Nature Communications

https://doi.org/10.1038/s41467-026-74005-z

Article in Press

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

Received: 4 August 2025

Accepted: 26 May 2026

Cite this article as: Yu, J., Zhou, J., Luo, S. *et al.* Genomic and ecological drivers of parallel arid adaptation in tree grapes (Vitaceae). *Nat Commun* (2026). https://doi.org/10.1038/s41467-026-74005-z

Jinren Yu, Ju Zhou, Shanshan Luo, Ilia J. Leitch, Kyaw Nyein, Rindra Manasoa Ranaivoson, Chuanyu Du, Russell L. Barrett, Jie Cheng, Chaobin Li, Yang Dong, Romer Narindra Rabarijaona, Alexandre Antonelli, Zhiduan Chen & Limin Lu

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. Leitch[4], Kyaw Nyein[1,2,3], Rindra Manasoa Ranaivoson[5], 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*]

[1]State Key Laboratory of Plant Diversity and Specialty Crops and Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences; Beijing 100093, China.

[2]China National Botanical Garden; Beijing 100093, China.

[3]University of Chinese Academy of Sciences; Beijing, China.

[4]Royal Botanic Gardens, Kew; TW9 3AE Richmond, Surrey, UK.

[5]Department of Plant Biology and Ecology, Faculty of Sciences, University of Antananarivo; Antananarivo 101, Madagascar.

[6]National Herbarium of New South Wales, Australian Botanic Garden; Locked Bag 6002, Mount Annan, NSW 2567, Australia.

[7]School of Biological, Earth, and Environmental Sciences, University of New South Wales; Kensington, NSW 2052, Australia.

[8]Department of Computational and Systems Biology, John Innes Centre, Norwich Research Park, Colney Lane. Norwich NR4 7UH, UK.

[9]Gothenburg Global Biodiversity Centre, Department of Biological and Environmental Sciences, University of Gothenburg; 41319 Gothenburg, Sweden.

[10]Wuhan Botanical Garden, Chinese Academy of Sciences; Wuhan 430074, China.

[11]Department of Biology, University of Oxford; Oxford OX1 3RB, UK.

*Corresponding author. Email: liminlu@ibcas.ac.cn

Abstract

Understanding plant adaptation is critical under intensifying global aridification. Succulence, a key drought-resistance innovation, has evolved repeatedly across plant lineages, yet its intrinsic genomic drivers remain underexplored. Integrating comprehensive evidence from genomics, ecology, and morphology, we investigate adaptation to aridity in the tree grape genus, *Cyphostemma* (Vitaceae), whose species span environmental gradients from rainforests to deserts and exhibit wide genomic and phenotypic variation. Utilizing genome assemblies of representative *Cyphostemma* species, we demonstrate that specific long terminal repeat retrotransposon (LTR-RT) lineages thrived through the radiation of *Cyphostemma* and led to substantial intron expansion, a phenomenon rarely studied in eudicots. The intronic LTR-RT insertions likely enhanced tolerance of genome structural changes, facilitating succulence evolution. Genomes of succulents were further expanded by intergenic LTR-RTs, which exhibit recurrent evolutionary advantages in arid and seasonal habitats. Our study reveals how genomic landscapes are shaped by both intrinsic LTR-RT dynamics and extrinsic environmental forces. Critically, we suggest that stochastic dynamics of LTR-RT communities enhance genomic evolvability, enabling adaptive evolution in plants.

Introduction

Succulent plants, distinguished by their unique morphology and ecophysiology, are iconic components of global arid ecosystems[1,2] and serve as key model systems for studying organism-environment interactions[3,4]. Succulence has evolved independently in over 80 plant families[5], highlighting diverse evolutionary pathways leading to this adaptive trait[6,7]. Despite profound phylogenetic and morphological diversity, succulents share common features for water storage and conservation, including enlarged cell volumes and reduced stomatal density[6,8,9]. Critically, succulents also exhibit a higher frequency of larger genome sizes (GS) compared to non-succulent relatives in eudicots, suggesting a potential role of GS in drought adaptation[10], where new adaptive traits may arise through nucleotypic effects and/or genetic variation during genome expansion[11,12].

Plant genomes can expand through two primary mechanisms: whole genome duplication (WGD) and the proliferation of repeat sequences. Repeat sequences, especially long terminal repeat retrotransposons (LTR-RT), are major drivers of GS diversity in plants[13,14]. LTR-RTs are retrotransposons with internal coding regions flanked by long terminal repeats, and they proliferate via a copy-and-paste mechanism that integrates new reverse-transcribed copies into the genome. These LTR-RT lineages, characterized by diverse insertion patterns and complex interactions, act like a "community" within genomes[15]. Unlike the large jumps in GS caused by WGD, the flexible expansion/contraction dynamics of LTR-RT communities enable rapid and adaptive reshaping of genome size and function. This plasticity allows plants to respond to environmental pressures with remarkable sensitivity, underscoring LTR-RT activity as an important evolutionary force[16,17]. Nevertheless, critical gaps remain concerning the mechanisms linking environmental selection to LTR-driven GS diversity and whether genome expansion plays a role in facilitating the evolution of arid-adapted traits such as succulence.

The plant genus *Cyphostemma* (ca. 200 species), including the largest known genome in the grape family (Vitaceae; 8.68 pg/1C in *C. villosum*), exhibits striking diversity in morphology, ecology, and

GS[10,18]. Notably, most *Cyphostemma* species with enlarged genomes are succulent and thrive in arid environments, distinguishing them from their vine relatives[10]. Previous studies have reported frequent LTR-RT insertions in the genomes of Vitaceae species, including an uncommon pattern of intron expansion driven by LTR-RT activities[19–21]. Intronic LTR-RT insertions can directly influence gene expression and function through altered transcription and/or splicing, potentially leading to morphological and physiological variation and, in some cases, contributing to adaptive innovation in response to aridity[6].

Here, we investigate the contribution of LTR-RTs to arid adaptation, using newly sequenced and assembled genomes of two ecologically divergent *Cyphostemma* species, *C. dehongense*, a non-succulent climber from Asian rainforests, and *C. currorii*, a succulent shrub from the Namib desert. Utilizing a robust phylogenetic framework of *Cyphostemma* based on dense taxon sampling and an expanded nuclear gene dataset, we explore the activity of LTR-RTs and their adaptive effects, genome expansion mechanisms, and relationships with environmental variables across clades. Synthesizing multiple lines of evidence from genomics, ecology, and functional morphology, our results suggest parallel succulence evolution and a pivotal role of LTR-RTs in arid adaptation of *Cyphostemma*.

Results and Discussion

The spatiotemporal framework

To analyse LTR-RT dynamics across *Cyphostemma* species, we constructed a robust phylogenetic framework based on single-copy orthologs from 151 taxa (Supplementary Data 1). The phylogeny includes four well-supported clades: the earliest-diverging Cymosum clade from continental Africa, followed by a sister lineage comprising the Malagasy clade and Asian clade, which together are sister to the continental African clade (Supplementary Figs. 1 and 2). Although cyto-nuclear discordance and nuclear gene tree conflicts were detected among the main clades (Supplementary Fig. 3), a low signal of

hybridization and a higher incidence of incomplete lineage sorting (ILS) contributing to incongruence (Supplementary Fig. 4 and Supplementary Fig. 5) suggest that hybridization had a minimal impact on the subsequent phylogeny-based analyses.

Divergence time estimation using clock-like nuclear genes from 172 Vitaceae taxa and two outgroup taxa from Leeaceae (Supplementary Fig. 6 and Supplementary Data 2) indicates that *Cyphostemma* originated in the late Cretaceous (ca. 71.6 Ma). The four major *Cyphostemma* clades diverged rapidly during the warm Eocene (Supplementary Fig. 6), consistent with the observed high level of ILS (Supplementary Fig. 4). The succulent taxa emerged later in the Oligocene, coinciding with global shifts toward drier and cooler climates[22]. Biogeographic reconstructions indicate a tropical African origin for the crown group of *Cyphostemma*. However, dispersal history during the diversification of the four major clades is ambiguous (Supplementary Fig. 7 and Supplementary Data 3), likely due to the rapid radiation and extensive extinction of early lineages obscuring historical biogeographic signals[18].

Genome and intron expansion via LTR-RTs

We compiled 139 genome size (GS) records of Vitaceae and outgroup species from the literature[10,23,24] and novel measurements (Supplementary Data 4) and mapped these data onto a published Vitaceae phylogeny[18] with non-focal taxa pruned, to contextualize GS diversity against a broader phylogenetic background. Our ancestral state reconstruction for GS showed generally larger genomes in *Cyphostemma* and *Tetrastigma* (both from the tribe Cayratieae), two of the most species-rich genera of Vitaceae (Fig. 1a). We also tested for evidence of adaptive evolution of GS using the best-fit Ornstein–Uhlenbeck (OU) model approach (Supplementary Data 5), which assumes a combination of selective factors acting on a lineage (selective regime) driving GS towards an optimum value (θ). For *Cyphostemma* and *Tetrastigma* lineages with larger genome sizes, we detect a trend of genome upsizing accompanied

by regime shifts towards different GS optima early in their diversification (Fig. 1a and Supplementary Fig. 8).

Comparative genomic analysis of publicly available chromosome-level genome assemblies from 55 super-rosid outgroup species with seven Vitaceae species shows a disproportionate expansion of intron size in Vitaceae species (Fig. 1b and Supplementary Data 6). The strong positive correlation between total intron length and GS suggests that intron expansion contributes considerably to GS expansion (Fig. 1c). Across super-rosids, intron length distributions show three peaks, with 'troughs' at c. 250 bp and c. 4,000 bp (see Fig. 1d). Species from Vitaceae, especially *Cyphostemma*, possess remarkably more genes with extra-long introns (> 4,000 bp) and fewer genes with short introns (≤ 250 bp) compared with other super-rosid species (Fig. 1d). Notably, Vitaceae genomes, especially those of the tribe Cayratieae, exhibit significantly longer average intron lengths and a higher ratio of intron length/exon length than the other super-rosid species analysed (Fig. 1b and Supplementary Data 6). An ANCOVA test indicated that this intron expansion is lineage-specific (p = 0.017) and is significantly associated with the length of intronic LTR-RTs insertions (p = 3.58e-09; Supplementary Data 7).

Synteny analysis, which assesses large-scale conservation of gene blocks to infer genome structural evolution across divergent lineages, further revealed that no additional ancient WGDs have occurred since the divergence between *Vitis* and *Cyphostemma* (Supplementary Fig. 9; see Supplementary Note 1 for detailed interpretation). This finding indicates that Vitaceae species only underwent the γ triplication event shared by core eudicots[25]. Together, these results demonstrate that LTR-RT proliferation, instead of WGD, is the primary driver of genome and intron expansion in *C. dehongense* and *C. currorii*. This pattern is remarkable, as LTR-RTs typically accumulate in intergenic regions, and introns exceeding 4,000 bp due to LTR-RT insertions are uncommon in eudicots (Fig. 1d), where average and median intron lengths are only 619.13 bp and 523.00 bp, respectively[26,27]. The extensive accumulation of LTR-RTs

within introns of Cayratieae thus points to a distinctive and underexplored evolutionary mechanism favouring intronic retrotransposon colonization.

LTR-RT diversity in *Cyphostemma*

We investigated the repeat sequence composition of 74 *Cyphostemma* species using shallow whole-genome sequencing data (Supplementary Data 8) within a phylogenetic framework (Fig. 2a and Supplementary Fig. 10). Across the genus, genome-wide LTR-RT content is dominated by a small number of LTR-RT families (Fig. 2b), indicating that genome expansion in *Cyphostemma* is primarily driven by the massive proliferation of a limited subset of LTR-RT families. *Tekay*, *Ogre*, and *SIRE* are the three most abundant LTR-RT families and together account for the majority of GS variation compared with other repeat sequences (Fig. 2a). Notably, LTR-RT composition varies substantially among major clades of *Cyphostemma* (Fig. 2a). In the continental African clade, repeat content is dominated by *Tekay* and *SIRE*, with relatively minor contributions from *Ogre*. In contrast, the Malagasy clade is characterized by the predominance of *Ogre*, accompanied by a moderate proportion of *SIRE* and an almost complete absence of *Tekay* insertions. *Ale* ranks as the fourth most abundant LTR-RT family in the genus (Fig. 2a and Supplementary Data 8). Although *Ale* does not constitute a high percentage in genomes, it is consistently abundant across all clades of *Cyphostemma* (Fig. 2b and c), suggesting a widespread but comparatively moderate contribution to genome expansion.

We also investigated the dynamics of the four most abundant LTR-RT families in detail using the genome assemblies of *C. currorii* and *C. dehongense* to estimate their insertion times and death rates. Compared with the non-succulent *C. dehongense*, the ages of *Ogre* and *Ale* LTR-RTs were shown to be significantly younger in the succulent *C. currorii* (Fig 3a, Supplementary Data 9–11). Although the number of *Tekay* and *SIRE* lineages in *C. dehongense* is too small to support statistically robust

comparisons of insertion age distributions, their high copy numbers in *C. currorii* (Supplementary Data 9) are consistent with recent or ongoing proliferation. Within the *C. currorii* genome, *Tekay* exhibits an elevated death rate than *SIRE*, indicative of a faster turnover (Supplementary Fig. 11 and Supplementary Data 11).

Our data also indicate that *C. currorii* has more recently active *Tekay* elements than *C. dehongense* (Fig. 3a, c), despite exhibiting higher overall CHH methylation levels of LTR-RT elements compared to *C. dehongense* (Supplementary Fig. 12a). These patterns appear contradictory, as elevated CHH methylation is generally associated with the suppression of active LTR-RTs and has previously been shown to play a critical role in regulating transposon silencing near genes[28]. One possible explanation lies in the structural composition of the *Tekay* family in *C. currorii*. The high percentage of non-autonomous elements (i.e., those lacking complete, functional open reading frames (ORFs)), so either possessing incomplete or no ORFs in the *Tekay* family (Fig. 3b and Supplementary Data 9), may reduce their susceptibility to being targeted for CHH methylation-mediated silencing (Supplementary Fig. 12b). Under a genomic background of elevated CHH methylation, these non-autonomous *Tekay* elements may experience relatively weaker repression, potentially facilitating their persistence and accumulation. Such proliferation features may contribute to the dominance of *Tekay* elements observed more widely in the continental African clade, compared with the Malagasy clade where the *Ogre* family predominates (Fig. 2a, c), suggesting distinct genome expansion histories. Overall, these distinctive, clade-specific LTR-RT dynamics have generated distinct LTR-RT landscapes across different *Cyphostemma* species (Fig. 2a and Supplementary Data 8), which likely arose during the early divergence of the genus and remain evolutionarily conserved, parallel to trends documented in palms and many other plant lineages[29,30].

Dynamics and origin of intronic LTR-RTs

In the genome assemblies of *C. dehongense* and *C. currorii*, the three most abundant LTR-RT families, *Tekay*, *Ogre*, and *SIRE*, primarily contribute to intergenic region expansion, whereas the introns are mostly elongated by *Ale* elements (Supplementary Fig. 13f), which account for 58.23% and 54.65% of the total intronic LTR-RT length in *C. dehongense* and *C. currorii*, respectively (Supplementary Fig. 14 and Supplementary Data 12). *Ale* elements are particularly enriched in extra-long introns (> 4,000 bp) compared with shorter introns, indicating that they are the main contributors to substantial intron expansion in both *Cyphostemma* species (Supplementary Fig. 15).

Compared with *Tekay*, *Ogre*, and *SIRE*, *Ale* elements are generally younger, suggesting more recent activities (Fig. 3a). Notably, intronic *Ale* LTR-RTs display significantly higher expression levels than intergenic copies (Fig. 3c, Supplementary Data 10 and 11). This elevated expression is likely related to the genomic context, as insertions within introns increases the probability of co-transcription with host genes[31]. Moreover, we find significantly higher solo/intact (SI) and fragmented/intact (FI) ratios (see Methods for the definition of "solo element" and "fragmented element") in the intron-preferring *Ale* elements in *C. currorii* (Fig. 3d, e), consistent with stronger deletion pressure acting on *Ale* lineages located within introns compared with those in intergenic regions[32,33].

This pattern argues against the preferential deletion of intergenic *Ale* elements as the primary explanation of the observed intronic enrichment by *Ale* elements. The rapid turnover implied by the observed elevated deletion ratios suggests that intronic *Ale* elements may have more opportunities to impact gene functions in succulent *C. currorii* compared with other abundant LTR-RT families like *Tekay*, *SIRE*, and *Ogre*. However, *C. dehongense* shows no significant differences in SI or FI ratios between intronic and intergenic *Ale* lineages, possibly due to the high percentage of lineages with only a single element in this species (Supplementary Data 10).

We reconstructed a time-calibrated phylogeny for *Ale* elements from *C. dehongense* and *C. currorii* to trace the evolutionary origin of intronic *Ale* lineages (Fig. 3f and Supplementary Data 13). Our

generalized linear model (GLM) analysis between GC content and element age reveals a time-dependent accumulation of cytosine-to-thymine substitutions most likely due to the deamination of methylated cytosine bases over time (Supplementary Data 14), consistent with a previous study in *Brachypodium distachyon*[34]. Based on this pattern, we estimated the time tree of newly inserted *Ale* elements (those with ages very close to 0 million years) using the coding sequences of the reverse transcriptase gene and designating old elements (with ages > 0.005 million years) as "fossil" calibrations for tip dating (Supplementary Data 13). The time tree reveals that intronic *Ale* lineages originated in the Cretaceous, but their major diversification is estimated to have occurred during the Oligocene, giving rise to most of the extant abundant (with > 50 elements) intron-preferring *Ale* lineages (Fig. 3f). This timing of *Ale* element diversification also aligns with the rapid radiation and ecological adaptation of the genus *Cyphostemma* (Supplementary Fig. 6), suggesting that intronic *Ale* proliferation may have contributed to some of the genomic innovations during the diversification of the genus. In both species of *Cyphostemma* analysed in detail, genes with extra-long introns, mainly caused by *Ale* element insertions, exhibit significantly reduced expression (Fig. 4a) and more relaxed selection than other genes (Supplementary Fig. 16). This supports the hypothesis that long introns may be selected against in highly-expressed genes due to their increased transcription cost and risk of imprecise alternative splicing[35,36].

Evolution of succulence

Our analysis of *C. currorii* genes with extra-long introns (> 4,000 bp) elongated by LTR-RTs, compared with syntenic genes in *C. dehongense*, reveals distinct functional enrichment in pathways linked to transcription regulation, DNA repair, and RNA catabolism (Fig. 4b and Supplementary Fig. 17a). These genes show significantly lower expression levels compared to the non-enriched genes in *C. currorii*, likely mediated by their elevated CHG and CHH methylation levels (Supplementary Fig. 18). As DNA loss during DNA repair pathways has been reported to be an important mechanism for genome size reduction[37],

the reduced expression of genes associated with DNA repair observed in *C. currorii* may limit the extent of genome shrinkage and may even result in genome expansion due to the ongoing amplification and persistence of highly proliferative LTR-RT families like *Tekay*, *Ogre*, and *SIRE*[12,14].

In addition, we compared the length of LTR-RT insertions in all syntenic gene pairs between *C. dehongense* and *C. currorii*. Genes with longer LTR-RT insertions in *C. dehongense* tend to exhibit contraction or less elongation in *C. currorii* (Fig. 4c and Supplementary Fig. 17), consistent with a mechanism that constrains excessive intron expansion to maintain gene function[35]. However, genes related to succulence development, as identified in previous studies[6] (Supplementary Data 15), deviate from this trend. In particular, cyclin- and auxin-related genes regulating cell size and vascular patterning show significant, positive correlations between intronic LTR-RT insertion lengths in *C. dehongense* and the elongated intronic insertions in *C. currorii* (Fig. 4d and Supplementary Fig. 17). These correlations suggest sustained functional suppression linked to continuing LTR-RT insertion. Consistent with this interpretation, several succulence-related genes show gene ontology (GO) terms enriched in elongated genes (Fig. 4b), particularly those involved in mitosis processes regulated by auxin and cyclin (Supplementary Data 15). Furthermore, succulence-related genes with significantly divergent expression levels between succulent and non-succulent *Cyphostemma* species are exclusively associated with auxin, cyclin, and Ribosomal S6 kinases (S6K), which are important in cell cycle regulation (Supplementary Data 16). Reduced activity of auxin, cyclin, and S6K genes can lead to increased cell DNA content and hence cell expansion through endoreduplication[38–41], as observed in succulent lineages of *Cyphostemma*[10]. Elevated DNA content further correlates with reduced stomatal density[42,43], another characteristic feature of succulent plants that enhances water retention under arid conditions[2,44]. Additionally, suppressed auxin-related genes are suggested to regulate the formation and localization of vascular tissues, which are vital for water-use efficiency[6].

We examined leaf-surface features using scanning electron microscopy across 22 *Cyphostemma* and outgroup species. This analysis revealed four distinct leaf-surface types (Supplementary Fig. 19 and Supplementary Note 2), supporting the independent evolution of succulence multiple times in this genus and confirming general differences between succulent and non-succulent species. Succulent species (n = 13) exhibit significantly lower stomatal density and greater variability in stomatal size on the abaxial leaf surface compared to non-succulent species (n = 9) (Supplementary Fig. 20a and Supplementary Data 17). However, these characteristics are not significantly correlated with the measured GS (Supplementary Fig. 20b), suggesting alternative drivers, such as tissue-specific DNA content increase via endoreduplication and/or environmental factors, influencing this relationship[45,46].

Importantly, succulence develops independently multiple times against a background of a largely conserved LTR-RT family composition within clades, a pattern also observed in other plant groups[29,30]. This indicates that major shifts in overall LTR-RT landscapes are not required for the evolution of succulence. Instead, regulatory effects of LTR-RT insertions (e.g., impact on methylation levels and transcription) at particular genes related to succulence development may be sufficient to generate repeated origins of succulence within *Cyphostemma*.

Larger genomes selected by arid habitats

A recent macroevolutionary study demonstrated that succulent *Cyphostemma* species in arid habitats tended to have larger genomes and extended this pattern to eudicots more broadly[10]. Building on these findings, herein we have integrated more environmental factors (such as temperature and seasonality, see Supplementary Data 18) and undertaken a detailed analysis of the repeat compositions for *Cyphostemma* species to investigate the genomic mechanisms underlying GS variation and adaptive trait evolution. We first investigated the temporal sequence of genome expansion and succulence across the time-calibrated

phylogeny of *Cyphostemma*. Ancestral state reconstruction of GS reveals that 72.8% of succulent species underwent regime shifts towards larger GS during their evolution (Fig. 2a and Supplementary Fig. 10). However, regimes differ across clades: the continental African clade exhibits general GS expansion, whereas the Malagasy clade shows rapid initial GS increase followed by sharp declines (Fig. 2a and Supplementary Fig. 10). Critically, succulence frequently arose before regime shifts in GS, indicating that significant genome expansion is not a prerequisite for succulence evolution (Fig. 2a). Stochastic character mapping linked succulence to habitat shifts into arid environments (Supplementary Fig. 21), suggesting that genome expansion in succulent *Cyphostemma* lineages was likely driven by novel selective pressures in arid habitats.

Given the distinct GS regimes and LTR-RT compositions observed across clades (Fig. 2a), we used phylogenetic generalized least squares (PGLS) regressions to assess relationships between environmental variables and GS and LTR-RT proportions in each clade. The analyses incorporated uncorrelated environmental variables derived from climate databases based on filtered distribution data (Supplementary Fig. 22, Supplementary Data 18 to 21). For the continental African clade, higher GS is significantly correlated with a greater range in annual temperature (bio7) and reduced precipitation of the coldest quarter (bio19) (Fig. 5a and Supplementary Data 19). The best model also includes isothermality (bio3), which shows a positive, yet non-significant, relationship with GS. These findings align with the arid, seasonal habitat characteristics of succulent species in the continental African clade, where increased seasonality and reduced precipitation of the coldest quarter may play a role in driving GS regime shifts (Fig. 5b). For the Malagasy clade, the best model reveals a positive, but non-significant, relationship between precipitation seasonality (bio15) and GS (Fig. 5a and Supplementary Data 19). Similarly, succulent and non-succulent species on the island exhibit non-significant differences in precipitation seasonality (Fig. 5b).

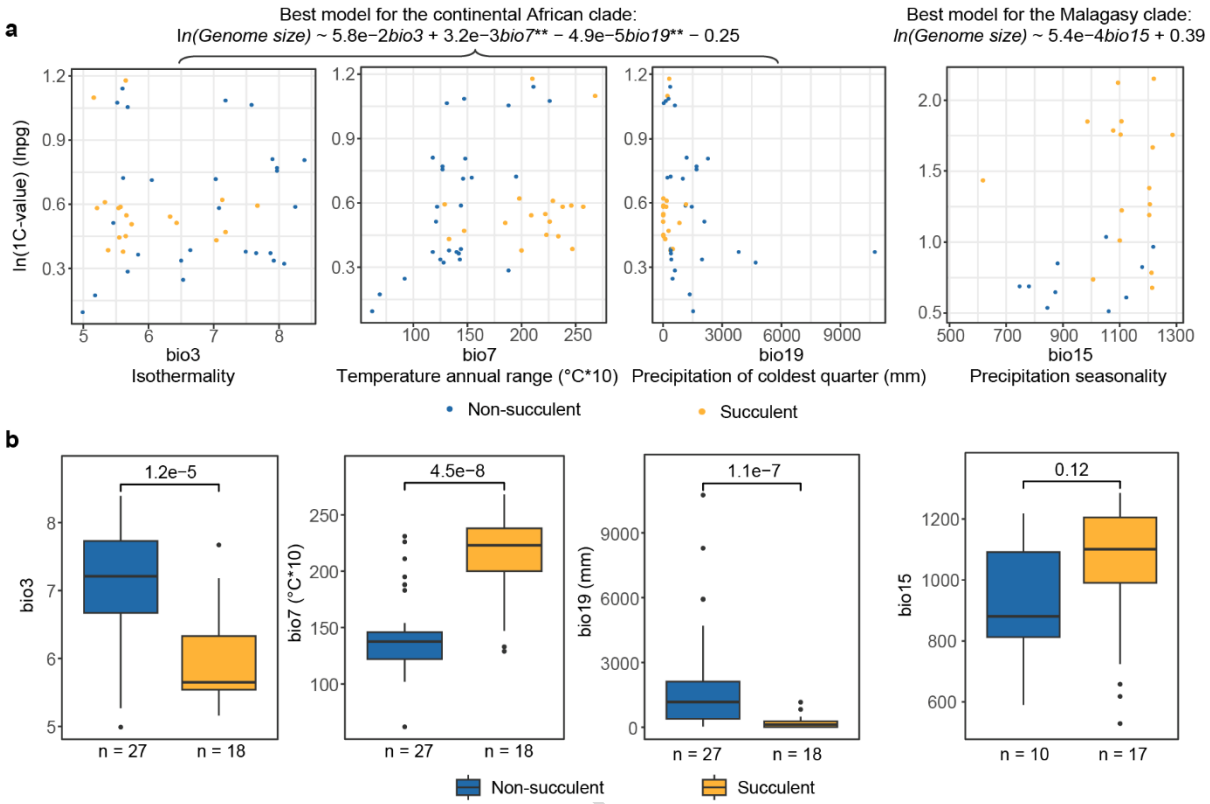

Fig. 5 | Environmental variables contributing to genome size variation. a, Relationships between genome size (1C-value) and environmental variables in the best regression model for the continental African clade (n = 45) and Malagasy clade (n = 27). "**" represents $p < 0.01$. Correlations between genome size and environmental factors were tested using the phylogenetic generalised least squares approach. **b**, Differences in the reported environmental variables in the habitats of succulent lineages of both clades, compared by unpaired two-sided Wilcoxon signed-rank tests. 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.

To test whether environmental factors contribute to genome expansion through the proliferation of specific LTR-RT families, we performed PGLS regressions between environmental variables significantly related to GS and the four most abundant LTR-RT families: *Tekay*, *SIRE*, *Ogre*, and *Ale* (Supplementary Data 22 to 24). Across both the continental African and Malagasy clades, none of the tested LTR-RT families shows a significant linear relationship with genome expansion (Supplementary Data 22). Within the continental African clade, annual temperature range exhibits a negative correlation with the proportion of *Tekay* (Supplementary Data 25). In contrast, no significant associations between

the proportion of LTR-RT families and environmental variables were detected in the Malagasy clade (Supplementary Data 26). Overall, these results provide limited evidence that environmental factors play a role in driving genome expansion through the selective amplification of specific LTR-RT families. Instead, our results suggest that environmental selection may act primarily on overall GS rather than on the composition of individual LTR-RT components.

We further investigated associations between GS and finer, local-scale environmental factors[47], identifying soil nitrogen as the dominant factor in both clades, but with different patterns (Supplementary Fig. 23 and Supplementary Data 27 and 28). In the continental African clade, GS shows differential responses to nitrogen levels across soil depths. In the Malagasy clade, the best model shows a non-significant negative correlation between GS and surface soil nitrogen (0–5 cm depth), consistent with shallow-rooted species in highly heterogeneous arid environments on the island. Despite ongoing genome downsizing (Fig. 2a and Supplementary Fig. 10), the Malagasy clade exhibits a limited capacity for GS reduction under current nitrogen scarcity, a paradox underscoring the need for broader comparative studies to disentangle the complex interplay between genomic constraints and ecological pressures in plant lineages growing in nutrient-limited environments.

Concluding Remarks

By integrating genomic, ecological, and phenotypic data, our study provides insight into how LTR-RT proliferation, a key mechanism of GS variation, likely contributed to the evolution of succulence in *Cyphostemma* and its association with arid habitats (Supplementary Fig. 24). Comparative analyses of the genome assemblies for the non-succulent *C. dehongense* and the succulent *C. currorii* highlight a potential role for previously underexplored intronic LTR-RT insertions in enhancing evolvability and contributing to the emergence of succulence. While the precise pathways linking LTR-RTs to succulence are yet to be

elucidated, we identify genes associated with auxin, cyclin, and S6K as candidates for future functional validation. Such genomic diversity and the resulting succulent traits may have originated in arid environments or, alternatively, emerged as pre-adaptations in ancestral lineages occupying more mesic environments. In the latter scenario, such pre-adaptations may have contributed to enhancing the survival of those species possessing more succulent traits as climate change led to increasingly arid conditions. Ultimately, extensive genome expansion driven by LTR-RT accumulation in both intergenic regions and introns and their impact on genome function and expression are likely to have further enhanced the adaptive potential and evolutionary persistence of *Cyphostemma* lineages in arid habitats.

Once dismissed as "junk DNA", active LTR-RTs are increasingly recognized as a dynamic reservoir for evolvability of adaptive innovation, and thus directional adaptation may arise from the stochastic evolution of LTR-RT communities. The parallel adaptive outcomes observed across *Cyphostemma* highlight the flexibility of evolutionary trajectories driven by LTR-RTs, which may prove vital for plant adaptation under global aridification. Future comparative genomic studies incorporating fine-scale environmental gradients and broader taxonomic sampling of succulent and non-succulent sister lineages from both continental Africa and Madagascar could help further clarify how genomic architecture mediates adaptive responses to ecological pressures in plants.

Methods

Taxon sampling and DNA sequencing

We collected shallow whole genome sequencing (WGS) data from 124 accessions of 118 *Cyphostemma* taxa, and 54 additional species of Vitaceae and two Leeaceae species as outgroup for biogeographic and repeat composition analyses. Plant materials from Madagascar were collected under the scientific permit issued by the Ministry of Environment and Sustainable Development in Madagascar (N_025/23/MEDD/SG/DGGE/DAPRNE/SCBE.Re), and materials from Kenya were collected under the permit

issued by the Ministry of Environment, Climate Change and Forestry in Kenya (RESEA/1/KFS 98 and RESEA/1/KFS 22). All collections comply with relevant local and national regulations. Remaining plant materials were sourced from previously established collections where no additional permits were required. DNA was extracted from silica-gel dried leaves of 115 individuals representing 109 *Cyphostemma* species and 11 outgroup species using the modified CTAB method. For each sample, 6–10 Gb of 150 bp paired-end reads were generated using an Illumina HiSeq2500 sequencer. We also compiled WGS data for nine *Cyphostemma* taxa and 43 outgroup taxa from SRA database of NCBI[48].

For genome assembly of *C. dehongense* and *C. currorii*, we first surveyed genome traits. We surveyed the genome traits of *C. dehongense* and *C. currorii* (i.e., genome size, heterozygous ratio, percentage of repetitive sequences) before the formal process of genome sequencing. Fresh young leaves were collected from *C. dehongense* and *C. currorii* cultivated in our greenhouse at China National Botanical Garden. DNA was extracted using QIAGEN® Genomic Kits following regular sequencing instructions from the manufacturer. The quality of the extracted DNA was evaluated using a 1% agarose gel to assess any degradation or contamination. The DNA concentration was measured using a Qubit 4.0 Fluorometer (Invitrogen, USA).

For both species, DNA sequences were fragmented into ~350 bp, and libraries were constructed after end repair, dA tailing, and adapter ligation. Raw reads were filtered using fastp v.0.23.2[49] with parameters "--average_qual 15 -l 150 -w 6", discarding sequences containing bases of low quality and adaptor sequences. Based on the cleaned data, the K-mers were counted using Jellyfish v.2.2.6[50], which were used to estimate the genome traits of the two species in GenomeScope v2.0[51]. Our genome survey revealed low heterozygous ratio and a high repetitive sequence percentage in both species: *C. dehongense* (~1.1 Gb; 0.47% heterozygosity; ~67.8% repetitive sequences) and *C. currorii* (~1.7 Gb; 0.63% heterozygosity; ~75.1% repetitive sequences). These results guided the subsequent long-read sequencing and assembly.

For *C. dehongense*, a PacBio SMRTbell library with ⩾15 kb size selection was prepared and sequenced on a PacBio Sequel platform (Pacific Biosciences). HiFi reads were generated with SMRT Link v.6.0 (https://www.pacb.com/smrt-link/). For *C. currorii*, an Oxford Nanopore Technology (ONT) library with 3–4 μg of DNA per sample was prepared, which was sequenced using a Nanopore PromethION sequencer instrument

(Oxford Nanopore Technology, UK). Raw fast5 sequencing files were converted to fastq format by the base-calling software Guppy v.5.0.13 (https://nanoporetech.com/document/Guppy-protocol) on the Oxford Nanopore Technologies™ sequencing platforms. Sequencing quality was assessed with NanoPlot v1.38.1[52]. For both species, Hi-C libraries were constructed following Rao et al.[53] and sequenced for 150 bp pair-end reads on the Illumina NovaSeq 6000 platform.

RNA sequencing and assembly

We sampled fresh mature leaves, stems, tendrils, and young roots from *C. dehongense*, and fresh mature leaves, stems, flowers from *C. currorii* cultivated in our greenhouse at China National Botanical Garden for genome annotation. To compare expression differences in succulent-related genes, we also sampled fresh mature leaves of 16 other *Cyphostemma* species (Supplementary Data 1).

For Illumina RNA-Seq data, total RNA was extracted from samples using the modified CTAB method. PolyA RNA-Sequencing libraries with an insert size of 150 bp were then constructed using the NEBNext Ultra RNA Library Prep Kit for Illumina (NEB, USA). The pair-end cDNA libraries were constructed and sequenced with an Illumina4000 instrument. Raw reads were trimmed using Trimmomatic v.0.39[54] with the parameters "SLIDINGWINDOW:4:20 LEADING:20 TRAILING:20 MINLEN:50". Potential organelle reads were filtered with Bowtie2 v.2.4.5[55]. Transcripts were assembled using Trinity v.2.15.1[56,57] with default settings, and the assembly quality was checked using TransRate v.1.0.1[58] and BUSCO v.5.2.2[59]. For each gene, the longest isoform was predicted and translated using TransDecoder v.5.5.0[57], and the redundant isoforms were removed using CD-HIT v.4.7[60] with a threshold value of 0.99.

Assembly of genome size records, morphological traits, and habitats

We collected GS records of Vitaceae and Leeaceae species from (1) previous literature, (2) the Plant DNA C-values Database (Release 7.1; Accessed Feb 2024)[61], and (3) new GS flow cytometry measurements. The flow cytometry measurements are conducted following Galbraith et al.[62]. We prepared nuclei suspensions from a chopped dried leaf sample of ~1 cm^2 in 2 ml Galbraith modified buffer, which was then filtered with a 30 μm nylon

mesh. Nuclei were stained with 50 μg/ml propidium iodide (PI) and 0.5 μg/ml RNase. The nuclear DNA content for each sample was estimated by recording > 5,000 particles using a BD LSRFortessa™ Cell Analyzer. For each individual, we first determined the appropriate voltage and reference standard (RS) by running a sample of the target species, as the genome size of Vitaceae varies greatly. The RS we selected from the following species: *Oryza sativa* subsp. *japonica* "Shig. Kato" (1C-value = 0.43 pg)[63], *Solanum lycopersicum* (wild; 1C-value = 0.75 pg), *Glycine max* "Williams 82" (1C-value = 1.11 pg)[64], *Zea mays* CE-777 (1C-value = 2.72 pg)[65], and *Capsicum annuum* "L. Zunla-1" (1C-value = 3.42 pg)[66]. After the RS was selected, we chopped the leaf samples of the target species and RS together to prepare the homogenate for flow cytometry. Gating was firstly performed on a dot plot displaying side scatter area (SSC-A) versus PI fluorescence area (PI-A) to exclude debris, dead cells, and electronic noise. The G1 peak of the target species and reference standard was defined according to the VL1-A histograms. (Supplementary Fig. S25). Coefficients of variations (CVs) for both G1 peaks were < 5% in all accepted samples. The absolute DNA amount (1C-value; pg) of the target species was calculated following the formula below:

$$1C\ of\ target\ species = \frac{G_1\ nuclei\ peak\ position\ of\ target\ species\ \times 1C\ \text{value}\ of\ RS}{G_1\ nuclei\ peak\ position\ of\ RS}$$

The final genome size estimate for each individual was calculated as an average value of three independent measurements.

For species with multiple records, we used the average value when multiple flow cytometry records existed; otherwise, we selected the most recent flow cytometry measurement. Habitat (habitat type and aridity scale) and morphology data (succulence, tuber, tendril, and trichome) were collected from floras, herbarium specimens, and field observation.

Genome assembly

We used Nextdenovo v.2.4.0[67] to assemble ONT sequencing data into contigs with the parameters "random_round = 20 minimap2_options_cns = -x ava-ont -t 10 -k17 -w17 nextgraph_options = -a 1". The contigs were corrected using Nextpolish v.1.4.0[68], with the parameters "lgs_options = -min_read_len 1k -max_read_len 100k -max_depth 100 lgs_minimap2_options = -x map-ont". The polished contigs were clustered, ordered, and anchored according to the interaction information shown by Hi-C short reads by pipeline ALLHiC v.5.18.2[69], with

the parameters "--enz DpnII --CLUSTER $chr_num --break N --fill Y". Assembly errors were manually corrected using Juicebox v.1.13.01[70]. Genome assembly completeness was evaluated using BUSCO v.5.2.2[59], compared with the udicotyledons_odb10 database.

Repeat annotation

Tandem repeats were extracted using the software TRF v.4.09.1[71]. RepeatMasker v.4.1.5[72] was used to identify repeats by searching against the genomes of *C. dehongense* and *C. currorii* with the RepBase of *Arabidopsis thaliana* (released on 26th Oct 2018)[73] with default settings. A complete library of LTR-RT was built combining the repeat sequences identified by RepeatModeler v.2.0.5[72], LTR_FINDER v.1.1[74], and LTRharvest v.1.6.4[75], and sequences < 100 bp or containing > 5% of gap "N" were discarded. Redundancies were discarded using LTR_retriever[76] following an 80-90-100 rule (LTR-RTs with > 80% sequence identity, covering > 90% of the longest sequence, and > 100 bp in CD-HIT clustering were regarded as redundancy). The family of remaining LTR-RT elements was identified using TEsorter v.1.4.6[77]. The complete library including intact LTR-RT elements was refined using LTR_retriever v.2.9.9 and custom Perl scripts by checking the intactness of LTR-RT elements. Truncated LTR-RT elements were identified as LTR-RTs with two complete or nearly complete LTRs from the same family, but missing terminal structural components, including the dinucleotide palindromic motifs flanking LTR and the 5-bp target site duplication (TSD) flanking the element. Solo LTR-RT elements were identified as a single LTR flanked by dinucleotide palindromic motifs and TSDs. LTR-RT fractions could not be classified as intact, truncated, or solo LTR-RT were defined as "fragmented LTR-RT".

Gene and non-coding RNA annotation

Gene prediction combined the methods of *ab initio* prediction, homology-based prediction, and transcriptome evidence, with an additional check for gene annotation with long introns. For *ab initio* prediction, we used AUGUSTUS v.3.2.3[78], Geneid v.1.4[79], Genescan v.1.0[80], GlimmerHMM v.3.04[81], and SNAP (version 2013-11-29)[82] to annotate genes on the repeat-masked scaffolds. For homolog prediction, gene models were annotated according to the protein similarity in alignments with GeneWise v.2.4.1[83], using reference genomes from

Arabidopsis thaliana, *Fragaria vesca*, and *Vitis vinifera* downloaded from the Phytozome database[84]. To incorporate transcriptome evidence, RNA-Seq reads from different tissues (leaf, stem, root, flower, and tendril) were aligned to the genome using Hisat v.2.0.4[85] with default parameters to identify exons and splice positions, and genome-based transcripts were assembled by Stringtie v.1.3.3[86]. PASA v.2.4.1[87] was used to predict gene models from repeat-masked genome assembly. To exclude annotation artefacts, we removed gene annotations lacking RNA-Seq support, detectable homology or conserved protein domains, or fragmented ORFs. Results from the above methods were integrated to generate the final annotated gene set in EvidenceModeler v.1.1.1[88].

The putative gene functions were assigned according to the best BLAST hit in the Swiss-Prot database and the non-redundant (NR) database in NCBI (E-value < 1e-5). The motifs and domains were annotated using InterProScan v.5.31[89] by searching against public protein databases, including ProDom, PRINTS, Pfam, SMART, PANTHER, and PROSITE. Gene Ontology (GO) terms were assigned to genes according to the corresponding InterPro entry. We also inferred putative gene pathways by obtaining the best matches in the KEGG database.

For non-coding RNA annotation, rRNAs were predicted using the software tRNAscan-SE, and other ncRNAs were identified by searching against the Rfam database with default parameters using infernal v.1.1.3[90].

Methylation level estimation

Fresh leaves of *C. dehongense* and *C. currorii* were collected from our greenhouse for bisulfite sequencing. Genomic DNA was extracted using the modified CTAB method. High-quality DNA was fragmented into 200–400 bp and bisulfite-treated using the EZ DNA Methylation-GoldTM Kit (Zymo Research, USA). The bisulfite-treated DNA fragments were purified and amplificated to prepare sequencing libraries using the Accel-NGS Methyl-Seq DNA Library Kit (Swift, USA). The raw bisulfite sequencing reads were filtered and mapped to the reference genomes of *C. dehongense* and *C. currorii* using Bismark v.0.24.0[91] with the parameters "-X 700 --dovetail". Methylated cytosines were identified with a binomial test employing the conversion rate as the expected probability, where FDR $p < 0.05$ suggested a methylation site. The mean methylation levels of genes and LTR-RTs were calculated following the methods in Schultz et al.[92].

Phylogeny reconstruction

The phylogeny of *Cyphostemma* was reconstructed based on 1,367 single-copy orthologs extracted from species of tribes Cayratieae and Cisseae. The phylogenetic trees of *Cyphostemma* were constructed in a previous study[93] with plastid coding sequences and 229 single-copy nuclear genes across the grape family. However, these phylogenetic trees failed to resolve the relationships among major clades of *Cyphostemma* and showed cyto-nuclear conflicts[10,94], impeding our downstream analyses within a robust phylogenetic framework. We herein attempted to build a well-supported phylogeny of *Cyphostemma* based on more single-copy orthologs from species of the tribe Cayratieae and Cisseae.

We selected transcriptomes of *Causonis ciliifera* and *Pseudocayratia speciosa* assembled from our previous study[95], and genomes of *C. currorii*, *C. dehongense*, *Tetrastigma hemsleyanum*[20], and *Cissus rotundifolia*[96] for ortholog identification. The accessions for genomes downloaded from public databases are listed in Supplementary Data 6. Using OrthoFinder v.2.5.4[97] with the multiple sequence alignment method and gene trees constructed by fasttree v.2.1.11[98], 6,063 potential "single-copy orthologs" were identified. After conducting reciprocal BLASTN v.2.14.1+[99], orthogroups containing genes with multiple hits within the same species or with an E-value > 10^{-6} were removed. The 5,223 orthogroups retained were aligned in MAFFT v.7.520[100] with the L-INS-i algorithm, and the alignments were trimmed by trimAl v.1.4.rev22[101] with the option "automated1". Orthogroups with alignment length < 450 bp were further removed, retaining 4,138 orthogroups. We used TreSpEx v.1.1[102] to filter out genes with misleading phylogenetic signals, i.e., orthogroups that do not support relationships confirmed by previous studies[103], orthogroups whose maximum likelihood (ML) tree built by IQ-TREE v.2.3.4[104] showed extremely long or short branches (longer than four times the average branch length or shorter than 0.00005 mutation units), orthogroups with a saturation slope < 0.4 or R^2 < 0.6, and orthogroups with a standard deviation of LB score > 39 or a mean value of the upper quantile of the LB score > 35. The thresholds used for filtering anomalies were determined according to the corresponding distribution plot. After the above filtering processes, 3,035 orthogroups were retained. Then we checked the synteny of the retaining genes using MCScanX v.1.0.0[105] with genomes of *C. currorii*, *C. dehongense*, *T. hemsleyanum*, and *C. rotundifolia*. Finally, a total of 1,367 orthogroups containing only syntenic genes were retained, which were used as target sequences of single-copy orthologs in the downstream analyses.

Coding sequences of the target single-copy orthologs were extracted from WGS data of *Cyphostemma* and outgroup species, using HybPiper pipeline v.2.1.6[106] with settings "no_stitched_contig"[106]. All the extracted sequences were aligned using MAFFT with the L-INS-i algorithm and trimmed with trimAl[101]. Alignments for orthologs were manually checked. Gene trees were constructed in IQ-TREE v.2.3.4[104] using the best model selected from K80, HKY, and GTR, with 1000 bootstrap replicates.

The multispecies coalescent (MSC) trees were generated using ASTRAL v.5.7.8[107] input with 1,367 gene trees. The nodes of gene trees with a bootstrap support (BS) value < 10% were collapsed following Zhang et al.[108] to avoid the influence of poorly supported clades. For the concatenated method, the phylogeny was constructed in IQ-TREE based on the concatenated alignments and the best partition scheme determined by ModelFinder[109], using a maximum likelihood (ML) approach with 5,000 ultrafast bootstrap[110] under the edge-unlinked partitioned model.

Gene tree conflict analysis

We estimated gene tree conflicts with gene tree concordance factor (gCF)[111], the internode certainty all (ICA) values[112], and the quartet sampling (QS) analysis[113]. The gCF values were calculated in IQ-TREE, with the MSC tree as the mapping tree. To quantify discordance among gene trees, we calculated the ICA value for each node using the Phyparts pipeline[114] based on the MSC tree and gene trees with nodes < 70% BS value collapsed, during which the number of concordant and conflicting bipartitions was also calculated. To overcome the effects of misleading gene tree topologies from alignment sparsity and weak support, we performed QS analysis on the MSC tree with 100 replicates and obtained the scores of confidence (QC, Quartet Concordance), consistency (QD, Quartet Differential), and informativeness (QI, Quartet Informativeness) of internal branches.

We also estimated the contributions of three prominent factors, i.e., hybridization, incomplete lineage sorting (ILS), and gene tree estimation error[115,116], that led to gene tree conflicts. The assessment was performed following the pipeline in Cai et al.[116]. To better detect the factors causing gene tree conflicts among major clades and reduce computation burden at the same time, we selected representative species for assessment (Supplementary Data 1), with *Tetrastigma serrulatum* as the outgroup. ML gene trees and a concatenated tree of representative species were built in IQ-TREE, and an MSC tree was constructed using ASTRAL, with the methods described above. The level of gene tree conflict was quantified by gCF values calculated with IQ-TREE. Introgression can be detected by the

deviation of the frequency of triplets from the MSC model. The triplets represent the possible topologies of three species: ((A,B),C), ((A,C),B) and ((B,C),A). If the topologies are only affected by ILS, which can be eliminated by the MSC model, the two minor discordant triplets will occur at the same frequency. However, introgression can lead to a discrepancy in the frequency of the two minor triplets. The level of introgression at each node was quantified by the reticulate index[116], which is the triplets exhibiting introgression among all the triplets. We obtained bootstrap species trees based on the bootstrap gene trees using ASTRAL, and simulated 1,367 gene trees for each bootstrap species tree under the MSC model using the R package Phybase v.1.4[117]. For each node, the triplet frequency of both the empirical and simulated gene trees was counted, and then the reticulate index was calculated, showing the nodes with significantly unbalanced triplets.

The level of ILS was estimated by the population mutation parameter "θ", which was calculated by dividing the branch length in mutation units (obtained from the dated ML tree) by the branch length in coalescent units (obtained from the MSC tree). To assess the effects of gene tree estimation error, we simulated 1,367 gene alignments with 1,771 bp (the average length of empirical gene alignments) based on the MSC tree with Seq-Gen v.1.3.4[118]. Then we constructed gene trees with the simulated alignments using RAxML v.8.2.12[119] with the "-f b" option, to calculate the frequency with which each node of the empirical species tree can be recovered by gene trees. Finally, we estimated the relative contributions of each factor using the linear regression model implemented in the R package relaimpo v.2.2-6[120].

Potential gene flow between the major clades of *Cyphostemma* was checked using the Species Networks applying the Quartets (SNaQ) algorithm implemented in the Julia package PhyloNetworks v.0.12.0[121], using the MSC tree as the starting tree. The network generated with the best maximum hybrid node number was used as the starting network for bootstrap analysis with 100 replicates. For each replicate, 100 SNaQ searches were performed.

Biogeographic analysis

One representative for each *Cyphostemma* species was selected for divergence time estimation. We extended the sampling to the family level (Supplementary Data 1 and 2) to incorporate reliable calibration points for Vitaceae. We used reliable fossil and secondary time calibrations to date the phylogeny of Vitaceae. The seed fossil of

Ampelocissus parvisemina from the late Paleocene in North Dakota of North America[122] was used to constrain the crown age of the *Ampelocissus–Vitis* clade as 56.8–62.0 Ma. The fossil of *Vitis glabra* Chandler from the late Eocene in the lower Bagshot beds of the London Clay of southern England[123,124] was used to constrain the lower bound of the crown age of subg. *Vitis* as 34.0 Ma. The upper stem age of Vitaceae was constrained as 91.7 Ma, following the secondary calibration from Magallón and Castillo[125]. We employed a conservative calibration approach, utilizing uniform prior distributions with soft bounds and a 2.5% prior density.

To reduce missing data, 229 single-copy orthologs identified in Vitaceae[93] were extracted with HybPiper for phylogeny reconstruction in ASTRAL. Divergence times were estimated based on 50 clock-like genes selected by SortaDate[126], using MCMCTree implemented in PAML v.4.10.7[127], which was run with a birth-death model, independent rates, and an HKY85 substitution model (alpha = 0.5). One sample was taken for every 1,000 generations, totalling 10,000 samples (the first 20% discarded as burn-in). Effective sample sizes for all the parameters were confirmed > 200 in Tracer v.1.7[128]. The estimation was conducted twice independently to ensure the convergence.

Ancestral area reconstructions were conducted on the pruned time tree of Vitaceae, retaining *Cyphostemma* and the outgroup species from Cayratieae. The distribution range was divided into six areas (see Supplementary Fig. 7): (A) tropical continental Africa, (B) southern Africa, (C) Mauritius, (D) Asia (including continental Asia, the Indian subcontinent, and the Malesian region), (E) Madagascar, and (F) Australasia (including continental Australia and New Guinea). The ancestral areas were estimated using the R package BioGeoBEARS[129] implemented in RASP v.4.2[130]. The best model was selected from DEC[131], DIVALIKE[132], and BAYAREALIKE[133] based on the corrected Akaike information criterion (AICc).

We also reconstructed the ancestral state of succulence and habitats of *Cyphostemma* species with a Bayesian stochastic character mapping approach implemented in the R package phytools v.1.0-1[134] based on the dated phylogeny with two MCMC chains and 5,000 runs (first 20% as burn-in).

Genome comparison in super-rosids

To investigate the expansion patterns of genome size and intron size in Vitaceae, we compared the genomes of Vitaceae species in the context of super-rosids. We downloaded high-quality chromosome-level genomes assembled using long-read sequencing data for representative super-rosids species from public databases (Supplementary Data 6). Genomes selected for comparative analyses met the following criteria: 1) reads depth > 30× (except for *Vicia faba* with a huge genome of 11.9 Gb); 2) assembly scaffold N50 > 5Mb; 3) complete BUSCO > 90%; 4) gene annotation corrected with external evidence; and 5) repeat annotations are provided and confirmed by multiple lines of evidence. We investigated the correlations between GS and total intron length, average intron length, and intron/exon ratio using the "lm" function in R. Repeat sequences in the downloaded genomes were identified and classified using the same method as for genomes of *C. dehongense* and *C. currorii*. We also conducted a two-sided ANCOVA in R to test whether genome size, intronic LTR-RT length, and specific lineage contribute to the intron/exon length ratio.

Genome size variation in Vitaceae and *Cyphostemma*

Ancestral states of GS in Vitaceae and *Cyphostemma* were reconstructed using the ML method with the R package phytools[134]. For the family-level reconstruction, to maximize species inclusion, we pruned the plastome phylogeny of 495 taxa from our previous study[18] to retain species with available GS data. For *Cyphostemma*, we used the MSC tree based on 1,367 single-copy orthologs.

For the GS evolution patterns in Vitaceae and *Cyphostemma*, we specifically tested (i) whether lineage-specific optimal GS values emerged during the diversification of *Cyphostemma*, and (ii) if regime shifts led to these optimal GS values. We selected the best-fit model of genome size evolution in Vitaceae and *Cyphostemma* with the R package OUwie v.2.6[135]. The BM1 model is the simplest Brownian motion model with a single stochastic evolution rate σ^2, while the BMS model allows different σ^2. Ornstein–Uhlenbeck (OU) models describe trait evolution with the rate of adaptive evolution (α) towards the optimum value (θ), together with σ^2. Different OU models have distinct assumptions about their parameters: OU1 has a single θ, OUM allows θ to vary across regimes, OUMV allows both θ and σ^2 to vary, OUMA allows θ and α to vary, and OUMVA allows θ, α, and σ^2 to vary. For models that allow multiple optimum values in different clades, we assumed that genome size evolved differently

in succulent and non-succulent taxa. The best models for genome size evolution in Vitaceae and *Cyphostemma* are OUMVA and OUMA, respectively, according to the AICc value (Supplementary Data 5).

For Vitaceae and *Cyphostemma*, the positions of discrete regime shifts in genome size were identified[136] in the R package bayou v.2.1[137]. This model assumes that, for each lineage represented by a branch, there is a regime selecting for an optimum trait value (θ) during the evolutionary history of the lineage. θ was determined by the parameter α, and the rate away from θ was described by a constant parameter σ^2. Two independent chains were run for 500,000 generations and sampled every 200 generations, with the first 20% of generations discarded as burn-in. We checked the effective sample size of each parameter in the model with the "summary" function in the R package bayou to ensure they were > 200. Finally, the regime shifts and the optimum genome size for each branch were calculated and visualized using the functions "plotSimmap.mcmc" and "plotBranchHeatMap", respectively, with a posterior probability cut off of 0.3.

Repeat sequence composition

We investigated repeat sequence compositions in *Cyphostemma* species with shallow WGS data (Supplementary Data 1) using RepeatExplorer2 in the Galaxy platform[138]. The depth of our WGS data ranges from 0.2–20×, meeting the requirements for RepeatExplorer2 (recommended depth: 0.1–0.5×)[138]. We trimmed the reads to 100 bp and discarded low-quality reads using the "Preprocessing of fastq paired-end reads" tool. For each species, 2,500,000 read pairs were randomly chosen using the software seqtk v.1.4[139], attaining 0.1–0.5× genome coverage following RepeatExplorer2 protocol[138]. Repeat sequences were identified by similarity-based clustering through TAREAN pipelines[140], followed by manual correction. Repeat abundances were normalized according to GS and were plotted with the R script "plot_comparative_clustering_summary.R" in RepeatExplorer2. For each clade of *Cyphostemma*, we calculated the mean proportions of different repeat sequences based on the data from all sampled species from that clade.

Identification of LTR-RT lineages

We identified LTR-RT lineages in *C. dehongense* and *C. currorii* by combining phylogeny, clustering similarity, and alignment dot plots of LTR-RTs. As suggested by Seberg and Peterson[141], "homology" is the decisive factor in assigning LTR-RTs into different lineages, and thus the classification of LTR-RT should follow their phylogenetic relationships. Therefore, we constructed an ML phylogeny of LTR-RT sequences in the non-redundant library assembled by LTR_retriever, and coarsely divided LTR-RTs into different lineages according to the tree first. The tree was built in IQ-TREE with the LG models and 5,000 ultrafast bootstraps, based on the amino acid (AA) sequences of reverse transcriptase (RT) identified by TEsorter, which evolves at a moderate rate and can better reflect homology. LTR-RT lineages were divided according to the tree shape, where in most cases, a single sequence represents a lineage, and multiple sequences were grouped into one lineage only when they are closely clustered, as the tree was constructed based on the non-redundant library. LTR-RTs in the non-redundant library lacking RT sequences were each considered to represent a lineage temporarily. The remaining LTR-RTs from the complete library were then classified into different families using TEsorter and into different lineages by BLASTN. Sequences that failed to match LTR-RT sequences in the non-redundant library under the 80-90-100 rule were considered to represent a new lineage. We concatenated the LTR-RT sequences from the same lineage and confirmed the sequence similarity within the lineage using an alignment dot plot in Geneious v.2021.2.2[142]. LTR-RT sequences distinct from other sequences in the same lineage were picked out to form a new lineage. Then the consensus sequences of lineages from the same family were aligned using MAFFT with the "--auto" parameter, and lineages whose consensus sequences show a similarity larger than 80% with each other were combined. Truncated, solo, and fragmented LTR-RT elements were assigned to lineages based on the same 80-90-100 rule in the similarity searching of BLASTN.

Investigation of LTR-RT traits

Traits, including the length of the total LTR-RT and LTR, percentage of internal region, GC content, distance to gene and exon, insertion time, and expression level, were recorded for intact, truncated, and solo elements. For intact elements, we also recorded the presence or absence of each type of coding region in the internal region, including capsid protein (GAG), aspartic proteinase (AP), integrase (INT), reverse transcriptase (RT), and RNase H (RH), based on Open Reading Frames (ORFs) identified by TEsorter.

The insertion time of intact LTR-RT elements in *C. dehongense* and *C. currorii* was calculated by dividing the divergence (mismatched nucleotide sites) between the LTRs by the nucleotide mutation rate of the genome. To obtain the mutation rate, we identified the syntenic genomic regions between *V. vinifera* and the two representative *Cyphostemma* species using LASTZ v.1.04.22 (http://www.bx.psu.edu/miller_lab/) with the parameters "T = 2 C = 2 H = 2000 Y = 3400 L = 6000 K = 2200". Sequence divergence between species was calculated based on the polymorphic sites, except for nucleotides classified as "N" or falling in an alignment gap. We set the divergence time of *V. vinifera* and two representative *Cyphostemma* species as 71.02 Ma according to the estimation of You et al.[18]. The calculated mutation rate in the genome of *C. dehongense* was 7.22×10^{-9}, and that of *C. currorii* was 3.72×10^{-8}.

The expression level of intact LTR-RTs for *C. dehongense* and *C. currorii* was calculated based on RNA sequencing data. The RNA reads obtained from different organs were mapped to the genome using Hisat2 v.2.1.0[143], and the output sam file was reordered by samtools v.1.18[144]. The reads mapped on each LTR-RT were counted using featureCounts v.2.0.6[145] with the default parameters, which were used to calculate the transcripts per million (TPM) value using a custom Python script. LTR-RTs with a TPM value of 0 were regarded as inactive and were discarded when plotting the expression level of LTR-RTs. The significances of differences in expression level across LTR-RT traits were compared using the Wilcoxon signed-rank test.

For each lineage, the solo/intact (SI) ratio, fragmented/intact ratio, and average value of LTR-RT features were calculated. All the calculations were conducted with custom Perl scripts. The information for truncated, solo, and fragmented LTR-RT elements is deposited in figshare.

Investigation of genomic LTR-RT dynamics

For family and lineage including > 50 intact elements, we plotted age distribution and insertion rate, and estimated death rate and half-life using the R package TE v.0.3-0[146].

Previous study suggests a negative relationship between element age and GC content due to the deamination of methylated cytosines[34]. We therefore checked correlations between age and GC content in each family and lineage including > 30 intact elements in *C. dehongense* and *C. currorii*. Since most lineages showed a significant negative relationship between element age and GC content when fitting a generalized linear model with the "glm"

function of R (Supplementary Data 14), we concluded that cytosines in the LTR-RT element have been replaced by thymine at a constant rate.

Following the concept of the fossilized birth-death model, we dated the phylogeny of *Ale* lineages. regarded the LTR-RT elements with age very close to 0 as "extant elements", and other elements as "fossils". According to the ML phylogeny of LTR-RT, we selected representative elements for each major clade to perform dating analysis, and for each lineage, we selected up to five elements to avoid over-representing. Because whether the elements possess complete ORFs affects the methylation level of the element, we only selected elements with complete ORFs. In total, we used 55 "extant elements" and 27 "fossils" from *C. dehongense* or *C. currorii* for *Ale* lineage sampling, and used four *Tork* elements as outgroup (Supplementary Data 13). The AA sequences of RT were aligned using MAFFT with L-INS-i algorithm, and sites with > 50% gaps were removed. The trimmed alignment was back-translated using the most frequent codon in plants. Time tree construction was performed in BEAST v.2.4.8[147] using a tip dating, GTR + R substitution model, and relaxed clock with lognormal distribution. The analysis was run with two MCMC chains for 500,000,000 generations, sampling every 1,000 generations. We then reconstructed the ancestral state of intron preference on this time tree using the Bayesian stochastic character mapping approach implemented in the R package phytools, with two MCMC chains and 5,000 runs.

Effects of LTR-RT insertions on genes

We compared the expression levels of genes with introns of different lengths in *C. dehongense* and *C. currorii*. We grouped genes in both species according to the length of their longest intron, as substantial intron expansions can cause gene malfunction by transcription errors and can serve as a proxy for selection force. To test whether genes with elongated introns are under relaxed selection, we calculated d_N/d_S ratio of orthologs in *C. dehongense*, *C. currorii*, *T. hemsleyanum*, *C. rotundifolia*, and *V. vinifera* in Vitaceae, and one outgroup species *A. thaliana*, using PAML v.10.4.7[127]. For *Cyphostemma* species, we tested the correlations between the d_N/d_S ratio and intron length with the R function "cor.test". We also compared LTR-RT insertion length in syntenic genes of *C. dehongense* and *C. currorii* determined by MCScanX to explore the possible intron elongation patterns, particularly focusing on genes possibly involved in succulence development, as suggested by literature[6,148].

To investigate the effects of elongated introns, we conducted GO enrichment analyses for genes in *C. currorii* with extra-long introns (> 4,000 bp) elongated by LTR-RTs compared with *C. dehongense*. All annotated genes in *C. currorii* were used as the background set. Enrichment analysis was performed in the R package clusterProfiler v.4.4.4[149], where the GO term enrichment with a *p* (adjusted by the "fdr" method) < 0.05 was considered significant. We tested expression differences in succulent and non-succulent species using transcriptomes of 18 *Cyphostemma* taxa (Supplementary Data 1), where genes with significant expression differences (FDR $p \leq 0.05$ and |log2FC (log-based fold change)| ≥ 1) were identified using the R package edgeR v.3.38.4[150].

Leaf-surface anatomy in *Cyphostemma*

We selected representative succulent and non-succulent *Cyphostemma* species with available GS data to measure their leaf-surface traits using SEM, with the non-succulent *Tetrastigma triphyllum* and *Cissus verticillata* as outgroup. We used guard cell length as a proxy of stomata size to avoid the influence of stomata movement, as suggested by a previous study[42]. Images of abaxial (lower) and adaxial (upper) surfaces of the middle portion of leaves were taken by a scanning electron microscope at 200× magnification to assess cell size and stomatal density. Three individual replicates per species and five random fields per leaf side were taken. Stomatal counts, guard cell length, and stomatal density were measured using ImageJ v.1.54f [151]. Stomatal index was calculated as follows:

$$stomatal\ index\ (\%) = \frac{S}{S+E} \times 100$$

Where S and E are the numbers of stomata and epidermal cells in the microscopic view field, respectively. The mean values and standard deviations of guard cell length were also calculated. The leaf surfaces were classified into four types according to their traits (see details in Supplementary Note 2).

Environmental factor analysis

We compiled the environmental data for *Cyphostemma* species according to their geographic occurrence from Global Biodiversity Information Facility (GBIF)[152], herbarium specimens, and our field observations. Occurrence records without coordinates but with locations below the district level were manually converted into coordinates in

Google Earth (https://www.google.com/earth/), considering habitats of species. Coordinates were cleaned using the R package coordinatecleaner v.2.0-20[153] and manually checked according to the distribution information in POWO.

We downloaded 36 environmental variables from the CHELSA database v.2.1[154], glUV database[155], and Global Aridity and PET Database v3[156], all at a 30 arc-second resolution. Based on the occurrence records, we extracted the environmental factors for *Cyphostemma* individuals and calculated the mean values for each species. Variables with little variation among *Cyphostemma* species were excluded. We also downloaded global total nitrogen content data of different soil depths (0–5 cm, 5–15 cm, 15–30 cm) from the SoilGrids250m 2.0 database[157] and global available phosphorus data from McDowell et al.[158]. Elevation data were also extracted[159] as a potential local factor influencing GS changes.

We tested correlations between GS and environmental factors using the PGLS approach, based on the MSC species tree of *Cyphostemma*. Due to different GS variation patterns in the continental African clade and Malagasy clade of *Cyphostemma*, the following analyses were conducted on the two clades separately. To eliminate multicollinear effects, we assessed the collinearity of environmental factors in the R package corrplot v.0.92[160] and kept only the best predictor of GS variation among the correlated variables. The best model was selected based on the AICc value from all environmental factor combinations in the PGLS analyses conducted by R package caper v.1.0.3. We also conducted the same PGLS analyses for soil nitrogen and phosphorus content and elevation.

To explore whether certain LTR-RTs promoted genome expansion, we tested relationships between GS and the proportion of LTR-RTs in the genome using the PGLS analysis. We only tested the four most abundant LTR-RT families (i.e., *Ale*, *Ogre*, *SIRE*, and *Tekay*) averagely occupying > 3% of the whole genome, which can lead to observable genome size changes. We tested all the combinations of the four LTR-RT families as the predictor of genome size in the PGLS analyses. To test whether the proliferation of LTR-RTs is related to the environment, especially aridity, we also tested the relationships between the proportion of the four most abundant LTR-RTs and the environment factors involved in the best model and global arid index using PGLS analysis, with all the covariate tested.

Statistical analysis

For RNA and bisulfite sequencing, three replicates were taken per sample. For all the statistical comparisons, the significances of differences between groups were calculated with the unpaired two-sided Wilcoxon signed-rank test in R, where a $p < 0.05$ was considered significant. Correlations were tested by Spearman's rank correlation test with the R function "cor.test". The numbers of samples involved in the statistical analysis were labelled near the corresponding graph.

Data availability

The genome and transcriptome sequencing data, and genome assemblies for *C. dehongense* and *C. currorii*, have been deposited in the National Genomics Data Center (NGDC)[161] under BioProject PRJCA058853 [https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA058853]. Accession numbers for ONT, PacBio, and Hi-C data are listed in Supplementary Data 29, for genome assemblies in Supplementary Data 6, and for transcriptome data in Supplementary Data 1. The shallow whole genome sequencing data for *Cyphostemma* species and outgroups generated in this study have been deposited in National Center for Biotechnology Information (NCBI) under the BioProject PRJNA1274375 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1274375/], and accession number for each sequence read archive (SRA) data is presented in Supplementary Data 1 and 2. The genome size records and environmental variables used for analyses in this study are provided in the Supplementary Information. The accessions of genome assemblies and annotations of super-rosid species used for comparative genome analyses are provided in Supplementary Data 6. Detailed information on LTR-RTs elements, distribution records of *Cyphostemma* species, annotations for repeat sequences and genes, methylation sequencing data, and photos of SEM leaf observation are available in the figshare database under the DOI: https://figshare.com/s/71cf4e8d081c59fe52d7 [https://figshare.com/s/71cf4e8d081c59fe52d7] [162]. The data and code to generate all the box plots, violin plots, dot plots, bar charts, line charts, and mean values in tables are provided in the figshare deposition.

Code availability

The code developed in this study to perform analyses can be accessed in figshare: https://figshare.com/s/71cf4e8d081c59fe52d7 [https://figshare.com/s/71cf4e8d081c59fe52d7] [162].

References

1. Grace, O. M. Succulent plant diversity as natural capital. *Plants People Planet* **1**, 336–345 (2019).

2. Griffiths, H. & Males, J. Succulent plants. *Curr. Biol.* **27**, R890–R896 (2017).

3. Arakaki, M. *et al.* Contemporaneous and recent radiations of the world's major succulent plant lineages. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 8379–8384 (2011).

4. Evans, M. *et al.* Insights on the evolution of plant succulence from a remarkable radiation in Madagascar (*Euphorbia*). *Syst. Biol.* **63**, 698–711 (2014).

5. Nyffeler, R. & Eggli, U. An up-to-date familial and suprafamilial classification of succulent plants. *Bradleya* **28**, 125–144 (2010).

6. Heyduk, K. The genetic control of succulent leaf development. *Curr. Opin. Plant Biol.* **59**, 101978 (2021).

7. Ogburn, R. M. & Edwards, E. J. The ecological water-use strategies of succulent plants. in Advances in Botanical Research (eds. Kader, J.-C. & Delseny, M.) vol. 55 179–225 (Academic Press, 2010).

8. Ogburn, R. M. & Edwards, E. J. Quantifying succulence: a rapid, physiologically meaningful metric of plant water storage. *Plant, Cell Environ.* **35**, 1533–1542 (2012).

9. Mozzi, G. *et al.* Divergent structural leaf trait spectra in succulent versus non-succulent plant taxa. *Ann. Bot.* **134**, 491–500 (2024).

10. Ranaivoson, R. M. *et al.* Plastid and nuclear phylogenomics of *Cyphostemma* (Vitaceae) provide new insights into genome size evolution across sub-Saharan Africa. *J. Integr. Plant Biol.* **68**, 1399–1420 (2026).

11. Pacey, E. K. *et al.* Polyploidy increases storage but decreases structural stability in *Arabidopsis thaliana*. *Curr. Biol.* **32**, 1–7 (2022).

12. Bhadra, S., Leitch, I. J. & Onstein, R. E. From genome size to trait evolution during angiosperm radiation. *Trends Genet.* **39**, 728–735 (2023).

13. Wang, D. *et al.* Which factors contribute most to genome size variation within angiosperms? *Ecol. Evol.* **11**, 2660–2668 (2021).

14. Novák, P. *et al.* Repeat-sequence turnover shifts fundamentally in species with large genomes. *Nat. Plants* **6**, 1325–1329 (2020).

15. Venner, S., Feschotte, C. & Biémont, C. Dynamics of transposable elements: towards a community ecology of the genome. *Trends Genet.* **25**, 317–323 (2009).

16. Galindo-González, L., Mhiri, C., Deyholos, M. K. & Grandbastien, M. A. LTR-retrotransposons in plants: engines of evolution. *Gene* **626**, 14–25 (2017).

17. Bennetzen, J. L. & Wang, H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.* **65**, 505–530 (2014).

18. You, Y. *et al.* Transition of survival strategies under global climate shifts in the grape family. *Nat. Plants* **10**, 1100–1111 (2024).

19. Jiang, K. & Goertzen, L. R. Spliceosomal intron size expansion in domesticated grapevine (*Vitis vinifera*). *BMC Res. Notes* **4**, 52 (2011).

20. Zhu, S. *et al.* Chromosome-level reference genome of *Tetrastigma hemsleyanum* (Vitaceae) provides insights into genomic evolution and the biosynthesis of phenylpropanoids and flavonoids. *Plant J.* **114**, 805–823 (2023).

21. Costa, J. H. *et al.* The alternative oxidase family of *Vitis vinifera* reveals an attractive model to study the importance of genomic design. *Physiol. Plant.* **137**, 553–565 (2009).

22. Prothero, D. R. The late Eocene-Oligocene extinctions. *Annu. Rev. Earth Planet. Sci.* **22**, 145–165 (1994).

23. Chu, Z. F., Wen, J., Yang, Y. P., Nie, Z. L. & Meng, Y. Genome size variation and evolution in the grape family Vitaceae. *J. Syst. Evol.* **56**, 273–282 (2018).

24. Gichuki, D. K. *et al.* Genome size, chromosome number determination, and analysis of the repetitive elements in *Cissus quadrangularis*. *PeerJ* **7**, e8201 (2019).

25. 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).

26. Hirsch, C. D. & Springer, N. M. Transposable element influences on gene expression in plants. *Biochim. Biophys. Acta* **1860**, 157–165 (2017).

27. He, B. *et al.* Evolution of plant genome size and composition. *Genomics, Proteomics Bioinforma.* **22**, qzae078 (2024).

28. Gent, J. I. *et al.* CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Res.* **23**, 628–637 (2013).

29. Schley, R. J. *et al.* The ecology of palm genomes: repeat-associated genome size expansion is constrained by aridity. *New Phytol.* **236**, 433–446 (2022).

30. Neumann, P. *et al.* Impact of parasitic lifestyle and different types of centromere organization on chromosome and genome evolution in the plant genus *Cuscuta*. *New Phytol.* **229**, 2365–2377 (2021).

31. Stritt, C., Thieme, M. & Roulin, A. C. Rare transposable elements challenge the prevailing view of transposition dynamics in plants. *Am. J. Bot.* **108**, 1310–1314 (2021).

32. Lisch, D. How important are transposons for plant evolution? *Nat. Rev. Genet.* **14**, 49–61 (2013).

33. Bennetzen, J. L. Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr. Opin. Genet. Dev.* **15**, 621–627 (2005).

34. Stritt, C., Wyler, M., Gimmi, E. L., Pippel, M. & Roulin, A. C. Diversity, dynamics and effects of long terminal repeat retrotransposons in the model grass *Brachypodium distachyon*. *New Phytol.* **227**, 1736–1748 (2020).

35. Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V. & Kondrashov, F. A. Selection for short introns in highly expressed genes. *Nat. Genet.* **31**, 415–418 (2002).

36. Lynch, M. Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 6118–6123 (2002).

37. Vu, G. T. H., Cao, H. X., Reiss, B. & Schubert, I. Deletion-bias in DNA double-strand break repair differentially contributes to plant genome shrinkage. *New Phytol.* **214**, 1712–1721 (2017).

38. Jones, A. R. *et al.* Cell-size dependent progression of the cell cycle creates homeostasis and flexibility of plant cell size. *Nat. Commun.* **8**, 15060 (2017).

39. Komaki, S. & Sugimoto, K. Control of the plant cell cycle by developmental and environmental cues. *Plant Cell Physiol.* **53**, 953–964 (2012).

40. Perrot-Rechenmann, C. Cellular responses to auxin: division versus expansion. *Cold Spring Harb. Perspect. Biol.* **2**, a001446 (2010).

41. Jovtchev, G., Schubert, V., Meister, A., Barow, M. & Schubert, I. Nuclear DNA content and nuclear and cell volume are positively correlated in angiosperms. *Cytogenet. Genome Res.* **114**, 77–82 (2006).

42. Beaulieu, J. M., Leitch, I. J., Patel, S., Pendharkar, A. & Knight, C. A. Genome size is a strong predictor of cell size and stomatal density in angiosperms. *New Phytol.* **179**, 975–986 (2008).

43. Simonin, K. A. & Roddy, A. B. Genome downsizing, physiological novelty, and the global dominance of flowering plants. *PLoS Biol.* **16**, e2003706 (2018).

44. Nakayama, H. Leaf form diversity and evolution: a never-ending story in plant biology. *J. Plant Res.* **137**, 547–560 (2024).

45. Jordan, G. J., Carpenter, R. J., Koutoulis, A., Price, A. & Brodribb, T. J. Environmental adaptation in stomatal size independent of the effects of genome size. *New Phytol.* **205**, 608–617 (2015).

46. Trávníček, P. *et al.* Diversity in genome size and GC content shows adaptive potential in orchids and is closely linked to partial endoreplication, plant life-history traits and climatic conditions. *New Phytol.* **224**, 1642–1656 (2019).

47. Liu, Y., Wang, Y., Willett, S. D., Zimmermann, N. E. & Pellissier, L. Escarpment evolution drives the diversification of the Madagascar flora. *Science.* **383**, 653–658 (2024).

48. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic Acids Res.* **39**, 2010–2012 (2011).

49. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

50. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

51. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).

52. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).

53. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

54. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

55. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

56. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

57. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).

58. Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M. & Kelly, S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.* **26**, 1134–1144 (2016).

59. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).

60. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

61. Pellicer, J. & Leitch, I. J. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytol.* **226**, 301–305 (2020).

62. Galbraith, D. W. *et al.* Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science.* **220**, 301–305 (1983).

63. Sakai, H., Tanaka, T., Antonio, B. A., Itoh, T. & Sasaki, T. The first monocot genome sequence: *Oryza sativa* (rice). *Adv. Bot. Res.* **69**, 119–135 (2014).

64. Schmutz, J. *et al.* A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* **46**, 707–713 (2014).

65. Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4069–4074 (2011).

66. Qin, C. *et al.* Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 5135–5140 (2014).

67. Hu, J. *et al.* NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol.* **25**, 107 (2024).

68. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).

69. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).

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

71. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).

72. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* **25**, 4.10.1-4.10.14 (2009).

73. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 4–9 (2015).

74. Xu, Z. & Wang, H. LTR-FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, w265–w268 (2007).

75. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).

76. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).

77. Zhang, R.-G. *et al.* TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic. Res.* **9**, uhac017 (2022).

78. Hoff, K. J. & Stanke, M. Predicting genes in single genomes with AUGUSTUS. *Curr. Protoc. Bioinforma.* **65**, e57 (2019).

79. Blanco, E., Parra, G. & Guigó, R. Using geneid to identify genes. *Curr. Protoc. Bioinforma.* **64**, e56 (2007).

80. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).

81. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).

82. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).

83. Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10**, 547–548 (2000).

84. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, 1178–1186 (2012).

85. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).

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

87. Xu, Y., Wang, X., Yang, J., Vaynberg, J. & Qin, J. PASA – a program for automated protein NMR backbone signal assignment by pattern-filtering approach. *J. Biomol. NMR* **34**, 41–56 (2006).

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

89. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

90. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337 (2009).

91. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).

92. Schultz, M. D., Schmitz, R. J. & Ecker, J. R. "Leveling" the playing field for analyses of single-base resolution DNA methylomes. *Trends Genet.* **28**, 583–585 (2012).

93. Wen, J. *et al.* Transcriptome sequences resolve deep relationships of the grape family. *PLoS One* **8**, e74394 (2013).

94. Rabarijaona, R. N. *et al.* Species delimitation and biogeography of *Cyphostemma* (Vitaceae), emphasizing diversification and ecological adaptation in Madagascar. *Taxon* **72**, 766–790 (2023).

95. Yu, J. *et al.* Distinct hybridization modes in wide- and narrow- ranged lineages of *Causonis* (Vitaceae). *BMC Biol.* **21**, 209 (2023).

96. Xin, H. *et al.* A genome for *Cissus* illustrates features underlying the evolutionary success in dry savannas. *Hortic. Res.* **9**, uhac208 (2022).

97. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).

98. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

99. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

100. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

101. 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).

102. Struck, T. H. TreSpEx—detection of misleading signal in phylogenetic reconstructions based on tree information. *Evol. Bioinforma.* **10**, 51–67 (2014).

103. Ma, Z. Y. *et al.* Phylogenomic relationships and character evolution of the grape family (Vitaceae). *Mol. Phylogenet. Evol.* **154**, 106948 (2021).

104. 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).

105. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).

106. Johnson, M. G. *et al.* HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Appl. Plant Sci.* **4**, 1600016 (2016).

107. Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541–i548 (2014).

108. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**, 15–30 (2018).

109. Chernomor, O., Von Haeseler, A. & Minh, B. Q. Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol.* **65**, 997–1008 (2016).

110. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

111. Minh, B. Q., Hahn, M. W. & Lanfear, R. New methods to calculate concordance factors for phylogenomic datasets. *Mol. Biol. Evol.* **37**, 2727–2733 (2020).

112. Salichos, L., Stamatakis, A. & Rokas, A. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol. Biol. Evol.* **31**, 1261–1271 (2014).

113. Pease, J. B., Brown, J. W., Walker, J. F., Hinchliff, C. E. & Smith, S. A. Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. *Am. J. Bot.* **105**, 385–403 (2018).

114. Smith, S. A., Moore, M. J., Brown, J. W. & Yang, Y. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* **15**, 150 (2015).

115. Pease, J. B., Haak, D. C., Hahn, M. W. & Moyle, L. C. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biol.* **14**, e1002379 (2016).

116. Cai, L. *et al.* The perfect storm: gene tree estimation error, incomplete lineage sorting, and ancient gene flow explain the most recalcitrant ancient angiosperm clade, Malpighiales. *Syst. Biol.* **70**, 491–507 (2021).

117. Liu, L. & Yu, L. Phybase: an R package for species tree analysis. *Bioinformatics* **26**, 962–963 (2010).

118. Rambaut, A. & Grassly, N. C. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics* **13**, 235–238 (1997).

119. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

120. Grömping, U. Relative importance for linear regression in R: the package relaimpo. *J. Stat. Softw.* **17**, 1–27 (2006).

121. Solís-Lemus, C., Bastide, P. & Ané, C. PhyloNetworks: a package for phylogenetic networks. *Mol. Biol. Evol.* **34**, 3292–3298 (2017).

122. Chen, I. & Manchester, S. R. Seed morphology of modern and fossil *Ampelocissus* (Vitaceae) and implications for phytogeography. *Am. J. Bot.* **94**, 1534–1553 (2007).

123. Tiffney, B. H. & Barghoorn, E. S. Fruits and seeds of the Brandon Lignite. I. Vitaceae. *Rev. Palaeobot. Palynol.* **22**, 169–191 (1976).

124. Manchester, S. R. Fruits and seeds of the middle Eocene nut beds flora, Clarno Formation, Oregon. *Paleoutographica Am.* **58**, 1–205 (1994).

125. Magallón, S. & Castillo, A. Angiosperm diversification through time. *Am. J. Bot.* **96**, 349–365 (2009).

126. Smith, S., Brown, J. & Walker, J. So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. *PLoS One* **13**, e0197433 (2017).

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

128. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).

129. Matzke, N. J. BioGeoBEARS: BioGeography with Bayesian (and likelihood) Evolutionary Analysis with R Scripts (Version 1.1.1). at https://github.com/nmatzke/BioGeoBEARS.

130. Yu, Y., Harris, A. J., Blair, C. & He, X. RASP (Reconstruct Ancestral State in Phylogenies): a tool for historical biogeography. *Mol. Phylogenet. Evol.* **87**, 46–49 (2015).

131. Ree, R. H. & Smith, S. A. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Syst. Biol.* **57**, 4–14 (2008).

132. Ronquist, F. Dispersal-vicariance analysis: a new approach to the quantification of historical biogeography. *Syst. Biol.* **46**, 195–203 (1997).

133. Landis, M. J., Matzke, N. J., Moore, B. R. & Huelsenbeck, J. P. Bayesian analysis of biogeography when the number of areas is large. *Syst. Biol.* **62**, 789–804 (2013).

134. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).

135. Beaulieu, J. M., Jhwueng, D. C., Boettiger, C. & O'Meara, B. C. Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution.* **66**, 2369–2383 (2012).

136. Hansen, T. F. Stabilizing selection and the comparative analysis of adaptation. *Evolution.* **51**, 1341–1351 (1997).

137. Uyeda, J. C. & Harmon, L. J. A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. *Syst. Biol.* **63**, 902–918 (2014).

138. Novák, P., Neumann, P. & Macas, J. Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nat. Protoc.* **15**, 3745–3776 (2010).

139. Li, H. Toolkit for processing sequences in FASTA/Q formats. at https://github.com/lh3/seqtk (2012).

140. Novák, P. *et al.* TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res.* **45**, e111 (2017).

141. Seberg, O. & Petersen, G. Corresponse: a unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nat. Rev. Genet.* **10**, 276 (2009).

142. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).

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

144. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

145. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

146. Dai, X. *et al.* Birth and death of LTR-retrotransposons in *Aegilops tauschii*. *Genetics* **210**, 1039–1051 (2018).

147. Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).

148. Males, J. Secrets of succulence. *J. Exp. Bot.* **68**, 2121–2134 (2017).

149. Wu, T. *et al.* clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).

150. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).

151. Abràmoff, M. D., Magalhães, P. J. & Ram, S. J. Image processing with imageJ. *Biophotonics Int.* **11**, 36–41 (2004).

152. GBIF.org. GBIF Occurrence Download. https://doi.org/10.15468/dl.4rmpdj (2024).

153. Zizka, A. *et al.* CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. *Methods Ecol. Evol.* **10**, 744–751 (2019).

154. Karger, D. N. *et al.* Climatologies at high resolution for the earth's land surface areas. *Sci. Data* **4**, 170122 (2017).

155. Beckmann, M. *et al.* glUV: a global UV-B radiation data set for macroecological studies. *Methods Ecol. Evol.* **5**, 372–383 (2014).

156. Zomer, R. J., Xu, J. & Trabucco, A. Version 3 of the global aridity index and potential evapotranspiration database. *Sci. Data* **9**, 409 (2022).

157. Poggio, L. *et al.* SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *Soil* **7**, 217–240 (2021).

158. Mcdowell, R. W., Noble, A., Pletnyakov, P. & Haygarth, P. M. A global database of soil plant available phosphorus. *Sci. Data* **10**, 125 (2023).

159. Information., N. N. C. for E. ETOPO 2022 15 Arc-Second Global Relief Model. *NOAA National Centers for Environmental Information.* (2022) doi:https://doi.org/10.25921/fd45-gt74.

160. Wei, T. & Simko, V. R package 'corrplot': visualization of a correlation matrix. at https://github.com/taiyun/corrplot (2021).

161. The GSA Family in 2025: a broadened sharing platform for multi-Omics and multimodal data. *Genom. Proteom. Bioinform.* **23**(4), qzaf072 (2025).

162. Yu, J. *et al*. Genomic and ecological drivers of parallel arid adaptation in tree grapes (Vitaceae) data sets. figshare (2026) https://figshare.com/s/71cf4e8d081c59fe52d7.

Acknowledgments

We thank Bing Liu, Xiaolei Lin, Jianfei Ye, Zhangjian Shan, Chengxin Fu, Pan Li, Anna Trias-Blasi, Jun Wen, Langxing Yuan, and Viet-Cuong Dang for sample collection and/or field assistance, Daming Zhang for support in laboratory work, Yichen You and Yujie Zhao for cultivating plant materials in our greenhouse, Jian Zhang for suggestions on methylation analyses, Xiaoxue Li for helping with reviewing author checklist items, Liming Cai, Dario Cantu, Zhenchang Liang, Yingxiong Qiu, Haiping Xin, Yi Wang, and Shanshan Zhu for sharing genome assembly data, Patemoshela Kashikola for sharing the image of *C. currorii*, and staff from HZU, K, P, PE, US, TCD for the loan of or access to specimens.

Funding

This work was supported by National Natural Science Foundation of China 32221001 (to Z.C.) and 32270230 (to L.L.); National Key Research Development Program of China 2022YFC2601200 (to L.L.), 2023YFF0805800 (to Z.C.); International Partnership Program of the Chinese Academy of Sciences 063GJHZ2024053FN (to L.L.), 151853KYSB20190027 (to Z.C.); Sino-Africa Joint Research Center and CAS International Research and Education Development Program SAJC202527ZD01 (to Z.C.); Swedish Research Council 2024-04303 (to A.A.); Swedish Foundation for Strategic Environmental Research MISTRA Project BioPath (to A.A.); Kew Foundation (to A.A.); and CAS President's International Fellowship Initiative (to R.L.B and A.A.).

Author Contributions Statement

Conceptualization: L.L.; Data curation: J.Y., J.Z., S.L., R.M.R.; Formal analysis: J.Y., J.Z., S.L.; Funding acquisition: Z.C., L.L.; Investigation: J.Y., Z.J., S.L., K.N., R.M.R., R.L.B., R.N.R., Z.C., L.L.; Methodology: J.Y., L.L.; Resources: Z.C., L.L.; Supervision: I.L., A.A., Z.C., L.L.; Validation: J.Y., J.Z.,

R.M.R, C.D; Visualization: J.Y.; Writing – original draft: J.Y., L.L.; Writing –review & editing: J.Y., J.Z., S.L., I.L., C.D., R.L.B., J.C., C.L, Y.D., A.A., Z.C., L.L. All authors have read and approved the final manuscript.

Competing Interests Statement

The authors declare no competing interests.

Figure Legends

Fig. 1 | Genome size variation and intron expansion in Vitaceae. a, Predicted ancestral genome sizes (1C-values) in Vitaceae with Ornstein-Uhlenbeck modelling. Branch colour indicates reconstructed ancestral genome size. Solid red circles show regime shifts of genome size evolution (larger circles represent higher confidence). Lineages with significant genome expansion are shown in red on the right. Blue and red arrows indicate the phylogenetic positions of the two representative species, *C. dehongense* and *C. currorii*, respectively. The images of the two species are shown on the bottom left. Photo credit: *C. dehongense* − Limin Lu; *C. currorii* − Patemoshela Kashikola (iNaturalist observation). **b**–**c**, Disproportionate intron expansion in seven Vitaceae species (purple), compared with 55 super-rosid outgroup species (grey). Blue and red dots represent *C. dehongense* and *C. currorii*, respectively. **b**, Boxplots showing differences in the average length of introns, Intron length/Exon length ratios, and the total length of LTR-RTs in introns between species in Vitaceae (n = 7) and super-rosid outgroup species (n = 55). 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 signed-rank tests were employed. **c**, Correlation between total intron length and genome size for Vitaceae (purple) and super-rosid outgroup (grey) species. Linear regression trend lines are shown separately for Vitaceae and super-rosid species. **d**, Distributions of intron length in Vitaceae and super-rosid species, with three peaks delineated by dashed lines at intervening troughs.

Fig. 2 | Variation of genome size and repeat compositions in *Cyphostemma*. a, Changes in repeat compositions and selective regimes on genome size in selected species of *Cyphostemma*, shown on the pruned coalescent tree of this study. Branch colour shows the optimal genome size (θ; 1C-value) toward which each lineage evolves under the selective regime acting on that branch, as inferred using Ornstein-Uhlenbeck modelling. Solid red circles indicate regime shifts, and solid purple circles indicate the occurrence of succulence (larger circles represent higher confidence). Succulent species are highlighted

in bold. Columns on the right of the phylogeny represent different repeat clusters, with different colours showing the categories of clusters. From top to bottom, the proportion of repeats in the genome decreases. Black bars below indicate read counts in each cluster. **b**, Mean proportions of repeat components across all sampled species within each clade. Colour coding follows Fig. 2a. Repeat types occupying < 3% of total repeats are shown in light grey. **c**, Proportions of the four most abundant LTR-RT families in *Cyphostemma*, estimated from shallow whole-genome sequencing data, showing differences in LTR-RT components between the continental African clade and Malagasy clade. 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 signed-rank tests were employed.

Fig. 3 | LTR-RT dynamics in the genome of *C. dehongense* and *C. currorii*. **a**, Age distribution and **b**, intact element counts for the four most abundant LTR-RT families in *Cyphostemma* estimated from shallow whole genome sequencing data. The right panel (**b**) shows the proportions of LTR-RT elements with complete, incomplete, and no Open Reading Frames (ORFs). **c**–**e**, Features of the four most abundant LTR-RT families in *Cyphostemma*. **c**, Expression levels. TPM, transcripts per million. **d**, Solo/intact ratio, showing deletion via recombination between the LTRs. **e**, Fragmented/intact ratio, showing deletion force through other processes. *P*-values showing the significance of the differences between *Ale* lineages preferring intron and intergenic region are shown on the right. For *Tekay* and *SIRE*, sample sizes for *C. dehongense* are too small for boxplots, and individual value plots are shown instead. 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. Each dot represents one LTR-RT lineage in *C. dehongense* (*Tekay*, n = 2; *SIRE*, n = 3; *Ogre*, n = 6; *Ale* intronic, n = 17; *Ale* intergenic, n = 46) or *C. currorii* (*Tekay*, n = 144; *SIRE*, n = 62; *Ogre*, n = 44; *Ale* intronic, n = 27; *Ale* intergenic, n = 55). Unpaired two-sided Wilcoxon signed-rank tests were employed to compare the traits of LTR-RT lineages. **f**, Time-calibrated phylogeny of *Ale* lineages across *C. dehongense* (Cyde) and *C. currorii* (Ccur), with divergence times

and pie charts showing ancestral states at key nodes. Branch colour indicates the state of the stem group inferred from the reconstructed ancestral state of its corresponding crown group. Abundant *Ale* lineages with > 50 intact elements are shown in bold. Time period of the Oligocene is highlighted in pale grey. P., Paleocene; Eo., Eocene, O., Oligocene; M., Miocene.

Fig. 4 | Effects of LTR-RT insertions on genes. a, Gene expression levels in the leaf and stem of *C. dehongense* and *C. currorii*, grouped by the length of the longest intron, with gene counts and *P*-values showing significant differences in expression levels labelled. For each species, stem and leaf samples were taken from three individual replicates. For boxplots, the centre line represents the median, the box bounds indicate the first and third quartiles, and the maximum and minimum values are indicated by the outline of violin plots. Unpaired two-sided Wilcoxon signed-rank tests were employed to compare the expression level differences between gene groups. **b**, Gene ontology enrichment of *C. currorii* genes with extra-long introns (> 4,000 bp) elongated by LTR-RT insertions compared with *C. dehongense*. Solid circles represent enriched biological processes, where the size indicates the number of related genes and the colour indicates the statistical significance. **c**–**d**, Correlations between the length of intronic LTR-RT in *C. dehongense* and the expanded length of intronic LTR-RT in *C. currorii* (expressed as the ratio of intronic LTR-RT length in *C. currorii* to *C. dehongense*) for **c**, all syntenic genes and all succulence-related syntenic genes, and **d**, 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 tested by Spearman's rank correlation test.

Fig. 5 | Environmental variables contributing to genome size variation. a, Relationships between genome size (1C-value) and environmental variables in the best regression model for the continental African clade (n = 45) and Malagasy clade (n = 27). "**" represents $p < 0.01$. Correlations between genome size and environmental factors were tested using the phylogenetic generalised least squares approach. **b**, Differences in the reported environmental variables in the habitats of succulent lineages of

both clades, compared by unpaired two-sided Wilcoxon signed-rank tests. 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.

Editorial Summary:
Succulence is a key drought-resistance trait, yet its genomic basis is underexplored. Here, authors report that long terminal retrotransposons inserted into introns may facilitate succulence evolution and genomic adaptability in tree grapes.
Peer Review Information: *Nature Communications* thanks Qing-Feng Wang and the other anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available."