

Supplementary Materials for

Genomic and ecological drivers of parallel arid adaptation in tree grapes (Vitaceae)

Jinren Yu^{1,2,3,4}, Ju Zhou^{1,2,3}, Shanshan Luo^{1,2,3}, Ilia J. Leitch⁴, Kyaw Nyein^{1,2,3}, Rindra Manasoana Ranaivosoa⁵, Chuanyu Du^{1,2,3}, Russell L. Barrett^{6,7}, Jie Cheng^{1,2,8}, Chaobin Li^{1,2}, Yang Dong^{1,2}, Romer Narindra Rabarijaona^{1,2}, Alexandre Antonelli^{4,9,10,11}, Zhiduan Chen^{1,2}, Limin Lu^{1,2*}

Corresponding author: Limin Lu, liminlu@ibcas.ac.cn

Table of Contents

Supplementary Note 1. Ancient whole genome duplications in Vitaceae.

Supplementary Note 2. Types of leaf surface in *Cyphostemma*.

Supplementary Figures 1–25

Supplementary References

Supplementary Note 1. Ancient whole genome duplications in Vitaceae.

We conducted synteny analysis and examined 4DTv distributions for Vitaceae species, using *Arabidopsis thaliana* as an outgroup, to investigate ancient whole genome duplication (WGD) events in the family. Our syntenic analysis confirms that no additional WGD events occurred following the divergence of *Vitis* and Cayratieae. The 4DTv distributions for *C. dehongense* and *C. currorii* exhibit peaks at identical 4DTv distances, indicating a shared duplication event. Given that *Vitis* is reported to have experienced only the γ triplication shared by core eudicots¹, and that no subsequent WGDs have been detected after its divergence from *Cyphostemma*, we interpret these shared peaks as signatures of the γ event.

Notably, the positions of the 4DTv peaks suggest a WGD event occurring after the divergence of *Cyphostemma* and *Arabidopsis*, which appears inconsistent with the conventional position of the γ event. We propose that this discrepancy reflects the two-step process of the γ event (Supplementary Fig. 9c). Under this model, an initial WGD was followed by a second polyploidization involving a third genome, potentially derived from a “ghost lineage” that diverged prior to the initial tetraploidization². In this scenario, *Arabidopsis* and *Cyphostemma* diverged at Node 1 in Supplementary Fig. 9c, whereas the second step of the γ event occurred after their divergence. This would result in signals of gene duplication after Node 1 and thereby explain the observed peaks in the 4DTv distributions without invoking a post-divergence WGD in Vitaceae (Supplementary Fig. 9b).

Supplementary Note 2. Types of leaf surface in *Cyphostemma*.

We classified the leaf surfaces of succulent *Cyphostemma* species into four different types according to trichomes, wax on the surfaces, and shape of epidermal cells and stomata

(Supplementary Fig. 19). Type I in the continental African clade exhibits similar morphology of epidermal cells and stomata on both adaxial and abaxial sides, with long trichomes and wax coating around the stomata, which sometimes cover the entire leaf surface. Type II, which is also in the continental African clade, has stomata of smaller sizes and epidermal cells on both sides but with thicker cell walls on the adaxial side, and this type is featured by diverse furcate trichomes on both sides. Type III in the Malagasy clade has larger cells and thicker cell walls on the adaxial side and wax coating around stomata on both sides, with almost no trichomes. Type IV in the Malagasy clade also has larger cells and thicker cell walls on the adaxial side, and it has remarkably compact stomata on the abaxial side, with dense, segmented trichomes on both sides. These traits can be clearly observed in the scanning electron microscopy photos of *Cyphostemma* species deposited in figshare.

Continental African clade



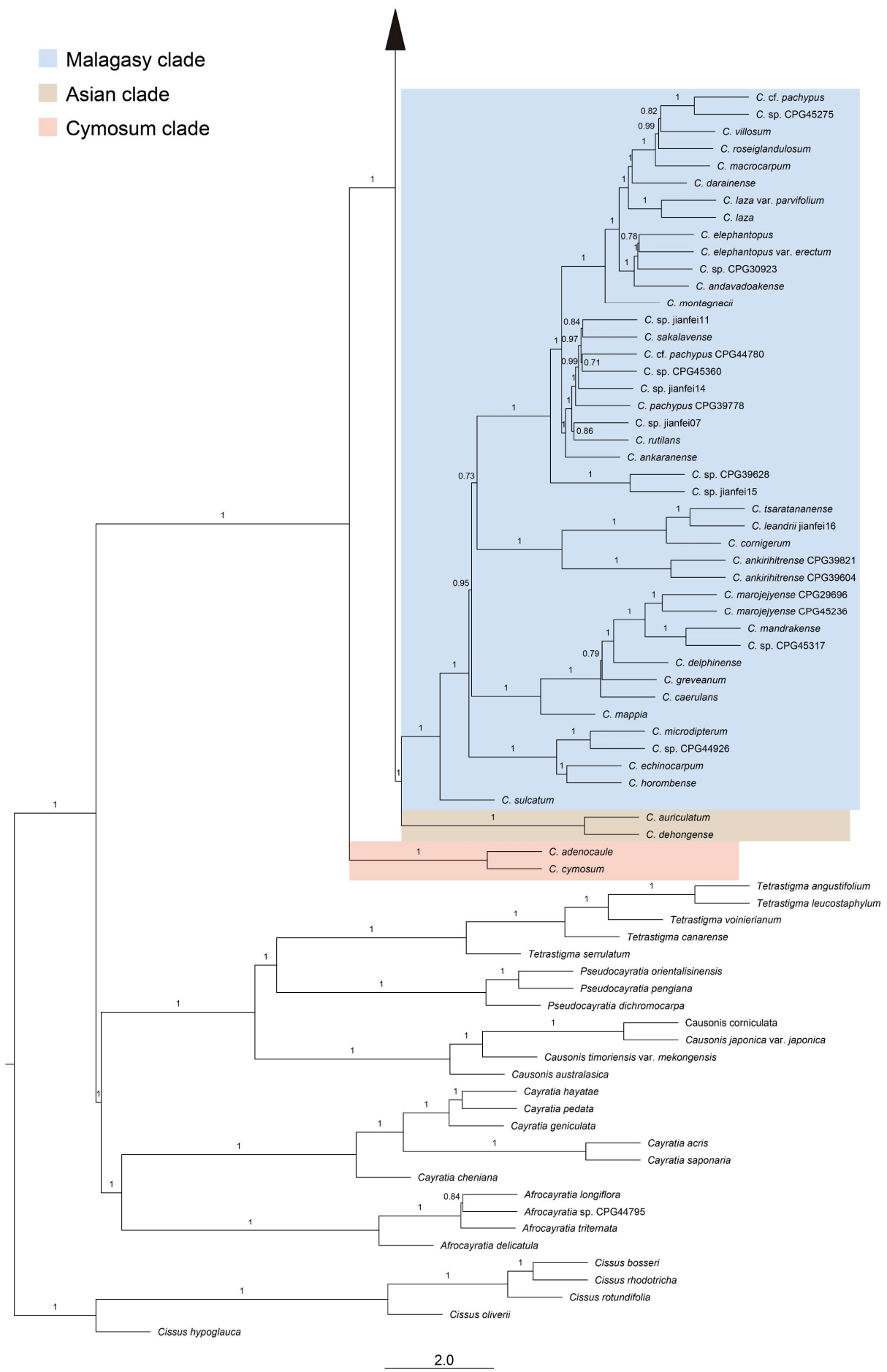
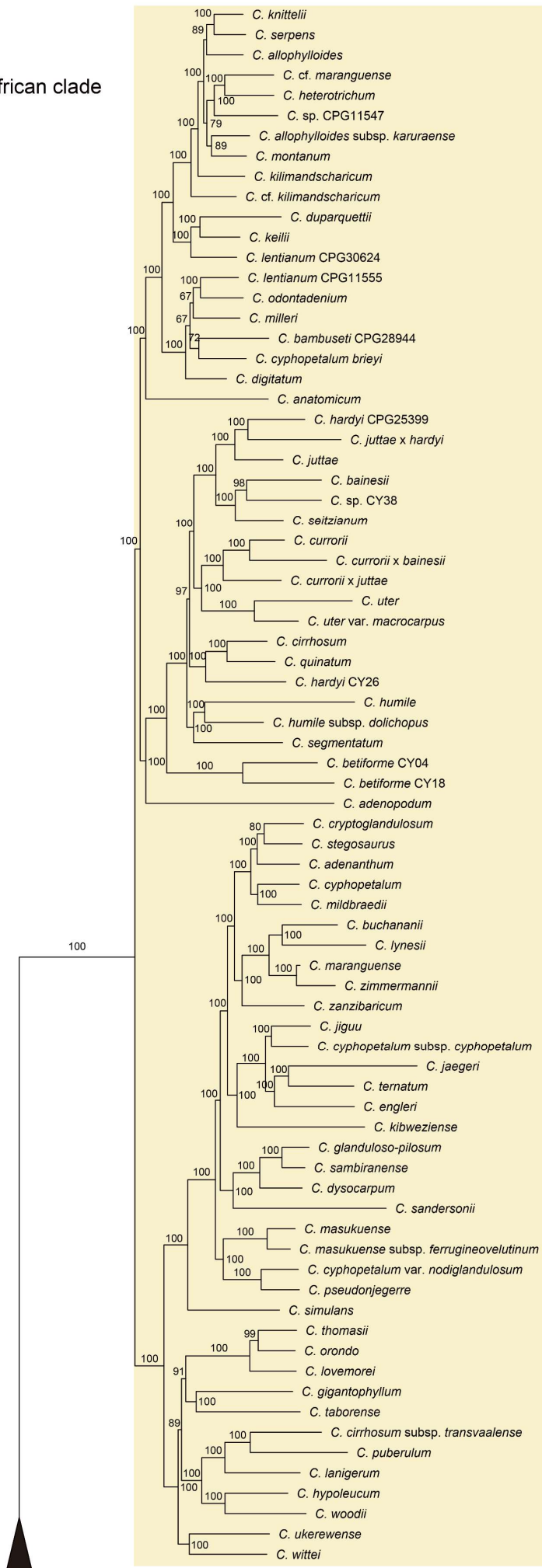


Fig. S1 | Phylogeny of *Cyphostemma* based on 1376 single-copy orthologs using the multispecies coalescent method. Local posterior probability support values are shown above the branches. The four main clades of *Cyphostemma* are indicated with different colors. The black arrows indicate a continued phylogeny.

Continental African clade



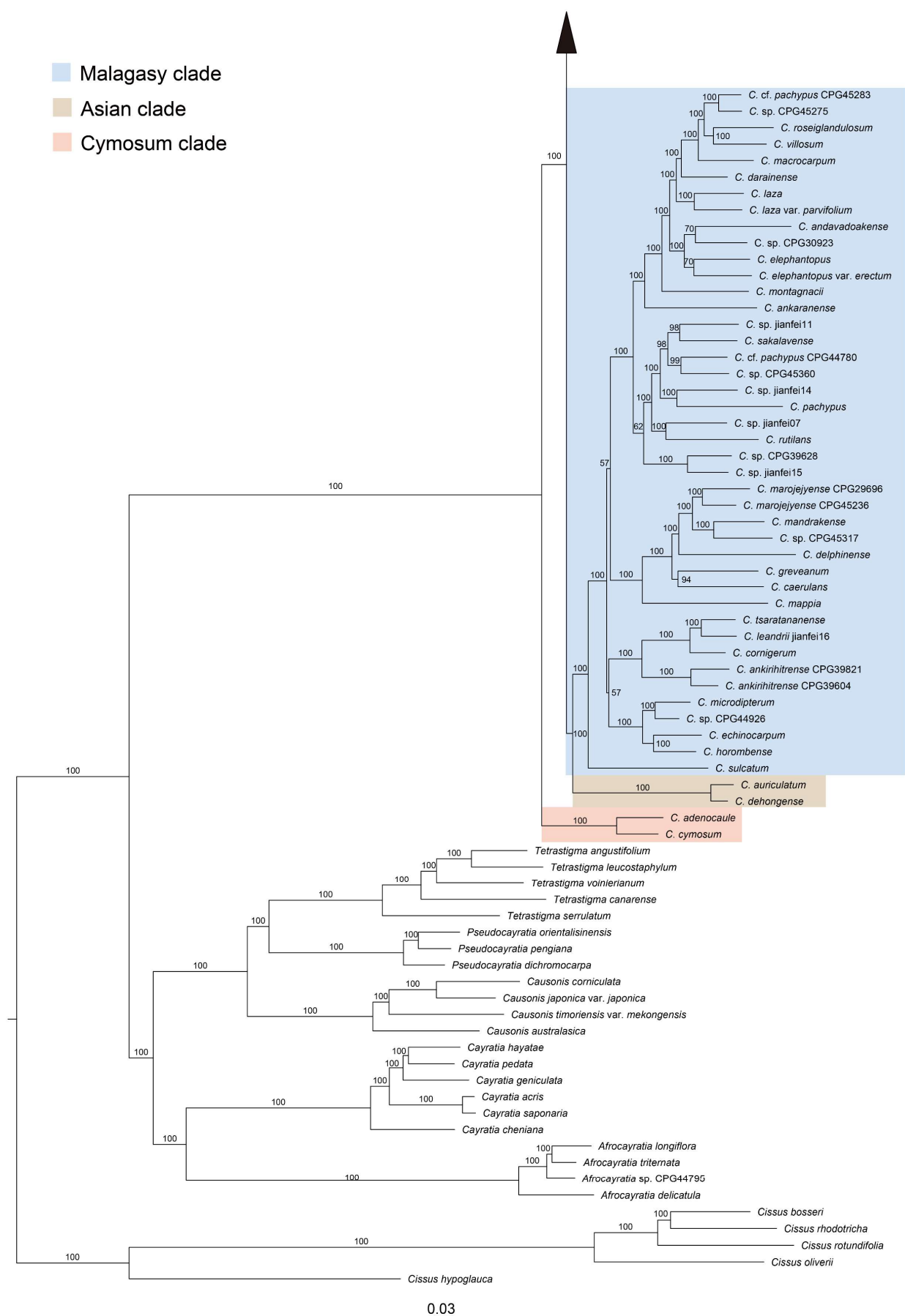
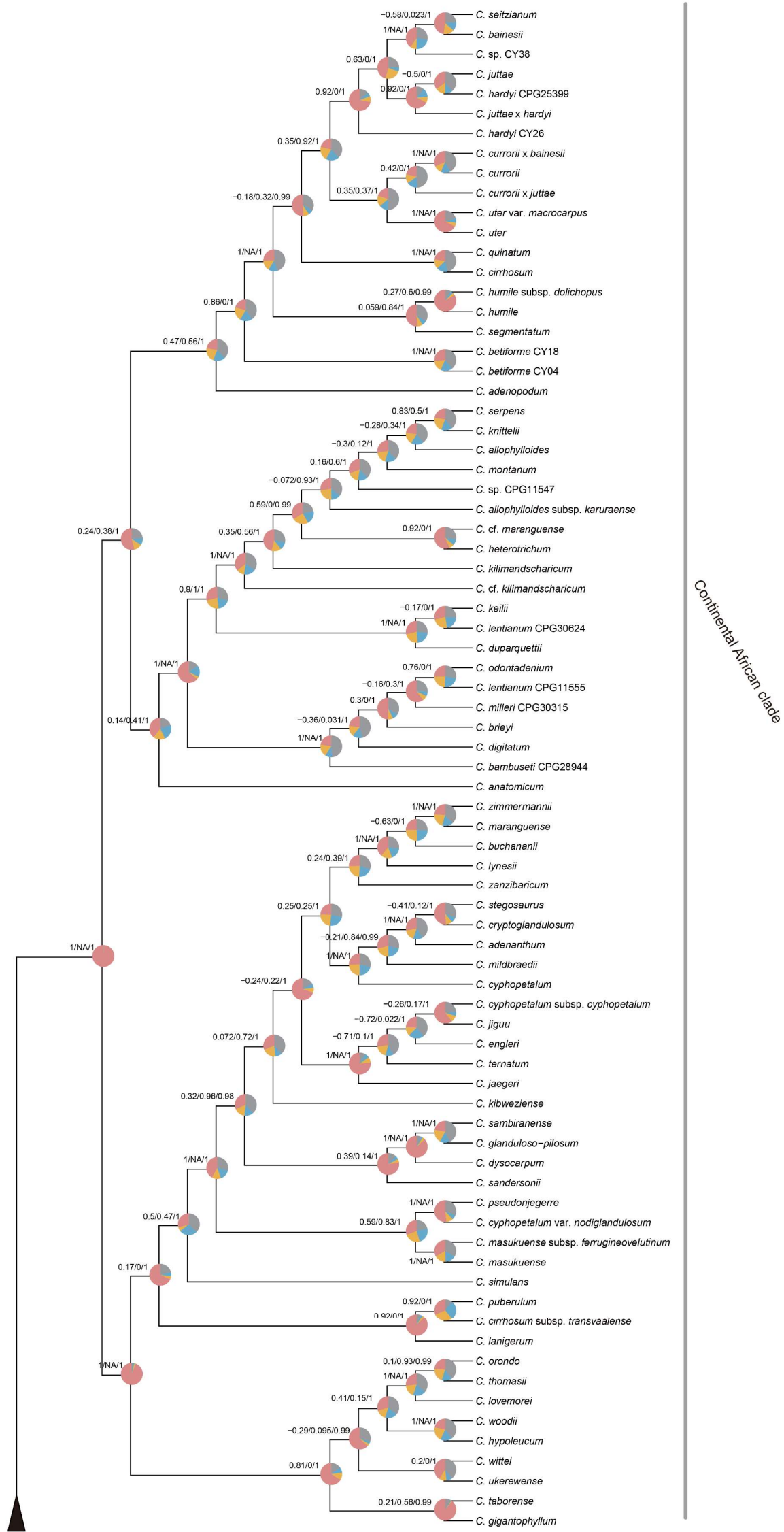


Fig. S2 | Phylogeny of *Cyphostemma* based on 1376 concatenated single-copy orthologs using the maximum likelihood method. Bootstrap support values are shown above the branches. The

four main clades of *Cyphostemma* are indicated with different colors. The black arrows indicate a continued phylogeny.



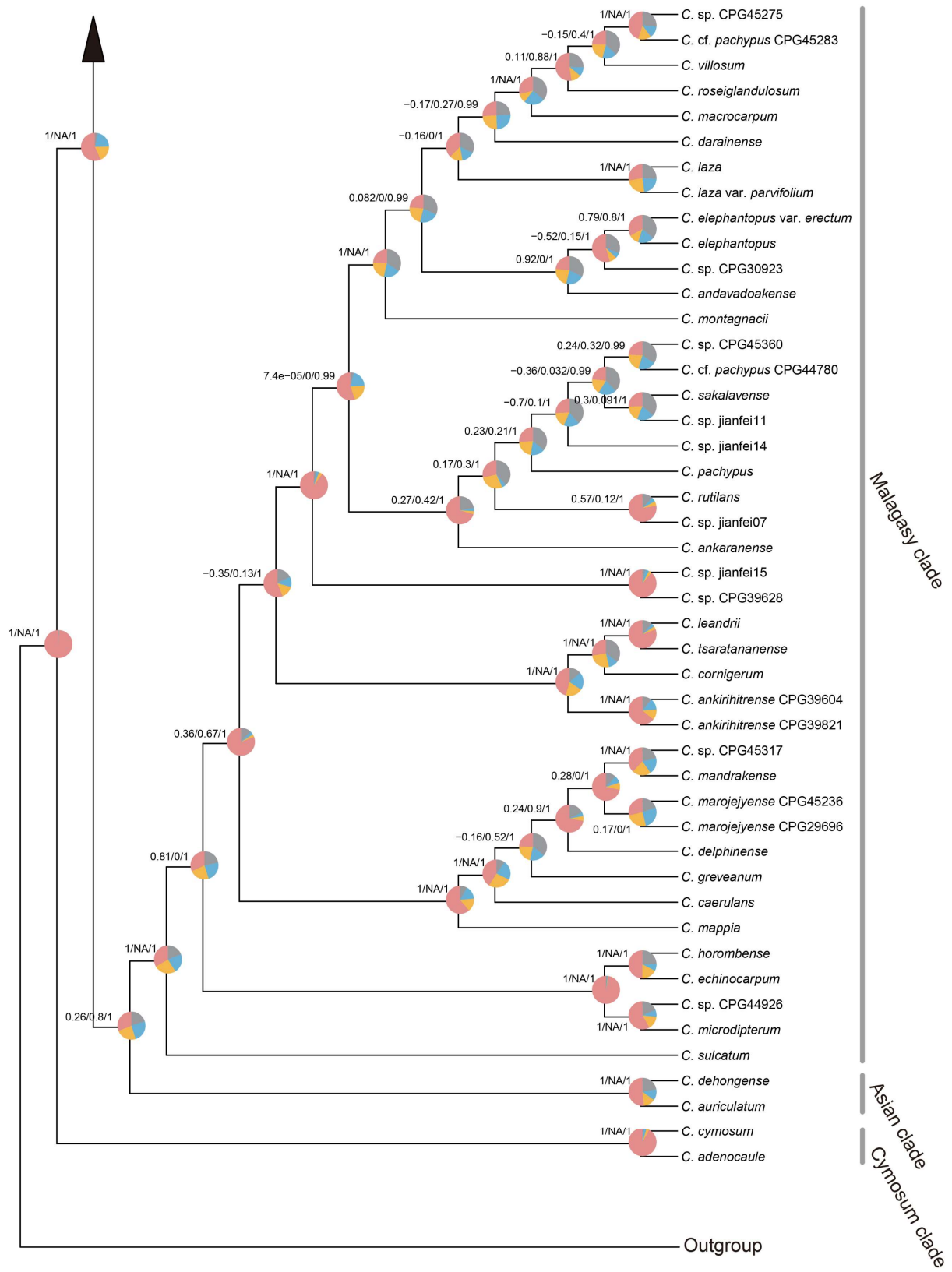


Fig. S3 | Gene tree discordances across the coalescent species tree of *Cyphostemma*. The pie chart on each node shows the proportion of gene trees supporting the congruent relationship (red), the dominant alternative (yellow), the remaining conflicting alternatives (blue), and genes with < 75% bootstrap support values for that node (grey). Quartet concordance/quartet differential/quartet informativeness values are shown near the nodes. “NA” means that all sampled quartet replicates support the major topology, and no meaningful quartet differences were produced. The black arrows indicate a continued phylogeny.

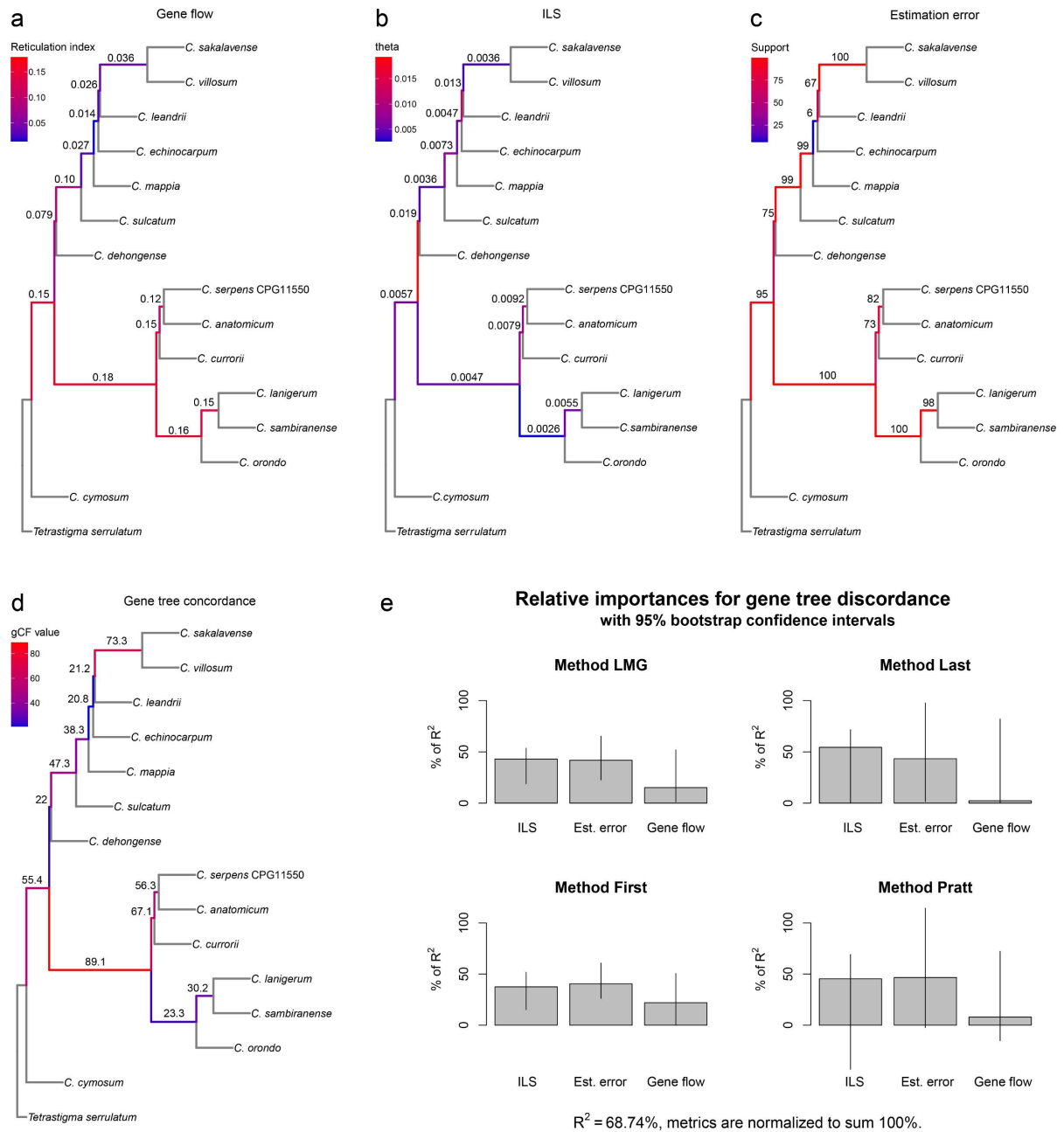


Fig. S4 | Relative importance of gene flow, incomplete lineage sorting (ILS), and gene tree estimation error for gene tree discordance. **a–c**, Pruned *Cyphostemma* phylogenies showing effects of three primary sources of gene tree discordance. **a**, Gene flow. Branches are colored by reticulation index, where the higher value indicates stronger hybridization. **b**, ILS. Branches are colored by the estimated population mutation parameter theta, where a higher theta value indicates

stronger ILS. **c**, Estimation error. Branches are colored by bootstrap support of simulated gene trees, where a higher support value indicates fewer estimation errors. **d**, Pruned *Cyphostemma* phylogeny shows gene tree variation. Branches are colored by gene concordance factor (gCF), where a higher value shows less gene tree discordance. **e**, Decompositions of ILS, estimation error (Est. error), and gene flow, calculated with Lindeman, Merenda, and Gold (LMG), Last, First, and Pratt methods, respectively. Error bars on the columns represent 95% confidence intervals.

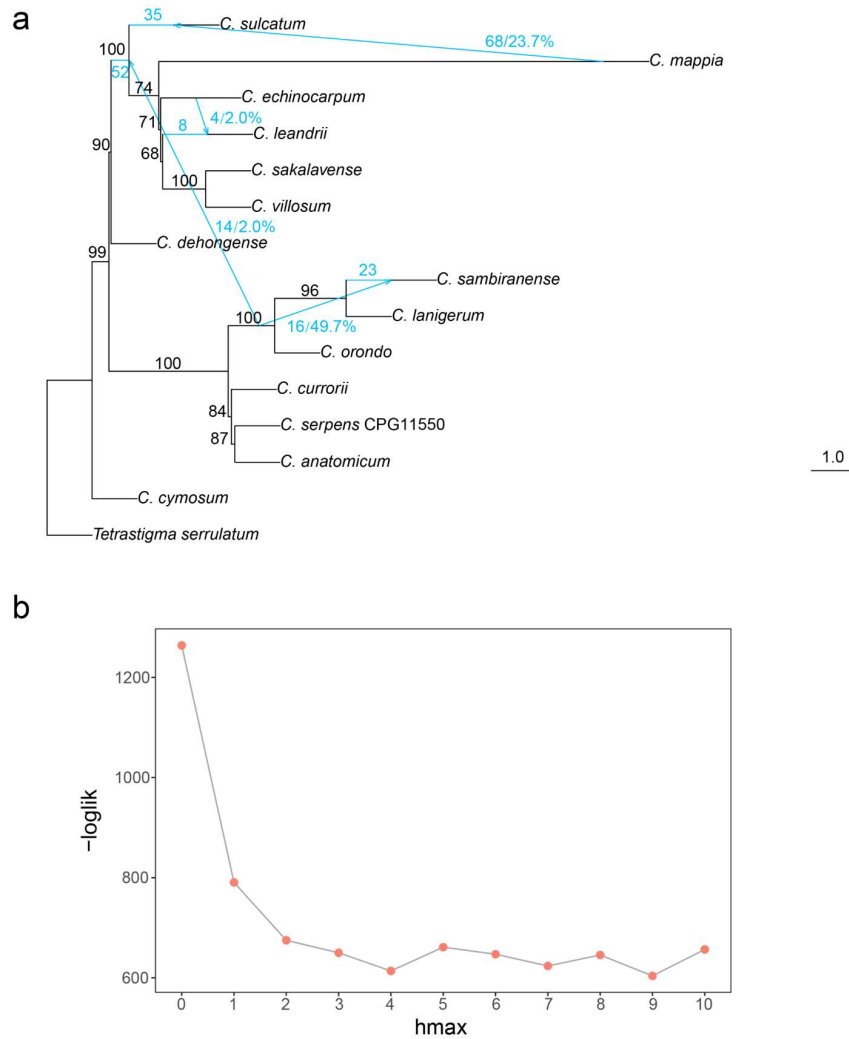
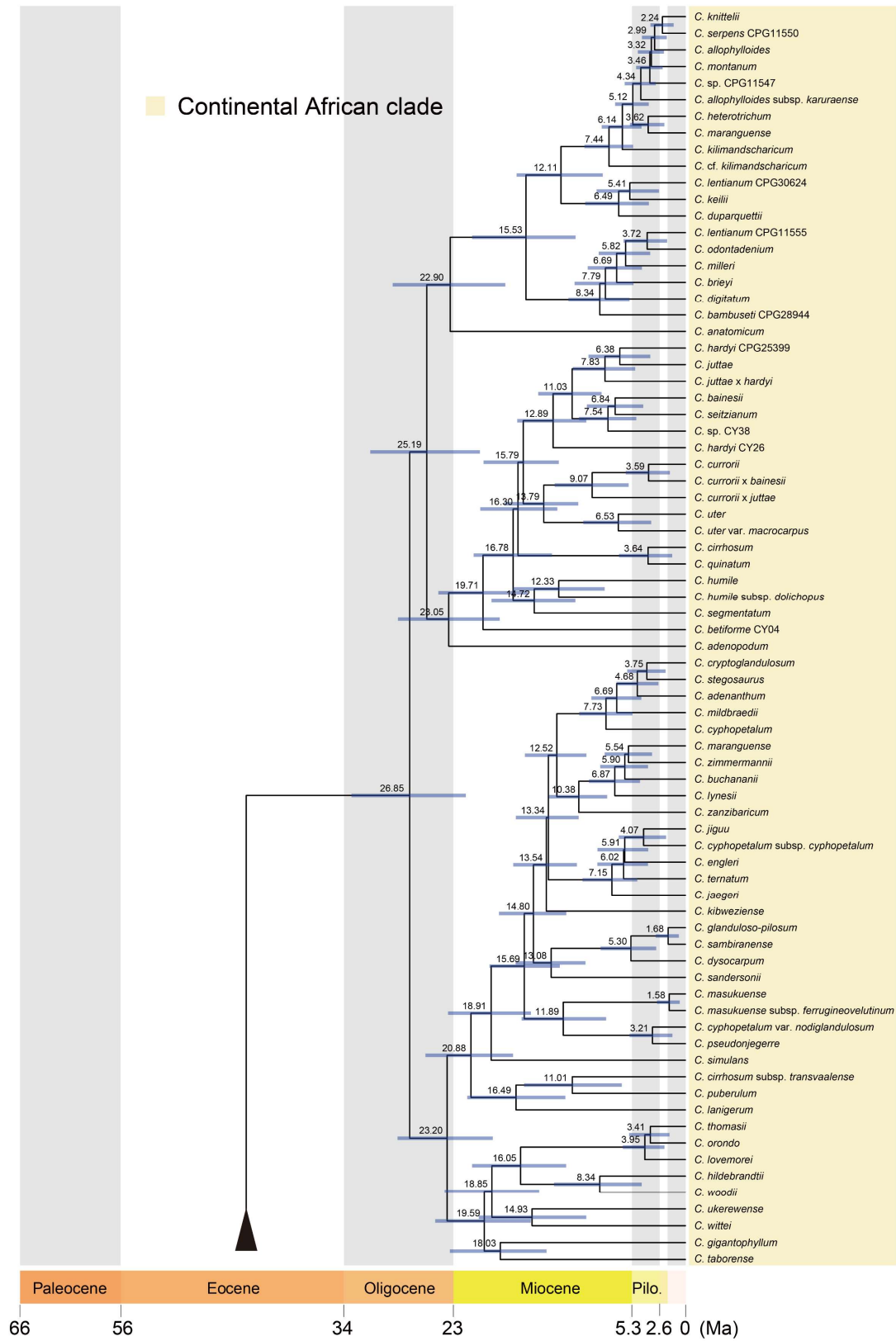
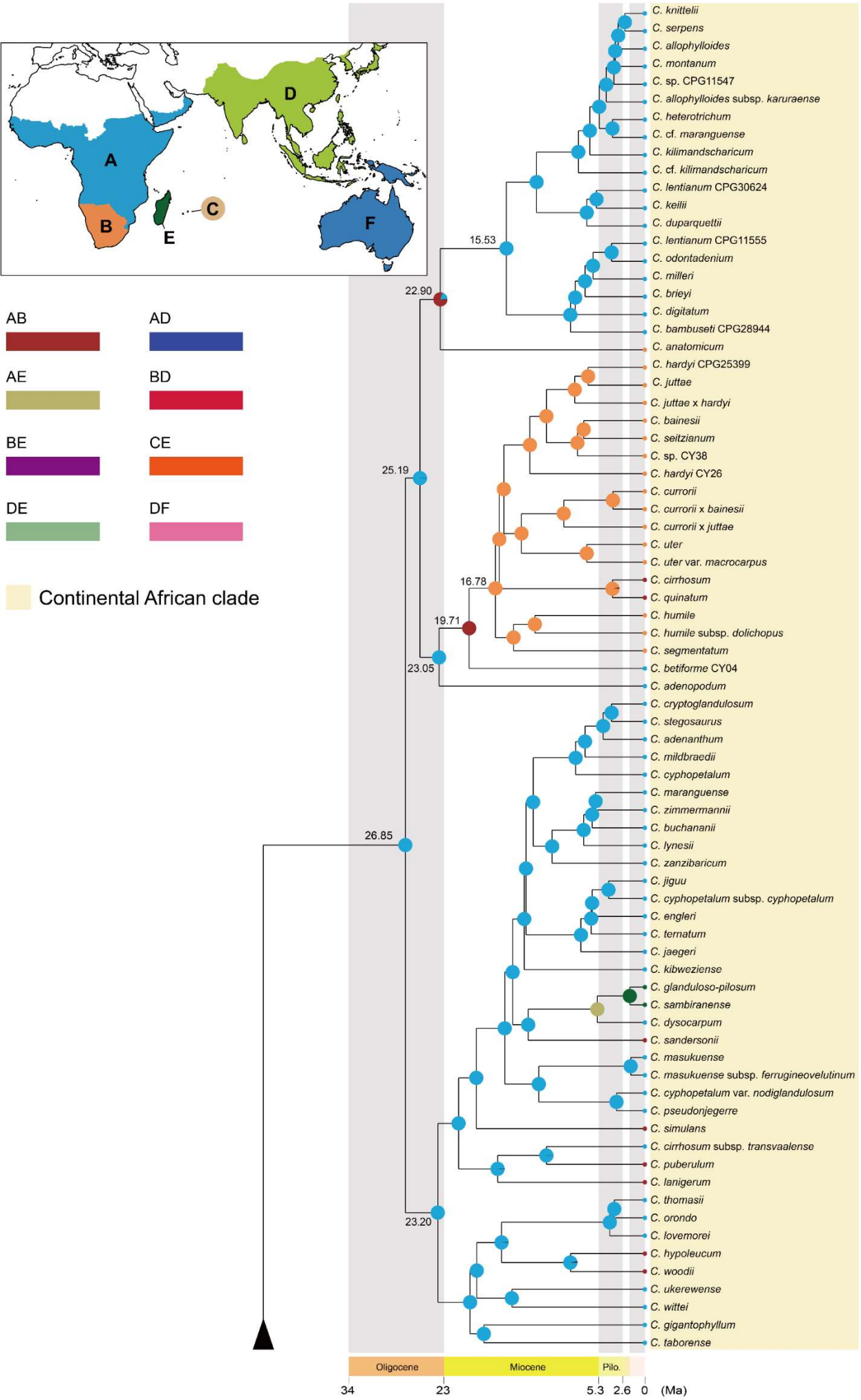


Fig. S5 | Estimated hybridizations among the major subclades of *Cyphostemma*. **a**, Phylonetwork of 14 species representing major subclades of *Cyphostemma* and the outgroup, estimated by SNaQ. Hybrid edges are indicated with blue lines, with bootstrap support values and minor inheritance probabilities shown along the lines in blue. Bootstrap support values of other edges are shown in black. The edge lengths are shown in coalescent units, with the length of each terminal branch set as 1. **b**, Negative log pseudolikelihood (-loglik) score profiles obtained by SNaQ, with hmax from 0 to 10.



with blue bars. Fossils used for calibration are shown with black stars on corresponding nodes. Ma, million years ago. “Pilo.”, Pilocene. The black arrows indicate a continued phylogeny.



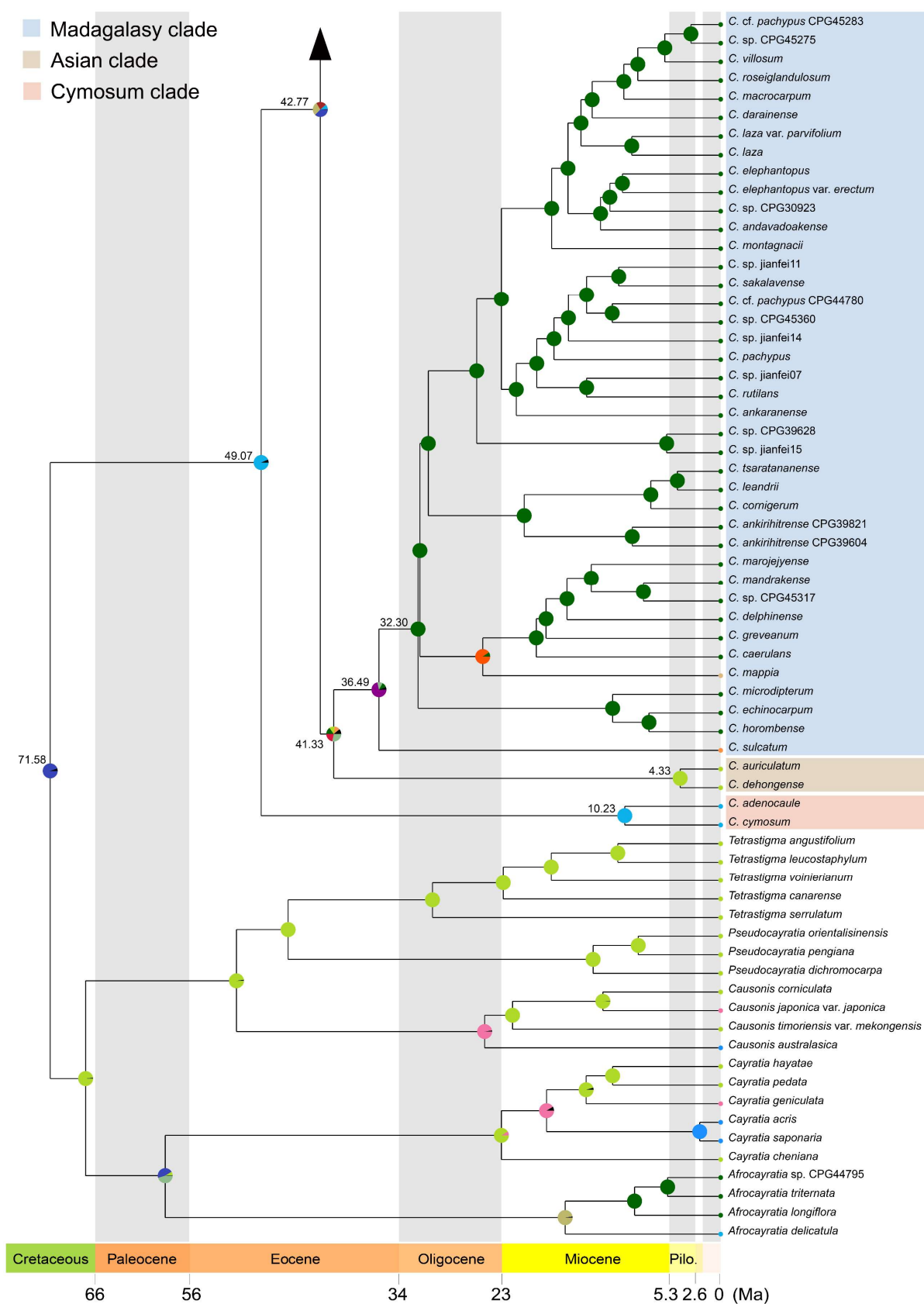


Fig. S7 | Ancestral area reconstructions based on the time tree of *Cyphostemma*. Pie charts on the nodes indicate the relative possibilities of ancestral areas estimated. A, tropical Africa; B, southern Africa; C, Mauritius; D, Asia (including the East Asian monsoon region, Sino-Himalaya, Indochina, and the Malesian region); E, Madagascar; F, Australasia (including continental Australia and New Guinea). Ma, million years ago. The black arrows indicate a continued phylogeny. The map is made with Natural Earth (<https://www.naturalearthdata.com>; accessed June 2024) in the public domain. “Pilo.”, Pilocene.

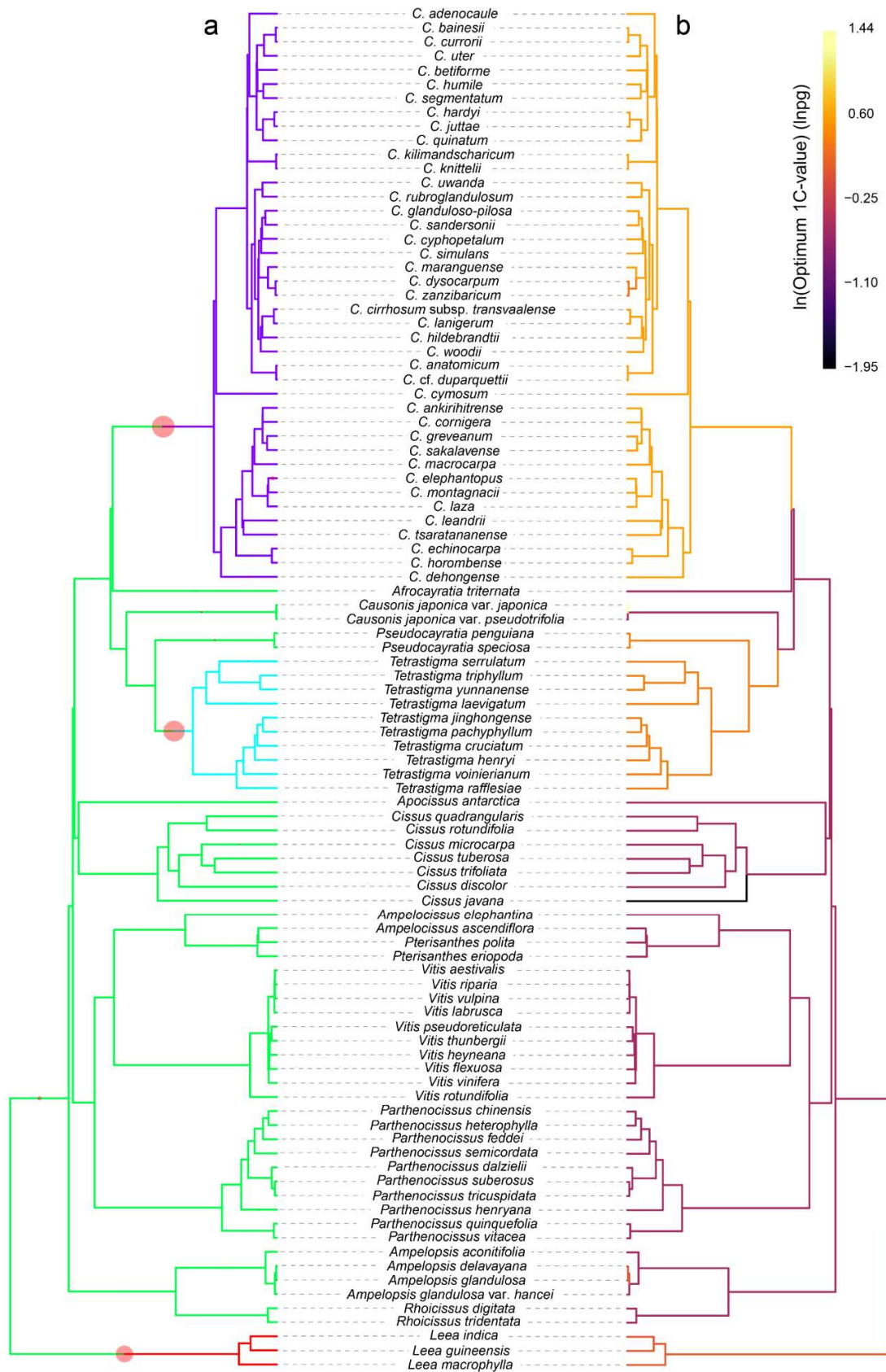


Fig. S8 | Shifts in the dynamics of genome size evolution towards an optimum value (i.e., regime shifts) in Vitaceae. **a**, Different dynamics of genome size across the tree, shown with different branch colors. Solid red circles indicate the positions of regime shifts, where larger circles represent higher confidence. **b**, Optimum genome size (1C-value) under the selective regimes on lineages, indicated by branch color.

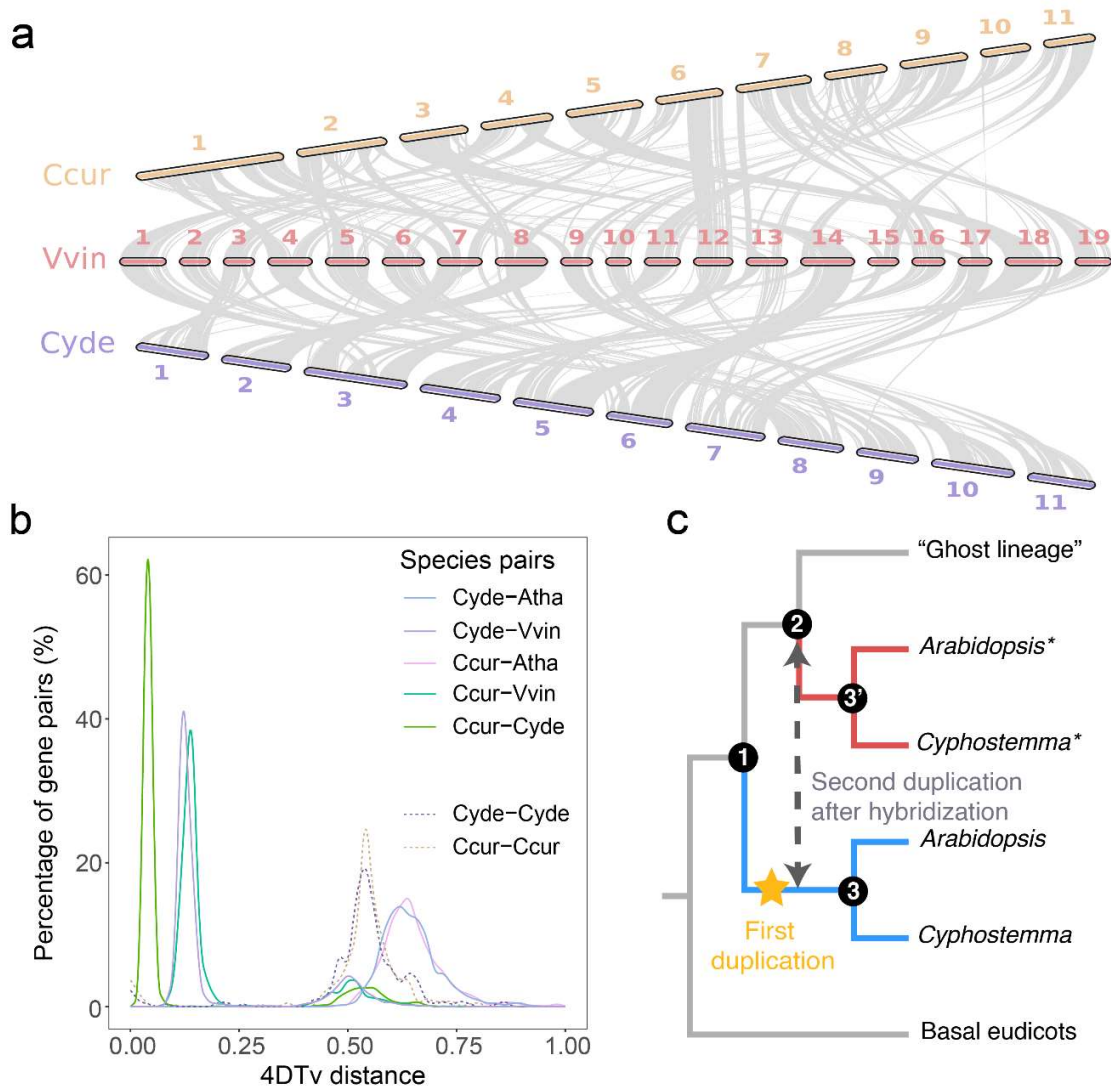


Fig. S9 | Analyses to detect ancient whole genome duplications in *Cyphostemma* species. a, Synteny between genomes of *C. dehongense*, *C. currorii*, and *Vitis vinifera*. Syntenic blocks between species are shown in gray. **b,** Distribution of 4DTv distance in syntenic genes between species pairs and within the genomes of *C. dehongense* and *C. currorii*. **c,** Schematic illustration of the two-step process of the γ event shared by core eudicots, inferred from 4DTv distribution (see Supplementary Note 1 for detailed interpretation). Atha, *Arabidopsis thaliana*. Cyde, *C. dehongense*. Ccur, *C. currorii*. Vvin, *Vitis vinifera*.

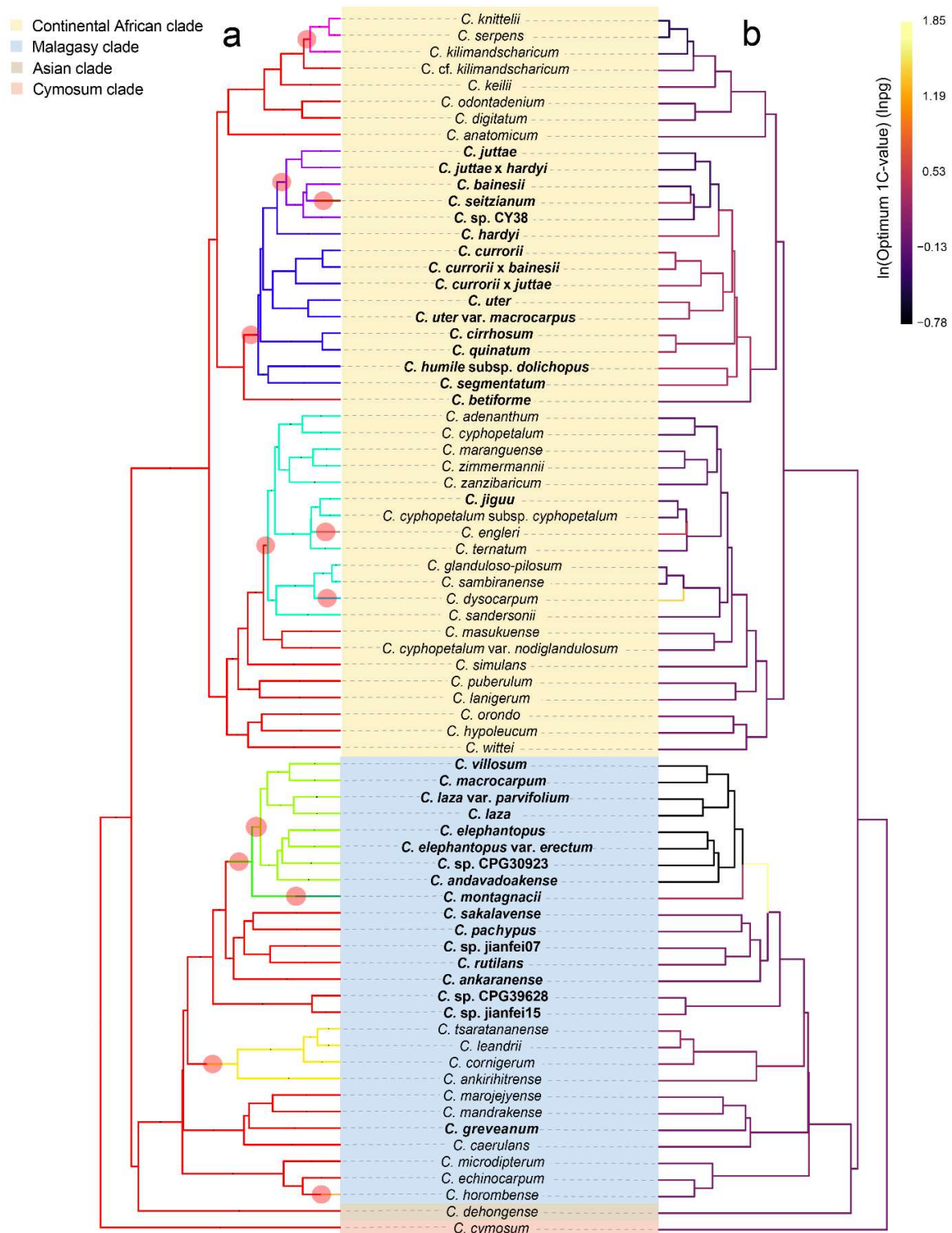


Fig. S10 | Changes in genome size variation regime in *Cyphostemma*. a, Different selective regimes on genome size, shown with different branch colors. Solid red circles indicate the positions

of regime shifts, where larger circles represent higher confidence in the predicted genome expansion events. **b**, Optimum genome size (1C-value) under the selective regimes estimated using Ornstein-Uhlenbeck model, indicated by branch color. Branch colors reflect the inferred evolutionary tendency toward an optimal genome size under selection rather than the reconstructed absolute genome size of ancestral nodes. Succulent species are labeled in bold.

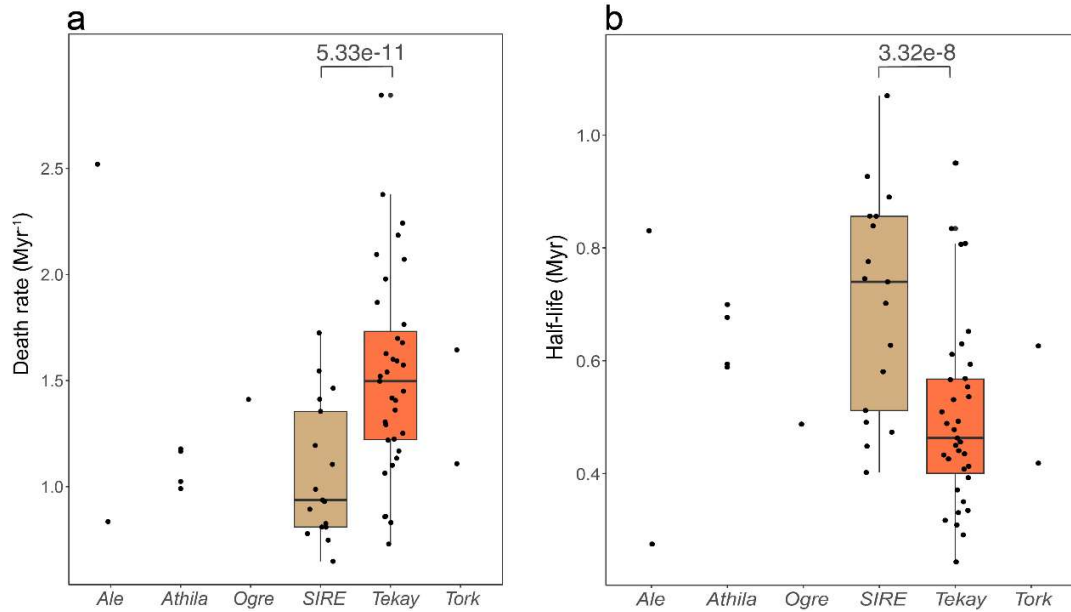


Fig. S11 | Insertion dynamics of major LTR-RT lineages in *C. currorii* calculated by the R package TE. a, Death rate. b, half-life. Myr, million years. For LTR-RT families with limited sample sizes (*Ale*, *Athila*, *Ogre*, and *Tork*), individual value plots are shown instead of boxplots. *Ale*, $n = 2$; *Athila*, $n = 4$; *Ogre*, $n = 1$; *SIRE*, $n = 17$; *Tekay*, $n = 34$; *Tork*, $n = 2$. The centre line represents the median, the box bounds indicate the first and third quartiles, and the whiskers extend to the minimum and maximum values. Unpaired two-sided Wilcoxon rank sum tests were employed.

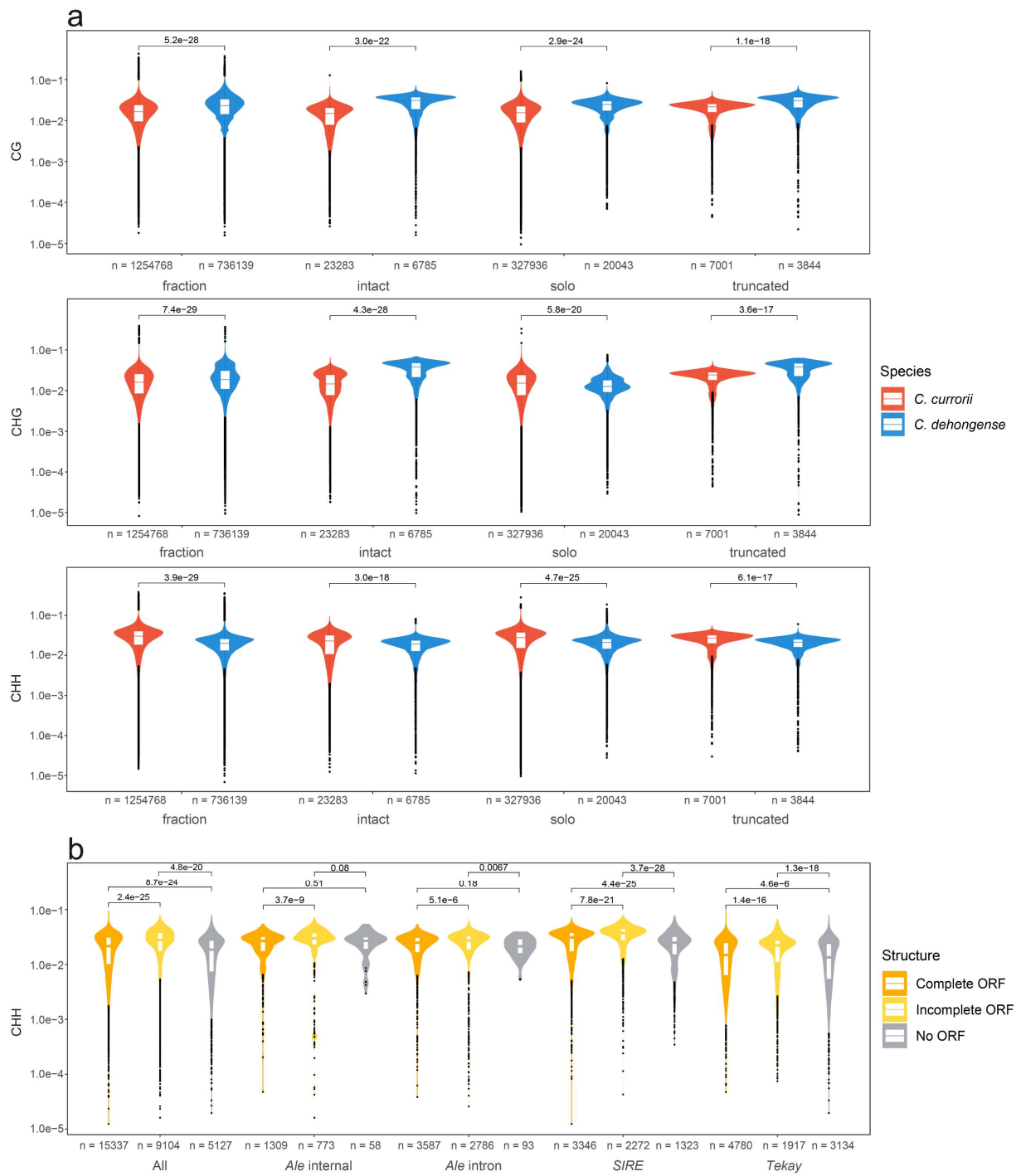


Fig. S12 | Methylation levels of LTR-RT elements. a, CG, CHG, and CHH methylation levels of LTR-RT fractions (*C. currorii*, $n = 1254768$; *C. dehongense*, $n = 736139$), intact LTR-RTs (*C. currorii*, $n = 23283$; *C. dehongense*, $n = 6785$), solo LTR-RTs (*C. currorii*, $n = 327936$; *C.*

dehongense, n = 20043), and truncated LTR-RTs (*C. currorii*, n = 7001; *C. dehongense*, n = 3844).

b, CHH methylation level for intact LTR-RT elements in *C. currorii*, showing differences in methylation level in elements with complete Open Reading Frames (ORF), incomplete ORF, and no ORF. Unmethylated elements are not shown in the figure. For boxplots, the centre line represents the median, the box bounds indicate the first and third quartiles, and the whiskers extend to the minimum and maximum values. Unpaired two-sided Wilcoxon rank sum tests were employed.

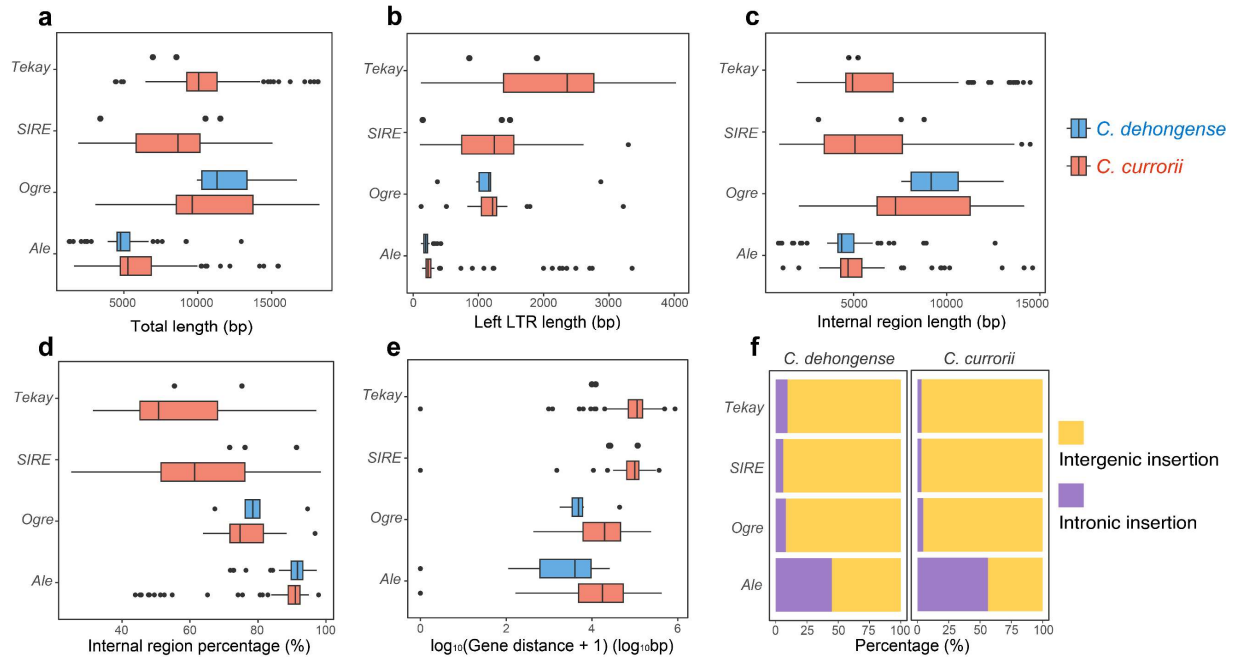


Fig. S13 | Features of LTR-RT lineages of the four most abundant families in *C. dehongense* and *C. currorii*. **a**, Length of LTR-RT elements. **b**, Left long-terminal repeat (LTR) length. **c**, Internal region length, which is the total element length minus the length of left and right LTRs. **d**, Percentage of the internal region in the total length of the element. **e**, Distance to the nearest genes. In box plots, each dot represents one LTR-RT lineage in *C. dehongense* (Tekay, n = 2; SIRE, n = 3; Ogre, n = 6; Ale, n = 63) or *C. currorii* (Tekay, n = 144; SIRE, n = 62; Ogre, n = 44; Ale, n = 82). For boxplots, the centre line represents the median, the box bounds indicate the first and third quartiles, and the whiskers extend to the minimum and maximum values. Unpaired two-sided Wilcoxon rank sum tests were employed. **f**, Percentages of intergenic and intronic insertions.

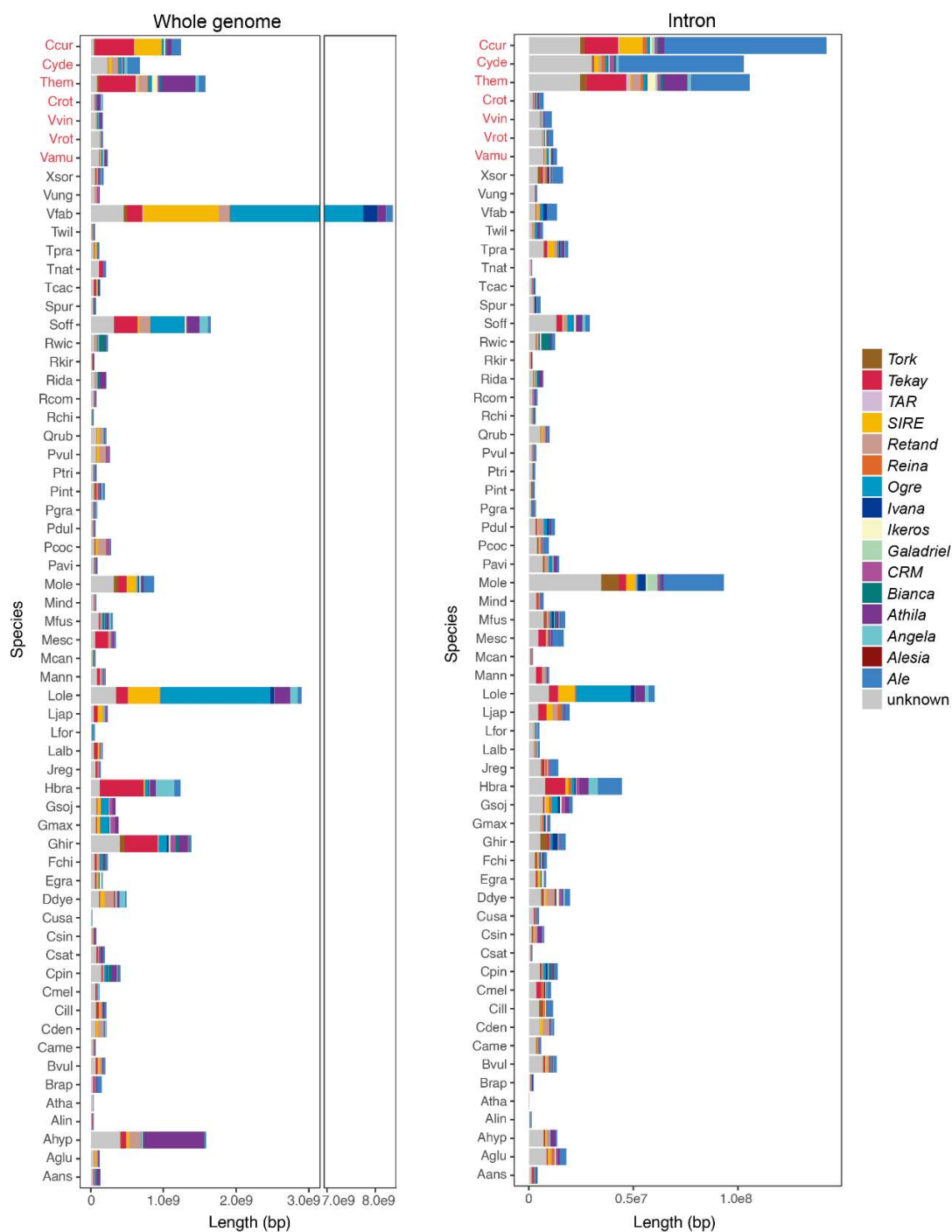


Fig. S14 | LTR-RT family compositions in the whole genome and introns of seven Vitaceae species and 55 other super-rosid species. Vitaceae species are shown in red. Full species names corresponding to abbreviations are provided in Supplementary Data 6.

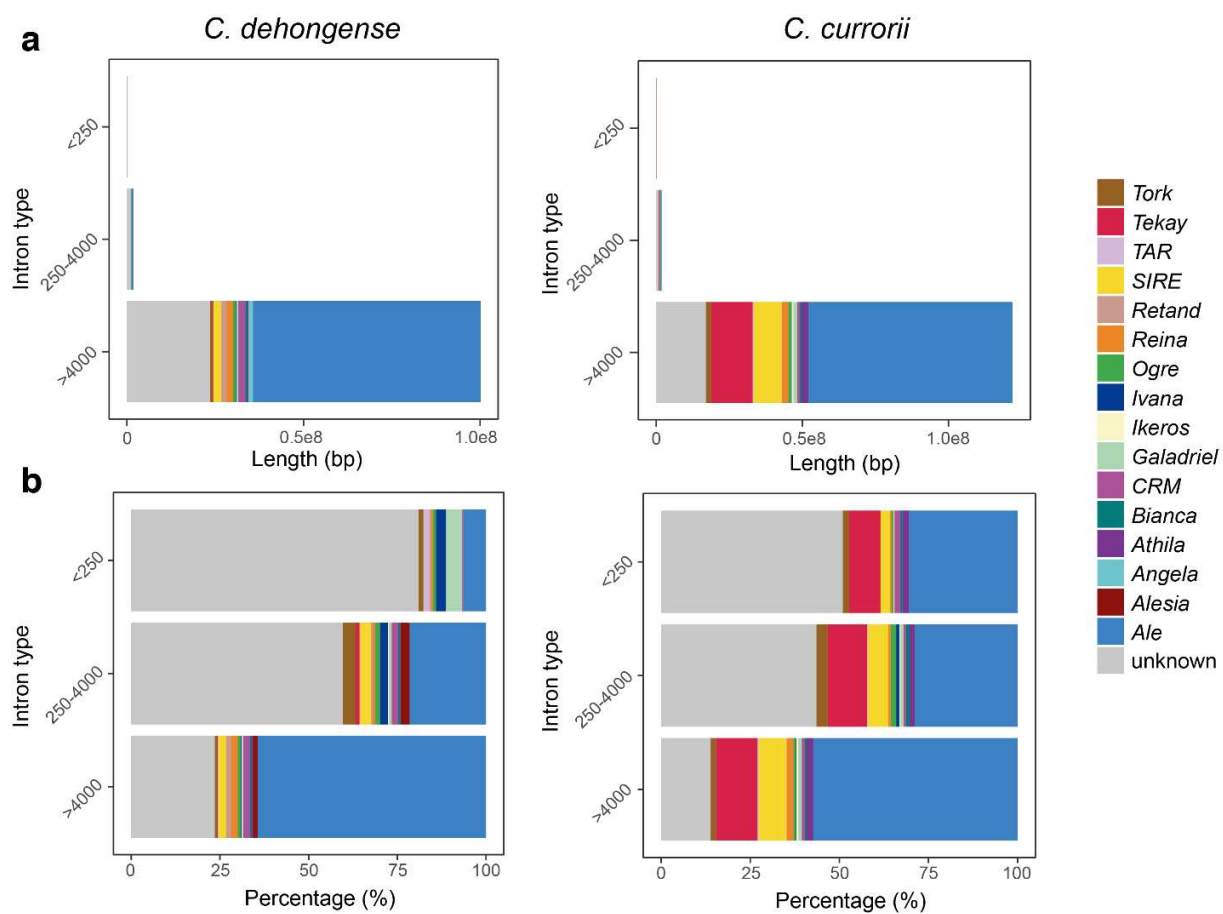


Fig. S15 | Intronic LTR-RT family compositions in genes of *C. dehongense* and *C. currorii*.

Compositions are shown by **a**, length, and **b**, percentage. Genes are grouped by the length of their longest intron.

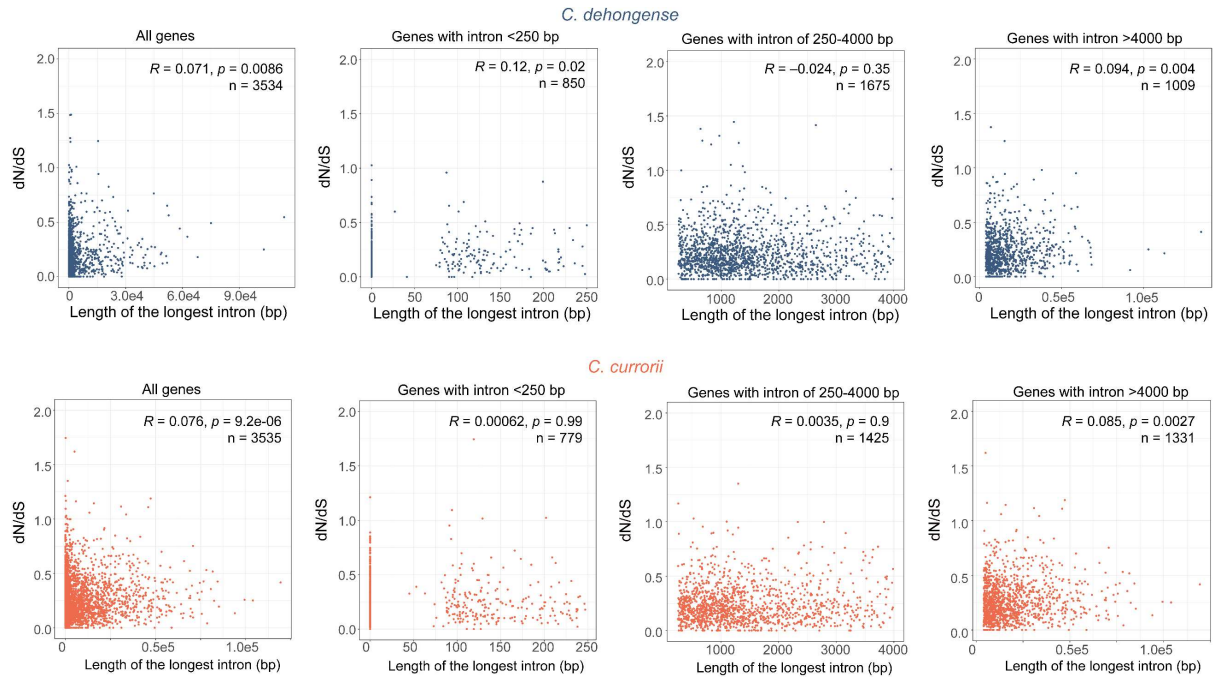


Fig. S16 | Ratio of nonsynonymous/synonymous sites (dN/dS ratio) in genes with introns of different lengths in the genomes of *C. dehongense* and *C. currorii*. A higher ratio indicates a more relaxed selection acting on genes.

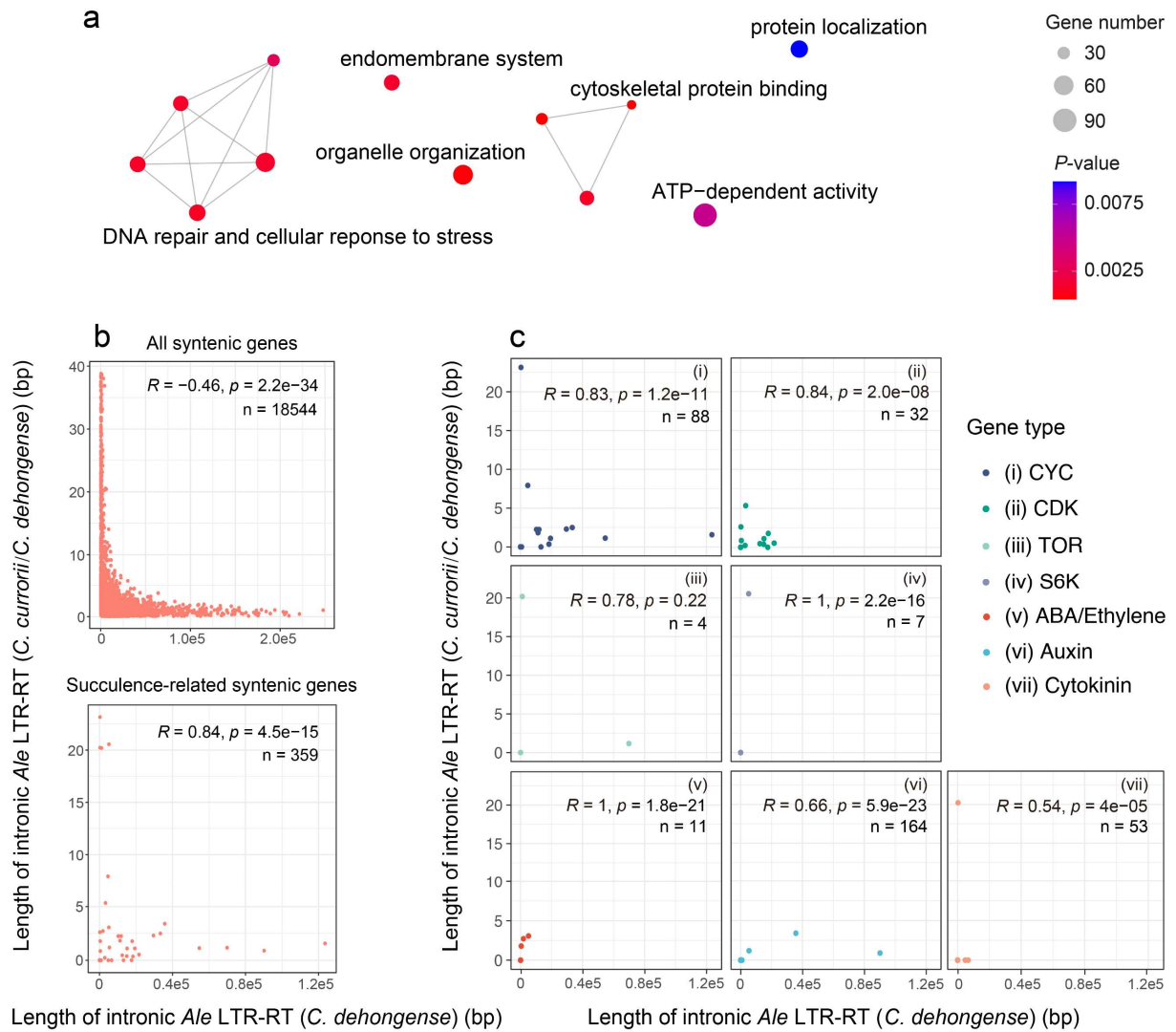


Fig. S17 | Effects of *Ale* LTR-RT insertions on genes. **a**, Gene ontology enrichment of *C. currorii* genes with extra-long introns (> 4000 bp) elongated by *Ale* LTR-RT insertions compared with *C. dehongense*. Solid circles represent enriched biological processes, where the size indicates the number of related genes and the color indicates statistical significance. **b–c**, Relationships between the length of intronic *Ale* LTR-RT in *C. dehongense* and the expanded length of intronic *Ale* LTR-RT in *C. currorii*. **b**, Correlations for all syntenic genes and succulence-related syntenic genes. **c**, Correlations for succulence-related syntenic genes of different functions. CYC, cyclin genes; CDK,

cyclin-dependent kinases; TOR, target of rapamycin; S6K, ribosomal S6 kinases; ABA, abscisic acid. Correlations were assessed by Spearman's rank correlation test.

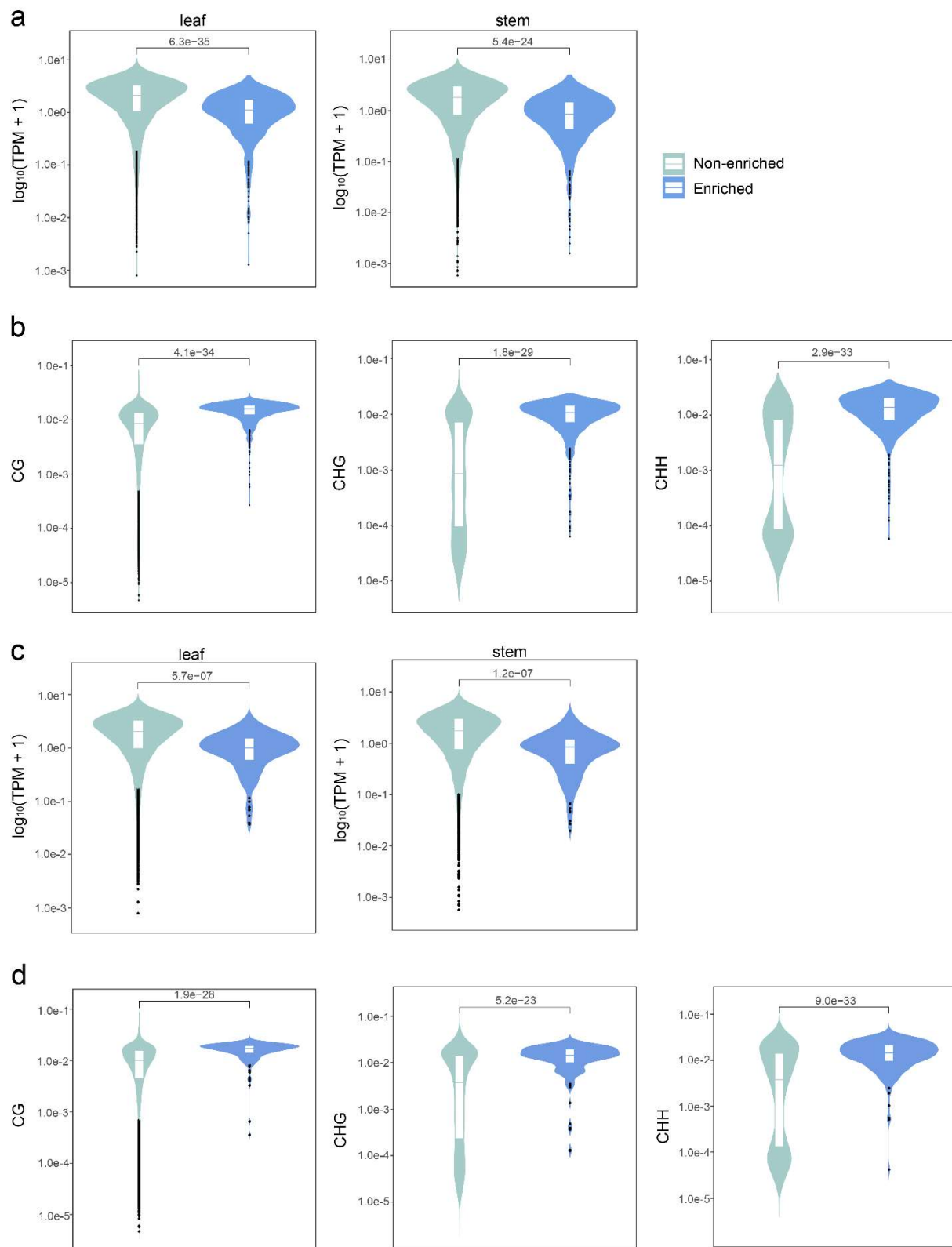
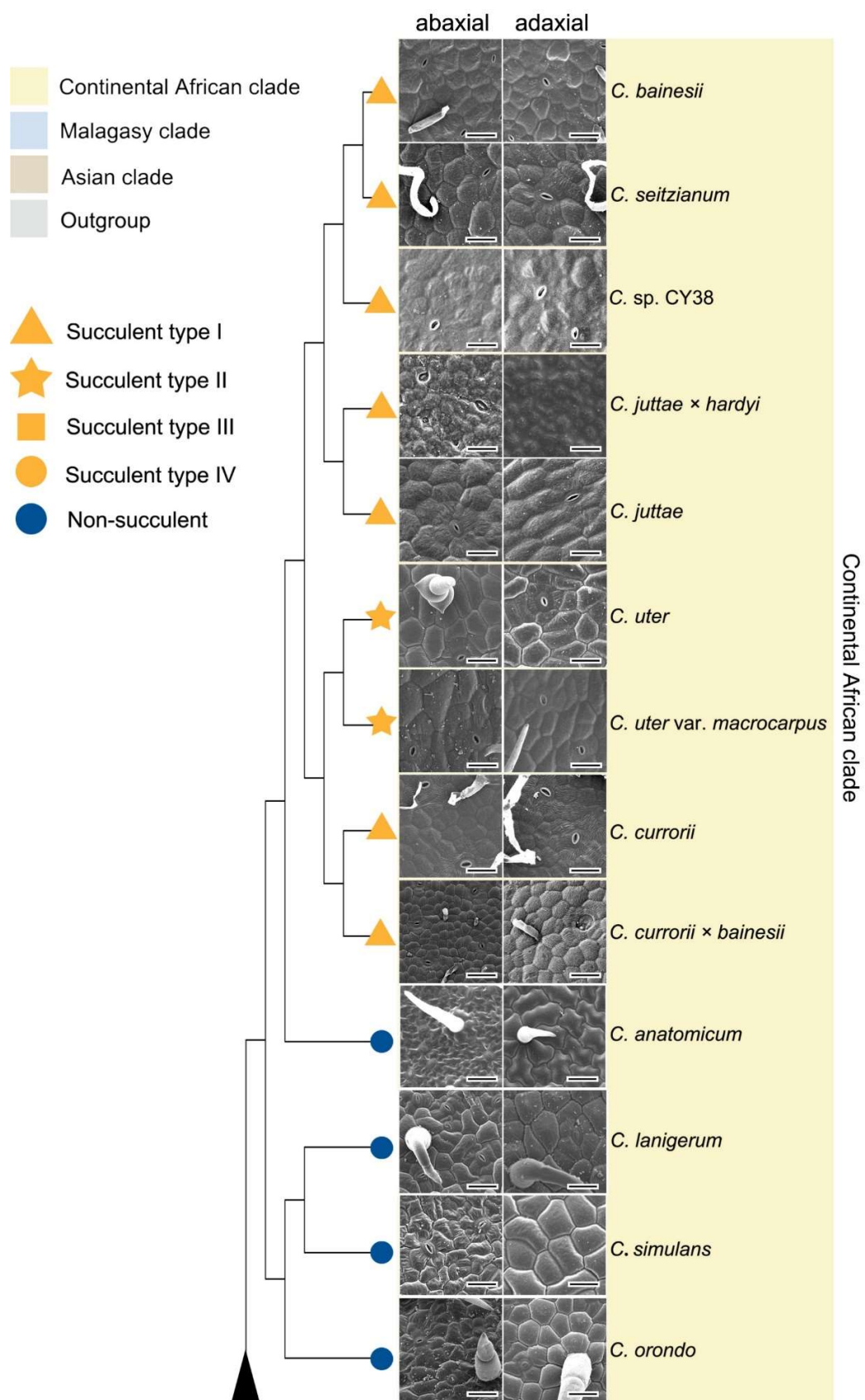


Fig. S18 | Comparison between enriched and non-enriched genes. **a**, Expression levels in leaf and stem and **b**, methylation levels of non-enriched (n = 17497) and enriched (n = 1047) genes in *C. currorii* with introns > 4000 bp, elongated with LTR-RT insertion. Differences in **c**, Expression levels and **d**, methylation levels between enriched genes involved in transcription regulation and DNA repair and all the other genes in *C. currorii* were especially calculated. Unexpressed or unmethylated genes are not shown in the figure. TPM, transcripts per million. The *P*-value above each box shows the significant difference in expression levels. For boxplots, the centre line represents the median, the box bounds indicate the first and third quartiles, and the whiskers extend to the minimum and maximum values. Unpaired two-sided Wilcoxon rank sum tests were employed.



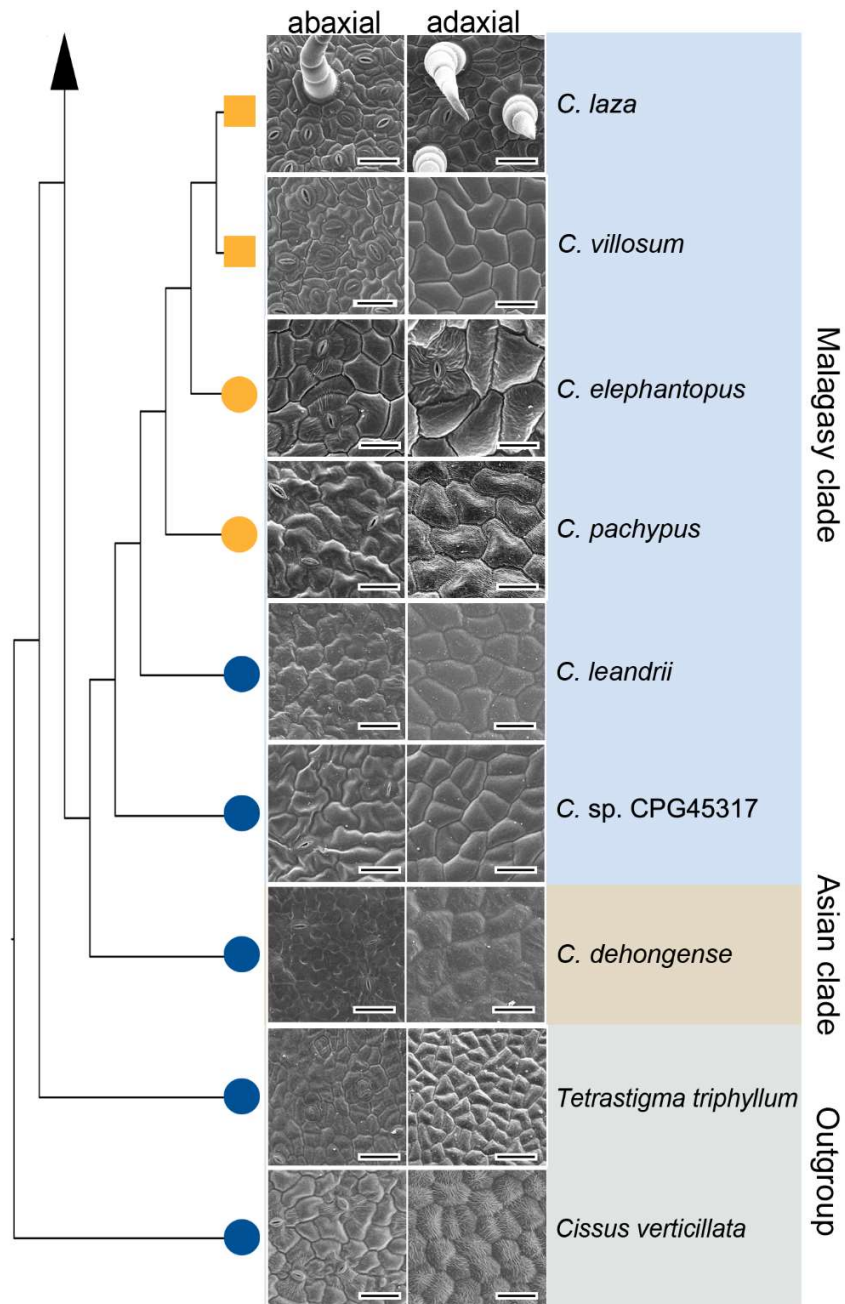


Fig. S19 | Scanning electron micrographs of abaxial and adaxial leaf surfaces of *Cyphostemma* and outgroup species in a phylogenetic framework. Scale bars = 100 μ m. The phylogeny was pruned from the coalescent tree generated by this study. For each species, three individual replicates and five random fields per leaf side were taken.

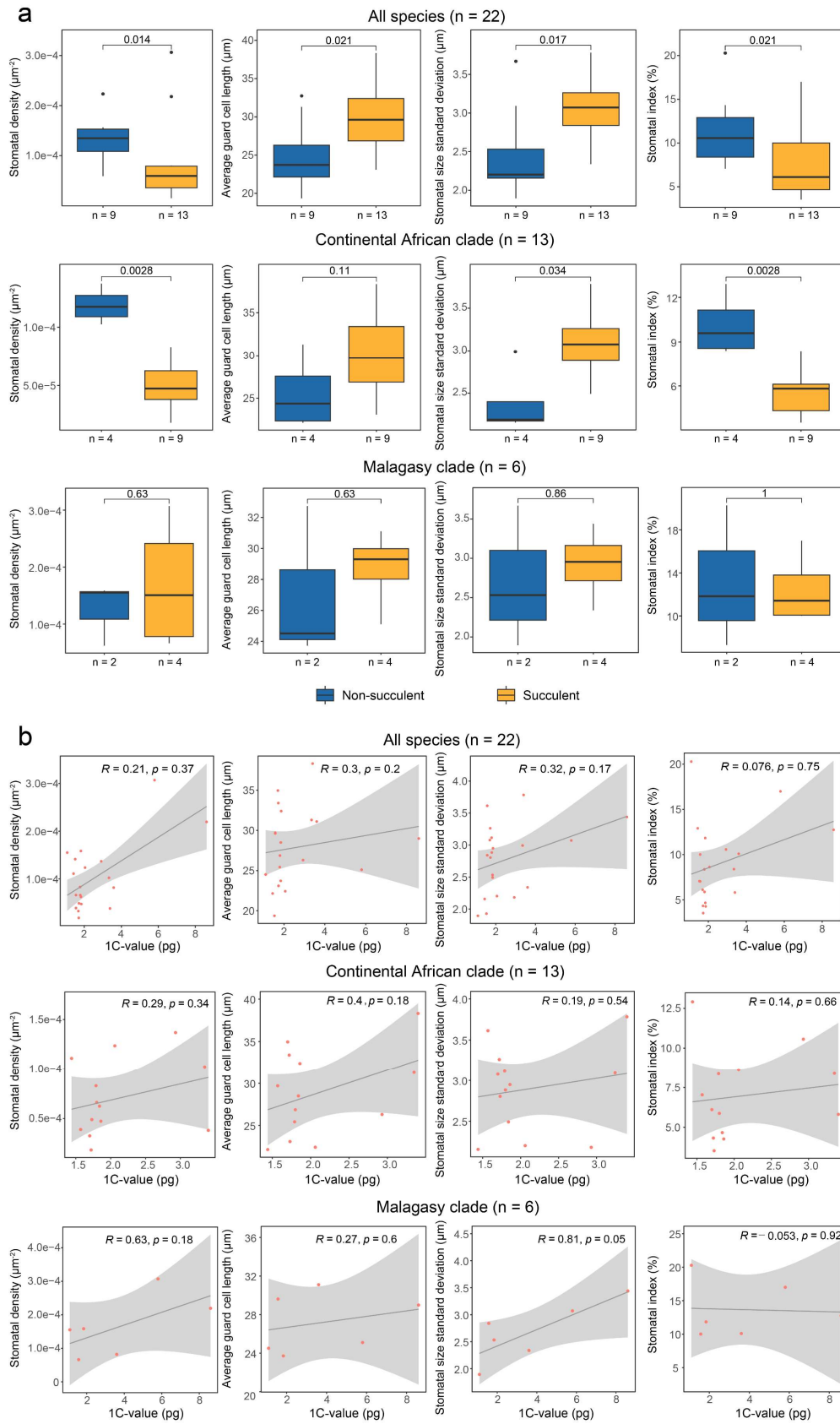


Fig. S20 | Leaf-surface features and their association with genome size (1C-value) in *Cyphostemma*. **a**, Differences in stomatal density, stomata length, stomata size variation, and stomatal index of abaxial leaf surface. For boxplots, the centre line represents the median, the box bounds indicate the first and third quartiles, and the whiskers extend to the minimum and maximum values. Unpaired two-sided Wilcoxon rank sum tests were employed. **b**, Correlation between leaf-surface features and genome size in succulent and non-succulent species in the continental African clade and the Malagasy clade. *P*-values and correlation coefficients Pearson's *r* are labeled above.

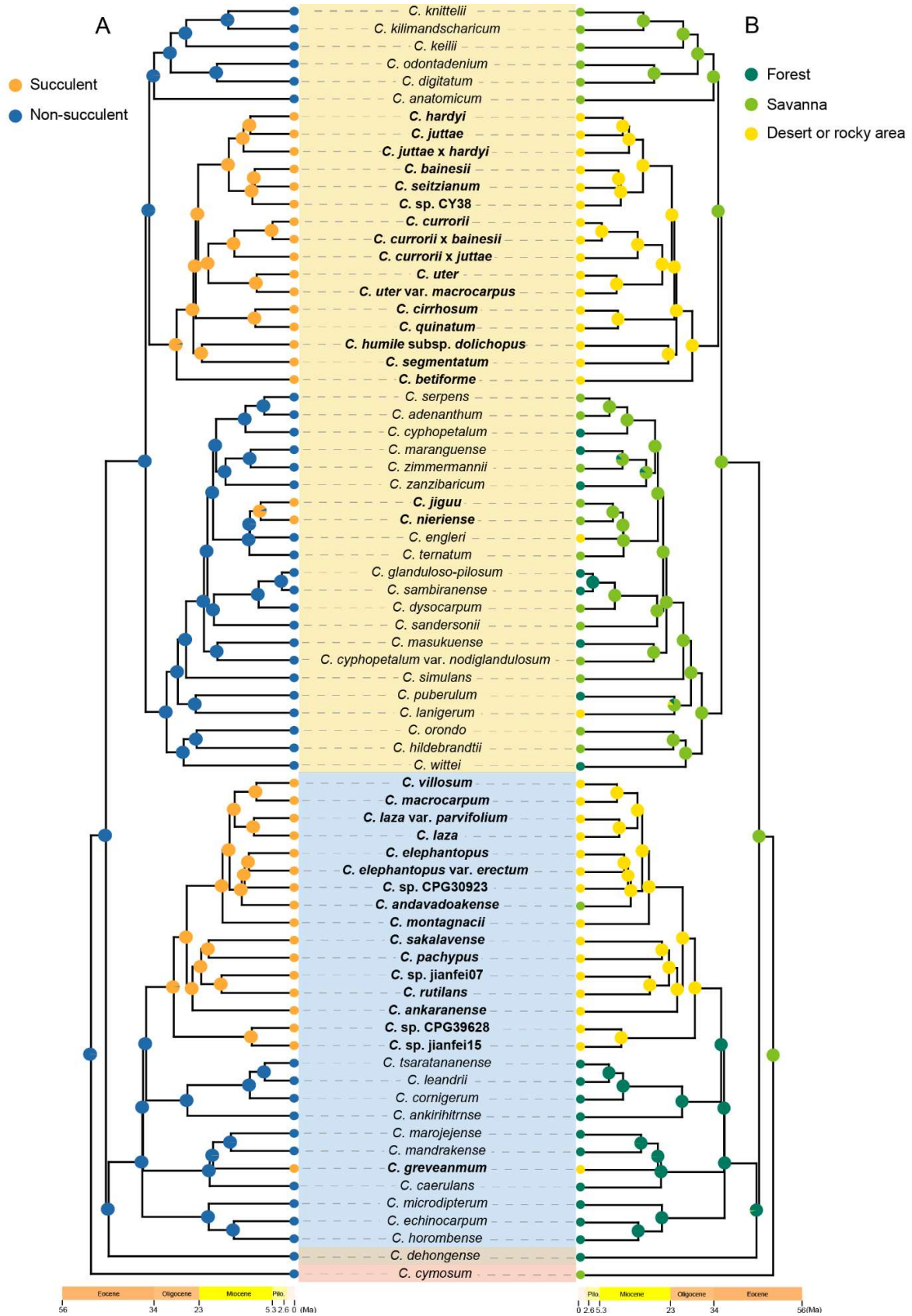


Fig. S21 | Ancestral state reconstruction using stochastic mapping based on the time tree of *Cyphostemma*. a, succulence. b, habitat. Different branch colors show distinct states. Succulent species are labeled in bold. “Pilo.”, Pilocene.

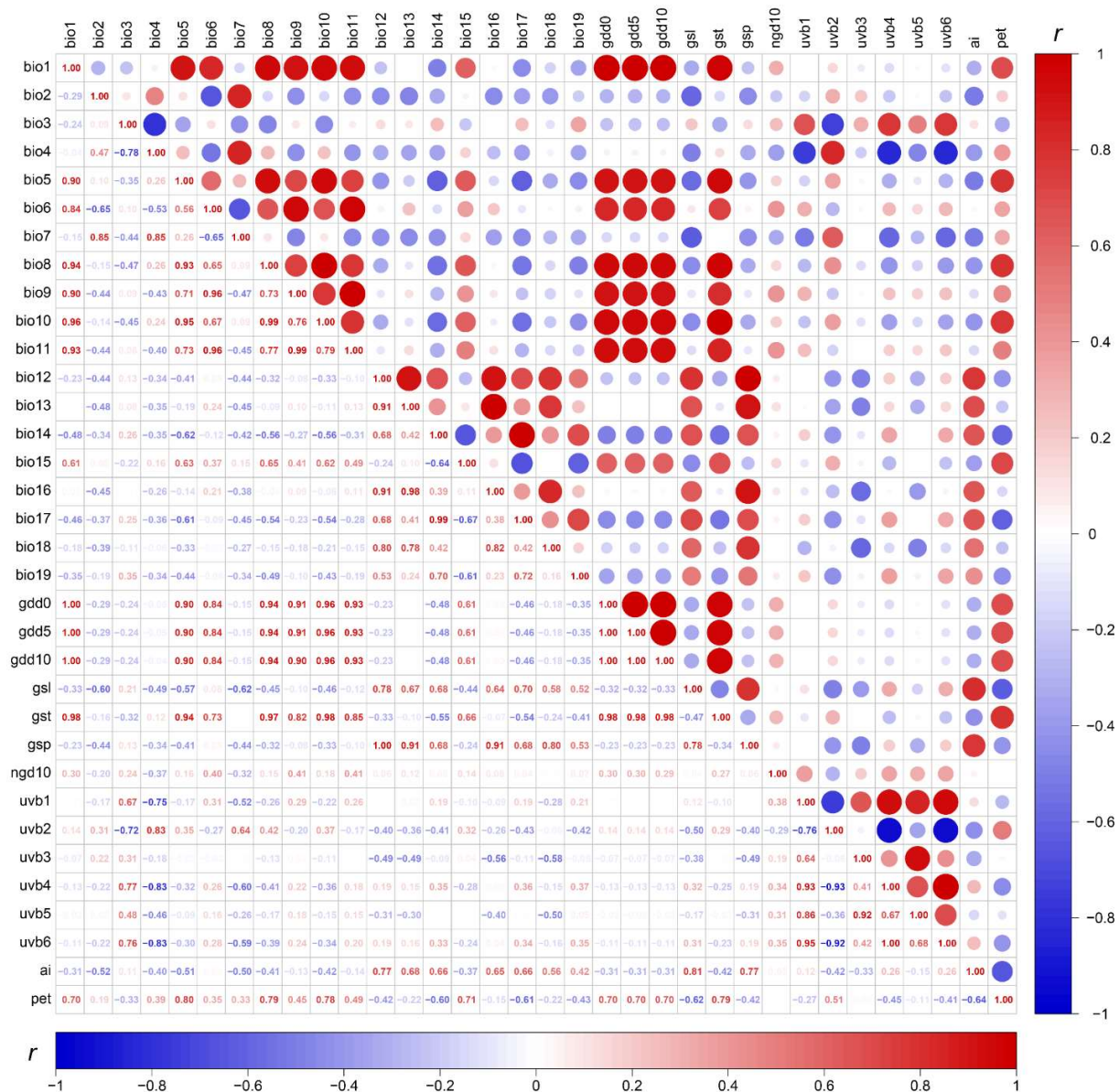


Fig. S22 | Correlations between all 36 environmental variables used in this study. Larger circles represent higher absolute values of the correlation coefficient Pearson's r , with red indicating positive and blue negative. The specific bioclimatic variables represented by the variable names are detailed in Supplementary Data 18.

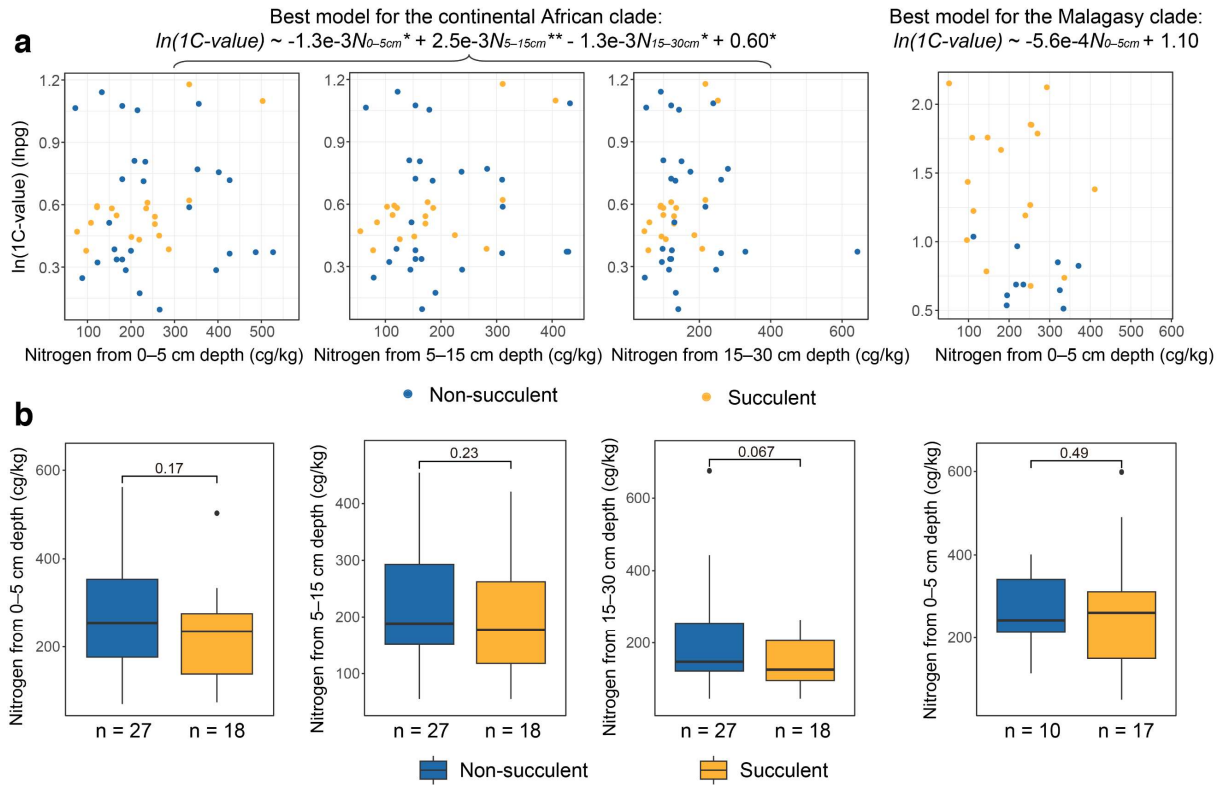


Fig. S23 | Local-scale environmental variables explaining genome size variation. a, Relationships between genome size (1C-value) and the nitrogen levels of the soil at different depths, which are in the best regression model for the continental African clade ($n = 45$) and Malagasy clade ($n = 27$). “***” represents $p < 0.01$, “*” represents $p < 0.05$. **b,** The differences in the local-scale environmental variables in the habitats of succulent lineages of both clades. For boxplots, the centre line represents the median, the box bounds indicate the first and third quartiles, and the whiskers extend to the minimum and maximum values. Unpaired two-sided Wilcoxon rank sum tests were employed.

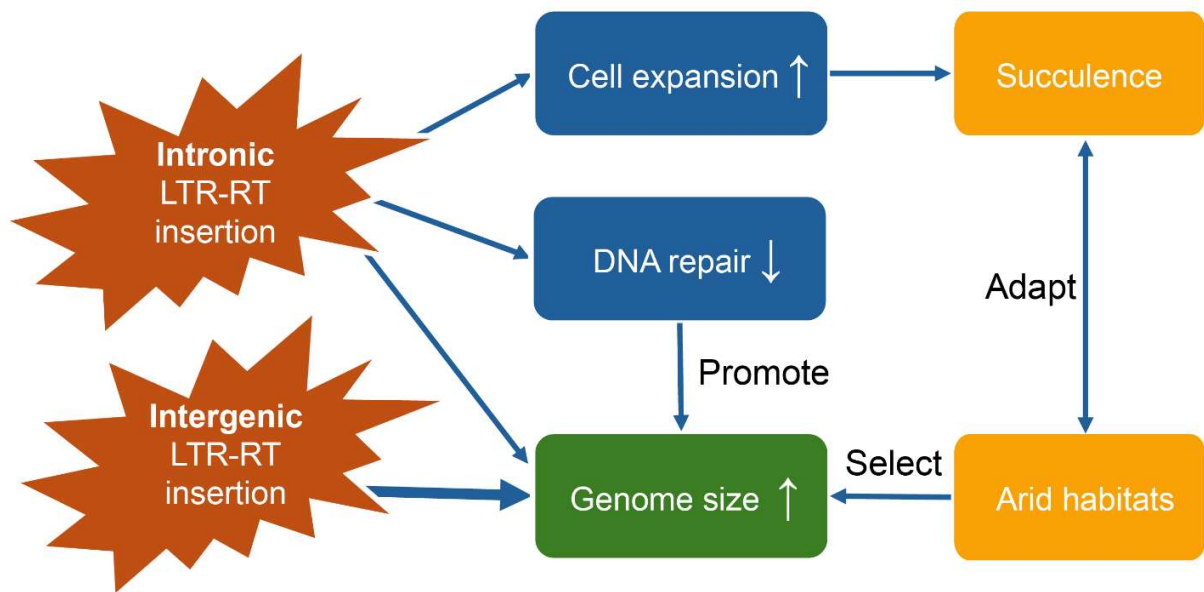


Fig. S24 | Conceptual diagram summarizing the main findings of this study. The thicker arrow from “intergenic LTR-RT insertion” indicates that these insertions represent the primary driver of genome expansion.

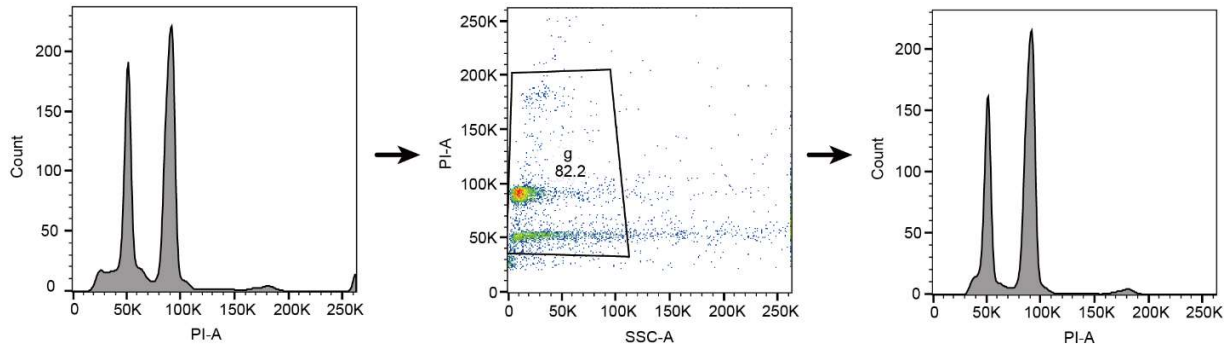


Fig. S25 | Gating strategy for flow cytometric genome size estimation. According to a side scatter area (SSC-A) versus PI fluorescence area (PI-A) dot plot, singlets (single cell signals) were identified. The VL1-A histograms of PI-A of single nuclei shows the G1 peaks of the target species (P1) and the reference standard (P2), showing that debris, dead cells, and electronic noise were excluded. Coefficients of variations (CVs) for both G1 peaks were < 5% in all accepted samples.

Supplementary References

1. Shi, T. & Van de Peer, Y. Revisiting ancient whole-genome duplications in the seed and flowering plants through the lens of dosage-sensitive genes. *Sci. Adv.* **12**, eaea9797 (2026).
2. Chanderbali, A. S. *et al.* *Buxus* and *Tetracentron* genomes help resolve eudicot genome history. *Nat. Commun.* **13**, 643 (2022).