}, There is a significantly higher copy number of \textit{LPR1}/\textit{LPR2} in aquatic plants than in terrestrial plants. Two-sided Mann–Whitney \textit{U}-test significance is shown on the top of each violin plot (exact \textit{P} value). \textbf{c}, Three premonocot duplicates of \textit{LPR1}/\textit{LPR2} remained in \textit{Acorus} (shown in dotted circles). ANA, Amborellales, Nymphaeales and Austrobaileyales.
in monocots, which is probably mediated by the movement of TEs. Indeed, mobile elements are normally silenced by epigenetic mechanisms due to their destructive potential. However, they can often be reactivated in the face of environmental stress and participate widely in chromosomal structural variation as well as genome instability. For example, in *Oryza*, sequence rearrangements are observed more frequently in repetitive regions56, which is in line with our results. Moreover, we observed that disrupted synteny during monocot evolution is associated with both the expression level and breadth (inverse of tissue specificity) of a gene. For example, a human–chimpanzee comparative study showed that chromosomal rearrangements, which disrupt synteny, are associated with elevated gene expression differences in the brain51. In *Brassica*, homoeologous chromosome rearrangements drive gene expression change in newly resynthesized *Brassica napus* allotriploids57. This could be appropriately addressed by changes in the cis-environment of a gene and considerable gene structural mutations, such as unpredictable sequence translocation or inversion when synteny is degraded by complex genetic forces as a whole58. In a commercial wine yeast strain, an inversion that involves *SSU1* and *GCRI* regulatory regions can activate *SSU1* expression; thus, this inversion facilitates sulphite resistance59. Another example in maize shows that an inversion in the *Tsu1* mutant with a breakpoint in the promoter of *Zmm19* significantly changes *Zmm19* expression, leading to kernels being completely enclosed in leaflike glumes59. Therefore, the genomic position is critical to gene expression. However, co-expressed gene clusters can often be observed in syntetic blocks in mammals60 but not in *Drosophilia melanogaster*61 or *Arabidopsis*, which probably differ in their constraints on development. However, future studies to test the relationship between co-expression and synteny conservation are needed in different plant species, particularly monocots and cereals. On the other hand, our results also showed that the genes from the outgroup (Nelumbo) with higher synteny retention in monocot species exhibit lower nucleotide diversity. This might be attributable to the functional constraints that play an important role in maintaining synteny because rearrangement can have an impact on gene expression62 and the abnormal chromosomal pairing and recombination of non-homologous regions can lead to copy number variation or gene loss. Collectively, the gene features observed here shed new light on the intricate evolutionary history of monocot families.

A deep investigation into genome evolution has allowed us to reveal the role of gene copy number variation in specific traits. For example, changes in the MADS-box regulatory gene family related to flower diversity63 and massive gene loss in *Cassutia australis* associated with its parasitic lifestyle64 have been reported. In our study, we inferred that substantial gene families probably drove traits associated with the emergence of monocots during flowering plant evolution. Our results displayed a significantly higher copy number of LPR1/LPR2 in aquatic plants than in terrestrial plants, which is consistent with previous findings in *Nelumbo*41. The expansion of LPR1/LPR2 is believed to be associated with a low-phosphate aquatic environment, especially in low Pi conditions41. Phosphorus (Pi) is one of the major nutrient limitations in many freshwater ecosystems, including streams and wetlands65. In *Arabidopsis*, *LPR1* and its homologue *LPR2* regulate root meristem activity related to Pi availability66. Although low Pi can inhibit primary root growth in wild-type *Arabidopsis*, increasing the gene products of *LPR1* by overexpression can further inhibit primary root growth under low Pi conditions; by contrast, the loss-of-function *lpr1* double-mutant showed enhanced primary root growth under low Pi conditions67. In *Nelumbo*, the increased copies of *LPR1/LPR2* were found to be highly expressed in its lateral and adventitious root primordia68. All these results probably suggest a shared evo–devo strategy in both early-branching aquatic monocots and other aquatic angiosperms to form ephemeral primary roots instead of taproots in response to low Pi in streams or wetlands69. Moreover, by utilizing lateral spreading, together with the development of adventitious roots, early monocots can adapt to wetland habitats with differential moisture contents close to that of the Earth’s surface70. Apart from *LPR1/LPR2*, we also found that losses of the non-monocot-conserved *DOT3* gene in monocots were linked not only to the emergence of ephemeral primary roots, but also to parallel venation in these clades71. Finally, we revealed that WOX1, an essential gene that regulates cotyledonal primordia initiation72,73, is completely lost in monocots, suggesting that an ancient loss occurred before modern monocots diverged. Such loss is very probably attributed to the formation of the single cotyledon character that is unique to monocots, which still needs more studies to be further investigated.

**Methods**

**Plant material, genome sequencing and RNA-seq of *Acorus*.** *Acorus tatarinowii* (NCBI Taxonomy ID: 123564) was collected from Shennongjia Nature Reserve (Hubei, China). DNA from leaves was extracted using Plant DNA Isolation Reagent (TIANGEN). For genome size estimation, genomic DNA was sheared into 250–280 bp fragments with the NEBNext Ultra DNA Library Prep Kit for Illumina (NEB) based on the manufacturer’s protocol. Paired-end reads (150 bp for each end) were sequenced on the Illumina HiSeq 4000 platform. DNA libraries were constructed based on the PacBio library preparation protocol and further sequenced on the PacBio Sequel platform (Pacific Biosciences) with the Sequel II Binding Kit 1.0, Sequel II Sequencing Kit 1.0 and Sequel II SMRT Cell 8M at Fraseragen. Subread data was obtained via SMRT LINK 7.0. Subreads with a quality score below 0.8 were excluded. The Hi-C DNA library was prepared at Fraseragen using a previously published protocol74. Generally, nuclear DNA was cross-linked inside tissue cell samples of young leaves. The extracted DNA was further digested using the restriction enzyme Mbol. Biotinylation was tagged at both sticky ends of the digested DNA fragments and then ligated randomly after dilution. The condensed, sheared and biotinylated DNA fragment libraries were prepared for paired-end sequencing with a 150-bp read length on an Illumina HiSeq platform.

**Chromosomal-level assembly of *Acorus tatarinowii*.** The genome size and sequence assembly of *Acorus* were annotated by jellyfish74, an assembly–estimator using k-mer frequency distribution (k-mer = 17 as the default) based on Illumina reads with default settings. For genome assembly, Nextdenovo software v.2.5.0 was applied for PacBio reads with the following parameters: read_cutoff = 1k; seed_cutoff = 40,150; blocksize = 1g; pa_correction = 2; seed_culltime = 2; sort_options = m -4 -g -l 10 -k 50; minimap2_options_raw = -x ava-pb -l 18; correction_options = -p 15. The corrected PacBio reads were further trimmed and assembled by Canu v.2.2 with the trimming parameters ‘genomeseqSize = 400m; correctedErrorRate = 0.12; corMaxEvidenceErate = 0.15; minReadLength = 1,000; minOverlapLength = 500; merylThreads = 40’ and the assembling parameters ‘genomeseqSize = 400m; maxThreads = 60; correctedErrorRate = 0.035’ (https://github.com/Nextenovo/NextDenovo). After mapping PacBio reads onto the polished contigs, redundant contigs were removed by purge_haplotigs based on read coverage. The Hi-C sequencing reads were aligned to the final contigs by BWA-MEM75. Finally, scaffolding of these contigs into pseudochromosomes was performed with LACHESIS76. Juicer was applied to construct high-resolution transcriptome sequencing, total RNA of young leaves and cotyledon character that is unique to monocots, which still needs more studies to be further investigated.

**Repeat, gene and functional annotations.** Before gene annotation, repeat sequences including TEs on the chromosome-level assembly were de novo predicted using Extensive de novo TE Annotator (EDTA, v.1.8.4) with default settings, and annotated by RepeatMasker (http://www.repeatmasker.org). Genes were predicted by combining: (1) RNA-seq evidence, (2) protein homology and (3) ab initio prediction. For gene annotation with transcriptional evidence, RNA-seq reads from our newly sequenced young leaf, old leaf, root and publicly available rhizome and leaf data (accession no. SR99647976 and SR99647979) were aligned by the HISAT2 pipeline to obtain transcript-based annotation77. CDSs were predicted using Transdecoder (https://github.com/TransDecoder). In addition, the de novo transcriptome assembly was trimmed by Trinity with default settings (https://github.com/trinityrnaseq/trinityrnaseq); PASA, which
integrated the de novo transcript assemblies, was applied to further update the assembly with default settings (https://github.com/PASApipeline/PASApipeline). Homology-based gene annotation was conducted using Genewise software with genome sequences and gene annotations from Acorus and seagrass, including Colocasia esculenta (accession no. ASNM445464v1), Zea mays (no. B73 RollingGen_v4), Oryza sativa (no. GCF_0000054325) and Zostera marina (no. GCA_001185155.1)\(^6\). Ab initio gene prediction was conducted using AUGUSTUS\(^7\) and GeneMark-ES\(^8\). The final consensus gene annotations were generated by EvidenceModeler with different weights among annotations (RNA-seq > gene homology > ab initio)\(^6\). Finally, protein-coding genes with more than 30% of the CDS overlapping with repeat sequences were considered repeat- or transposon-related genes and were discarded from downstream analyses. GO functional annotations were inferred using the ‘non-redundant’ database of plants in eggNOG 4.5 with default settings\(^9\).

**AMK reconstruction.** Ancient genomes are reconstructed in a six-step method as illustrated in Fig. 4a. The first step consists of aligning the genes (protein sequences) using BlastP with thresholds for cumulative identity percentage (CIP) ≥ 50% and cumulative alignment length percentage BLAST parameters (CALP) ≥ 50% (defined in Salz et al.\(^3\)) (https://github.com/nelmolublatae/amp_article/blob/main/6.CIP_CALP.ppl), which deliver conserved genes between the investigated species given the following formulas:

\[
\text{CIP} = \frac{\text{nb ID by (HSP/AL)} \times 100}{\text{Query length}}
\]

where CIP corresponds to the cumulative percentage of sequence identity observed for all the high-scoring pairs (HSPs) divided by the cumulative aligned length (AL) which corresponds to the sum of all HSP lengths. The ‘nb’ denotes number.

\[
\text{CALP} = \frac{\text{AL}}{\text{CIP}}
\]

where CALP is the sum of the HSP lengths (AL) for all HSPs divided by the length of the query sequence. With these parameters, BLAST produces the highest cumulative percentage identity over the longest cumulative length, thereby increasing stringency in defining conserved genes between two genome sequences\(^2\). The second step consists of removing species-specific and local (tandem) duplicates and retaining only the single-copy orthologues, which will reveal that proteogens conserved in all the investigated species or between a subset (at least two) of the investigated species. This step consists in extracting one-to-one gene relationships between species from the step 1 output file. The step three consists of clustering of chaining groups of conserved genes into synteny blocks (SBs). The third step consists of extracting all combinations of chromosome-to-chromosome relationships (for SBs sharing more than five orthologous genes) from the step 2 output file (or alternatively using tools such as DRIMM synteny software\(^2\)). In the fourth step, SBs from the previous output file are then merged into ancestral syntegomes (also referred to as CARs). This step consists of defining independent groups of SBs sharing synteny between the modern species investigated (or alternatively with tools such as MGRA software\(^2\) or ANGES software\(^2\)). The fifth step corresponds to CAR validation, in which CARs correspond exclusively to diagonals in dotplot-based comparative genomics deconvolutions of the synteny between the investigated species. Finally, the sixth step consists of deriving a parsimonious evolution model by introducing the smallest number of rearrangements (fissions, fusions and translocations) to explain the transition between the ancestral and modern genomes. This strategy has been previously applied to infer a pre-\(\tau\) AMK structured into 5 protochromosomes with 6,707 genes (available in Supplementary Table 3 from Murat et al.\(^1\)) at the MRCA of Anura\(^1\), Dioscorea\(^1\), and grasses. In the current study, we used this \(n = 5\) AMK, as a pivot to compare, in a BlastP and dotplot-based approach, the modern karyotypic structures of the Acorus genome and other early-branching monocot genomes, including Spirodela polyrhiza, Colocasia esculenta and Dioscorea (alata and rotundata). From the gene (protein sequences) alignments using BlastP, and CIP and CALP values obtained from the pre-\(\tau\) AMK, Colocasia and Dioscorea, stored in a tabular format to further extract from it conserved genes (step 1), one-to-one gene orthologous relationships (step 2), SBs (step 3) and CARs (step 4), as well as dotplot illustrations of the synteny between the investigated species, we proposed the karyotypic structures of the ancestral monocots (Step 5) and inferred an evolutionary scenario taking into account the lowest number of genomic rearrangements (including inversions, deletions, fusions, fissions and translocations) that may have occurred between the AMK and modern monocot genomes. All data described in this current study, such as conserved genes, SBs and ancestral chromosomal blocks, are available in Supplementary Tables 5 and 6.

**Gene and WGD analyses.** To identify the origins of genes from duplications and WGDs in Acorus, intraspecific and interspecific SBs were identified by McScan via JCVI\(^1\). To determine the WGDs in relation to the divergence of rice, asparagus and seagrass, 40-fold values for all syntenic paralogous pairs or orthologous pairs were estimated and corrected for possible multiple transversions at the same site according to a previous method\(^1\). \(K_v\) values of all syntenic paralogous/orthologous pairs were also calculated by codeML of the PAML package\(^1\). Histograms of 4dTv and \(K_v\) values for all syntenic paralogous/orthologous pairs were plotted with a bin size of 0.01 (Fig. 2a) using the alignments of Acorus\(^1\) and seagrass, raw 4dTv values for all syntenic paralogous pairs or orthologous pairs were discarded from downstream analyses. GO functional annotations were inferred using the ‘non-redundant’ database of plants in eggNOG 4.5 with default settings\(^9\).

**Synteny retention among genes.** To estimate the variation in synteny retention among different genes during monocot radiation, we used the outgroup taxon Nelumbo nucifera because of its greatest similarity of synteny structure in relation to monocots, and the availability of required datasets including whole-genome methylated genomic resequencing and expression profiles of all organs and tissues\(^1\)–\(^3\). The 29,582 Nelumbo genes sharing homologue(s) (BlastP \(E\) value < 10\(^{-10}\)) with at least 1 of the 42 monocots were used for the following analysis of synteny retention rates. For each Nelumbo nucifera gene, the number of monocots showing a synteny relationship was used to represent its relative synteny retention rate. To gain information about the number of gene duplications, nucleotide diversity, CDS length, gene length, the number of predicted PPIs, the average expression level, expression specificity, TE density and methylation levels on genes and flanking regions were obtained from our previous study\(^5\). Among the types of gene duplications, WGD genes (genes retained after ancient WGD)\(^4\), tandem duplicates (tandemly duplicated genes), single-copy genes (genes without homologues within Nelumbo), proximal duplicates (duplicated having one or a few intervening genes) and WGD&bndem duplicates (genes that underwent both WGD and tandem duplications) were classified using MCscanX in our previous study\(^5\). While the two-sided Mann–Whitney \(U\) test was applied to compare relative synteny retention rates of WGD&bndem duplications with non-WGD&bndem duplicates (tandem, proximal, single-copy and dispersed), Pearson correlations were calculated between synteny retention rates and different factors, such as gene length, for all Nelumbo nucifera genes using R (https://www.r-project.org/). Meanwhile, Nelumbo genes sharing homologue(s) with at least one monocot were further divided into four groups following a decreasing number of monocots with synteny.
retention. Levels of each gene trait among groups I, II, III and IV were compared using the Kruskal–Wallis test in GraphPad Prism v.9.

Evolution of functional genes at the emergence of monocots. To gain insight into OG evolution in the ancestral monocot, 28 representative taxa, including early-branching angiosperms, monocots and eudicots, were used for comparisons. First, OGs were obtained via OrthoFinder37. Single-copy genes identified from OGs were aligned using I-TASSER with the parameters described above. The species tree rooted with Ginkgo was used as an input to build an ultrametric tree (chronogram) by r8s, whereas fossil constraints were set to Arabidopsis–Nymphaea (125–247.2 Ma), Arabidopsis–Liriodendron (125–247.2 Ma), Arabidopsis–Oryza (125–247.2 Ma) and Arabidopsis–Aquilegia (~128.6 Ma) according to a previous study38. To estimate OG gain and loss along the ultrametric tree, we applied Dollo-Parsimony via COUNT software with default settings97.

To estimate the number of OGs with significant expansion and contraction along the ultrametric tree, CAFE was applied with a P value threshold of 0.05 (ref. 39). In parallel, to better detect OGs with significant copy number differences between monocots and non-monocot angiosperms, the copy numbers of these two clades were compared using the Mann–Whitney U test for each OG. OGs with a P value <0.01 and fold change of the average copy number ≥2 were considered significantly different.

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

Data availability. The datasets generated and analysed during the current study including PacBio Sequel II, Illumina, Hi-C data, genome assembly, annotation and RNA-seq reads have been deposited in China National GeneBank (CNGB, https://db.cngb.org/) under accession number CNP0001708. Public transcripts used in this study are available from NCBI under the accession number SRR0944796 and SRR0944797.

Source data are provided with this paper.

Code availability. The main custom scripts and workflow have been deposited in Github (https://github.com/nelumbolutea/amk_article).

Received: 27 October 2021; Accepted: 30 May 2022; Published online: 14 July 2022

References

1. Givnish, T. J. et al. Monocot plastid phylogenomics, timeline, net rates of species diversification, the power of multi-gene analyses, and a functional monophyly for the origin of monocots. Am. J. Bot. 105, 1888–1910 (2018).
2. Friis, E. M., Pedersen, K. R. & Crane, P. R. Araceae from the Early Cretaceous of Portugal: evidence on the emergence of monocotyledons. Proc. Natl Acad. Sci. USA 101, 16565–16570 (2004).
3. Bremer, K. Early Cretaceous lineages of monocot flowering plants. Proc. Natl Acad. Sci. USA 97, 4707–4711 (2000).
4. Coffard, C., Kardjilov, N., Manke, I. & Bernardes-de-Oliveira, M. E. C. Fossil evidence of core monocots in the Early Cretaceous. Nat. Plants 5, 691–696 (2019).
5. Zeng, L. et al. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. Nat. Commun. 5, 4956 (2014).
6. Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L. & Hernández-Hernández, T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytol. 207, 437–453 (2015).
7. Duvall, M. R., Luehr, G. H. Jr, Eguiarte, L. E. & Clegg, M. T. Phylogenetic analysis of rbcL sequences identifies Acorus calamus as the primordial extant monocotyledon. Proc. Natl Acad. Sci. USA 90, 4641–4644 (1993).
8. Chase, M. W. Monocot relationships: an overview. Am. J. Bot. 91, 1645–1655 (2004).
9. Tang, H., Bowers, J. E., Wang, X. & Paterson, A. H. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc. Natl Acad. Sci. USA 107, 472–477 (2010).
10. Jiao, Y., Li, J., Tang, H. & Paterson, A. H. Integrated syntic and phylogenomic analyses reveal an ancient genome duplication in monocots. Plant Cell 26, 2792–2802 (2014).
11. Soltis, P. S. & Soltis, D. E. Ancient WGD events as drivers of key innovations in angiosperms. Curr. Opin. Plant Biol. 30, 159–165 (2016).
12. Shi, T. & Chen, J. A reappraisal of the phylogenetic placement of the Aquilegia whole-genome duplication. Genome Biol. 21, 295 (2020).
13. Murat, F., Armero, A., Pont, C., Kloppe, C. & Salse, J. Reconstructing the genome of the most recent common ancestor of flowering plants. Nat. Genet. 49, 490–496 (2017).
73. Liu, B. et al. Estimation of genomic characteristics by analyzing k-mer distribution in de novo genome projects. *Nat. Biotechnol.* 31, 1119–1125 (2013).

74. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* 3, 95–98 (2016).

75. Ou, S. et al. Benchmarking transposase element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 20, 275 (2019).

76. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915 (2019).

77. Perti, M. et al. StringTie improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295 (2015).

78. Birney, E., Clamp, M. & Durbin, R. *Gene*Wise and Genowise. *Genome Res.* 14, 988–995 (2004).

79. 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).

80. Besemer, J. & Borodovsky, M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* 33, W451–W454 (2005).

81. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* 9, R7 (2008).

82. Besse, J. et al. eggnOGF 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* 44, D286–D293 (2016).

83. Salse, J., Abrouk, M., Murat, F., Quraishi, U. M. & Feulillet, C. Improved criteria and comparative genomics tools provide new insights into grass paleogenomics. *Brief. Bioinform.* 10, 619–630 (2009).

84. Pham, S. K. & Pevzner, P. A. DRIMM-Synteny: decomposing genomes into ancestral genomes maps. *Bioinformatics* 26, 2509–2516 (2010).

85. Lin, C. H., Zhao, H., Lowcay, S. H., Shahab, A. & Bourque, G. webMGR: an online tool for the multiple genome rearrangement problem. *Bioinformatics* 26, 408–410 (2010).

86. Jones, B. R., Rajaraman, A., Tannier, E. & Chauve, C. ANGES: reconstructing Ancestral Genomes maps. *Bioinformatics* 28, 2388–2390 (2012).

87. Tang, H. et al. Synteny and collinearity in plant genomes. *Science* 320, 486–488 (2008).

88. Tang, H. et al. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* 18, 1944–1954 (2008).

89. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591 (2007).

90. Garsmeur, O. et al. Two evolutionarily distinct classes of paleopolyploidy. *Mol. Biol. Evol.* 31, 448–454 (2014).

91. Wang, H. et al. GC gene body DNA methylation changes and evolution of duplicated genes in cassava. *Proc. Natl Acad. Sci. USA* 112, 13729–13734 (2015).

92. Edger, P. P., McKain, M. R., Bird, K. A. & VanBuren, R. Subgenome assignment in allopolyploids: challenges and future directions. *Curr. Opin. Plant Biol.* 42, 76–80 (2018).

93. Efremov, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238 (2019).

94. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534 (2020).

95. Cusí, Ó. et al. Evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 26, 1910–1912 (2010).

96. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271 (2006).
Competing interests
The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41477-022-01187-x.
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-022-01187-x.
Correspondence and requests for materials should be addressed to Jinming Chen, Jérôme Salse or Qingfeng Wang.
Peer review information Nature Plants thanks the anonymous reviewers for their contribution to the peer review of this work.
Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | Genome-wide Hi-C interaction heatmap. Genome-wide Hi-C interaction heatmap of Acorus (resolution: 500 kb).
Extended Data Fig. 2 | Subgenome fractionation of Acorus by comparing to outgroups. Differences in the number of collinear genes between LF (less fractionated) and MF (more fractionated) blocks when comparing to outgroup species, Amborella (A), Aristolochia (B), Aquilegia (C), Nelumbo (D), Spirodeia (E) and Elaeis (F).
Extended Data Fig. 3 | Syntenic gene retention in five outgroups. Comparison of the numbers of syntenic anchor genes in five outgroups (Amborella, Nymphaea, Aristolochia, Cinnamomum and Nelumbo) in relationship to monocot genome assemblies.
Extended Data Fig. 4 | Negative correlation between syntenic gene retention and paleopolyploidies (ancient WGDs). Significantly negative correlation (calculated by Pearson’s correlation) between the number of syntenic genes in *Amborella* (A), *Nymphaea* (B), *Cinnamomum* (C), *Aristolochia* (D) and *Nelumbo* (E) and the expected copy number of genes after paleopolyploidizations. The error bands represent 95% confidence intervals based on a binomial model.
Extended Data Fig. 5 | Syntenic gene retention and gene features. Violin plots showing different levels of CDS length (A), gene length (B), number of predicted protein-protein interactions (C), expression level (D), gene-upstream TE density (E), genic-region TE density (F), gene-downstream TE density (G), gene methylation (H) and nucleotide diversity (I) for *Nelumbo* genes from those with the greatest number of monocot species being syntenic (I) to those with the minimum (IV). One-way Kruskal-Wallis test significance is shown on the top of each violin plot (adjusted P values).
Extended Data Fig. 6 | Synteny between Acorus and the $n=5$ pre-$\tau$ AMK (from Murat et al.\textsuperscript{13}). CENTRE-The dotplot-based deconvolution of the synteny between Acorus (y-axis) and the $n=5$ pre-$\tau$ AMK (x-axis) defines 12 independent pairs of duplicated blocks covering the entire Acous genome (highlighted in rectangles), suggesting 12 CARs (referenced to as AMK1-1’-2-2’-3-3’-4-4’-5-5’-6-6’) at the basis of the speciation between Acorus and $n=5$ AMK (or any species within the $\tau$-WGD lineage). LEFT-From this ancestral state of 12 protochromosomes, the Acorus genome has been shaped through a lineage-specific WGD to reach a $n=24$ chromosomes intermediate, followed by 12 fusions to reach the 12 modern chromosomes. BOTTOM-From this ancestral state of the 12 chromosomes, the reported $n=5$ pre-$\tau$ AMK (Murat et al.\textsuperscript{13}) has been shaped through 6 ancestral chromosome fusions to reach an $n=6$ AMK intermediate (represented by six colors including orange, dark blue, pink, light green, light blue, and dark green) followed by one fission (dark green) and two fusions (dark green-orange, dark green-light blue) explaining the transition between the $n=6$ AMK and the previously reported $n=5$ pre-$\tau$ AMK (Murat et al.\textsuperscript{13}) at the most recent common ancestor of Ananas, palm and grasses.
Extended Data Fig. 7 | Evolutionary scenario of the monocot karyotypes. The figure illustrates the ancestral monocot karyotypes with all the proposed rearrangements (fusions, fissions) that shaped the modern genomes with the evolution of the number of chromosomes (in green circles) compared to what proposed in Xu et al.36 (in red circles).
Extended Data Fig. 8 | The pattern of WGDs in monocots. Dotplots illustration of the synteny between the reconstructed ancestral monocot karyotype (n = 6 AMK, x-axis) and modern species: Ata (Acorus tatarinowii), Cesu (Colocasia esculenta), Spo (Spirodela polyrhiza), Drot (Dioscorea rotundata), Dalata (Dioscorea alata) and AGKpre: pre-WGD (p) ancestral grass karyotype (n = 7) (y-axis). Signatures of reported WGD events are illuminated with red dots on the dotplots.
Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies 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 of collecting the data.

Data analysis

- For Plant Material, PacBio Sequel, Hi-C Sequencing and RNA-seq of Acorus tatarinowii
  SMRT LINK v7.0

- For genome size estimation:
  Jellyfish v2.3.0
  GCE v1.0.2

- For Chromosomal-level assembly of Acorus tatarinowii
  Nextdenovo v2.5.0
  Canu v2.2
  BWA v0.7.17
  LACHESIS v2017-12-21
  JucleBox v2.1.10.

- For Repeat, gene and functional annotations
  EDTA v1.8.4
  RepeatMasker v4.1.2
  HISAT2 v2.2.1
  StringTie v2.1.5
  Trinity v2.13.2
  PASA v2.3.1
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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The datasets generated and analyzed during the current study including PacBio Sequel II, Illumina, Hi-C data, genome assembly, annotation, and RNA-seq reads have been deposited in China National GeneBank (CNGB, https://db.cngb.org/) under accession number CNP0001708. Public transcriptomes used in this study are available from NCBI under the accession number accession Nos. SRR9644796 and SRR9644797.

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-list.pdf

Life sciences study design

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

**Sample size**

For genome sequencing and assembly of Acorus, only one individual is used to ensure the sample purity, and further low heterozygosity estimated by Kmers ensured the accuracy for genome assembly, which successfully allowed us produce a high-quality assembly. For RNA-seq, since the purpose is to confirm the expression bias towards LFs (subgenome dominance) being consistent among different tissues, a total of five tissue RNA-seq data representing different tissue types were considered as tissue replicates, and finally all tissue samples successfully concluded the same trend of LF > MF in expression.

**Data exclusions**

For genome sequencing, Sequel II Subreads with a quality score below 0.8 were excluded.

**Replication**

For genome sequencing, Sequel II generated data of 250 fold length of the Acorus genome size, which is enough to obtain a high-quality genome. For subgenome dominance tested by RNA-seq of tissues, a total of five tissues were considered as five tissue replicates, and all five samples successfully revealed the same consistent trend of LF > MF in overall expression. For phylogenetic tree of monocots and outgroups, 1000 bootstraps were used and we successfully obtained a species tree with high-confidence support.

**Randomization**

No randomization was applied in this manuscript since the genome assembly was not allocated into experimental groups.

**Blinding**

No blinding was applied in this manuscript since the genome assembly was not allocated into experimental groups.
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 | Methods |
|----------------------------------|---------|
| n/a | Involved in the study |
| ☒ | Antibodies |
| ☒ | Eukaryotic cell lines |
| ☒ | Palaeontology and archaeology |
| ☒ | Animals and other organisms |
| ☒ | Human research participants |
| ☒ | Clinical data |
| ☒ | Dual use research of concern |
| n/a | Involved in the study |
| ☒ | ChIP-seq |
| ☒ | Flow cytometry |
| ☒ | MRI-based neuroimaging |