Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Fabaceae pan-centromere analysis traces the evolutionary origin of centromeric satellites with widespread solo-LTR transition

Bin Guo

guobin@nwu.edu.cn

Northwest University

Rongxing Liu

Northwest University

Weikai Chen

Peking University https://orcid.org/0009-0009-8125-3993

Gang Zhou

Xinjiang Key Laboratory of Drug Quality Control/Xinjiang Institute for Drug Control

Jie Feng

Northwest University

Huan Wang

Peking University Institute of Advanced Agricultural Sciences/Shandong Laboratory of Advanced Agricultural Sciences in Weifang

Xizhen Yue

Northwest University

Ke Wang

Peking University Institute of Advanced Agricultural Sciences

Rui Li

Peking University Institute of Advanced Agricultural Sciences/Shandong Laboratory of Advanced Agricultural Sciences in Weifang

Jianing Wang

Northwest University

Tianzi Yao

Northwest University

Yongjie Qin

Northwest University

Hong Tao

Xinjiang Key Laboratory of Drug Quality Control/Xinjiang Institute for Drug Control

Yirui Zhao

Northwest University
He Wang
Northwest University
Yanping Fu
Northwest University
Wei He
Northwest University
Li Guo
Peking University https://orcid.org/0000-0001-6100-3481

Article

Keywords: Legume, Glycyrrhiza, centromere evolution, satellite DNA, retrotransposons, solo-LTR

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11542179/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

RESEARCH ARTICLE

Fabaceae pan-centromere analysis traces the evolutionary origin of centromeric satellites with widespread solo-LTR transition

Authors

Rongxing Liu[1†] , Weikai Chen[2†] , Gang Zhou[3†] , Jie Feng[1] , Huan Wang[2] , Xizhen Yue[1] , Ke Wang[2] , Rui Li[2] , Jianing Wang[1] , Tianzi Yao[1] , Yongjie Qin[1] , Hong Tao[3] , Yirui Zhao[1] , He Wang[1] , Yanping Fu[1] , Wei He[1] , Li Guo[2*] , Bin Guo[1*]

Affiliations

[1] Key Laboratory of Resource Biology and Biotechnology in Western China, Ministry of Education, Shaanxi Provincial Key Laboratory of Biotechnology, College of Life Sciences, Northwest University, Xi'an, 710069,China

[2] Peking University Institute of Advanced Agricultural Sciences, Shandong Laboratory of Advanced Agricultural Sciences in Weifang, Shandong, China

[3]Xinjiang Key Laboratory of Drug Quality Control, Xinjiang Institute for Drug Control, Urumqi, Xinjiang Uyghur Autonomous Region, 830011,China

[†] Co-first authors

[*]Corresponding authors: li.guo@pku-iaas.edu.cn; guobin@nwu.edu.cn

Abstract

Centromeres are essential for chromosome segregation, but their diversity and evolution across Fabaceae remain poorly understood. Here, we present a family-wide pan-centromere dissection across 20 legume genomes using one *Rosaceae* genome as outgroup, incorporating our complete, gap-free assemblies of two *Glycyrrhiza* genomes. We dissect 251 (215 gap-free) centromeres from 18 Fabaceae members and establish a 'solo-LTR-driven' model for centromere evolution. By developing LTRtrace, a pipeline to annotate centromeric solo-LTRs, we uncover a 'molecular fossil' designated as 'satLTR' where single solo-LTRs encapsulate multiple satellite repeats. By capturing the intermediate states of LTR-to-satellite conversion, such fossil demonstrates how solo-LTRs expand into massive centromeric satellite arrays via local tandem amplification and homogenization. Finally, we build an interactive platform (solo-LTR database) via integrating nearly complete genomic solo-LTR and LTR-RTs from 18 legume species., which enables the zero-code exploration of centromere-enriched repeats, serving as a crucial resource for designing synthetic centromeres. Together, our study reveals the Fabaceae centromere diversity, and provides mechanistic insights into the evolution of centromeres in legumes and a broad range of plant lineages.

Keywords

Legume, *Glycyrrhiza,* centromere evolution, satellite DNA, retrotransposons, solo-LTR

Introduction

Centromeres are critical genomic regions indispensable for accurate chromosome segregation, yet their sequences evolve at an exceptionally rapid pace, a phenomenon known as ‘centromere paradox’[1]. Despite the strict functional conservation of these chromosomal domains, their genetic composition is strikingly diverse. Plant centromeres are mostly composed of either satellite repeat arrays, retrotransposons, or a mixture of both. Recent telomere-to-telomere (T2T) genome assemblies have fundamentally reshaped our understanding of centromere diversity, revealing that centromere architecture is highly heterogeneous and evolutionarily dynamic. For example, T2T genome analyses of *Solanaceae* family have uncovered strikingly divergent evolutionary paths for centromeres. *Capsicum* centromeres completely lack high-copy tandem satellites, driven instead by recent invasions of Gypsy retrotransposons with minimal sequence conservation[2]. The allotetraploid *Nicotiana benthamiana* displays a complex centromeric mosaic, uniquely co-opting nuclear

mitochondrial DNA (NUMTs) and rDNA as functional kinetochore scaffolds[3]. By contrast, common tabacco *N. tabacum* exhibits frequent centromere repositioning and asymmetric subgenome contraction post-polyploidization, with CENH3 preferentially recruited by Tekay retrotransposons despite widespread satellite depletion[4]. This rapid structural fluidity extends beyond *Solanaceae* as haplotype-resolved genome assemblies in *Vitis vinifera* reveal extreme allelic divergence in centromere size and LTR (long terminal repeats) invasion patterns despite a uniform satellite framework[5].

Recent studies have shown that retrotransposon insertions, satellite amplification and even plastid DNA integration have led to a remarkable variation of centromere landscapes in plants. In particular, the interaction of retrotransposons with satellite repeats has profound effects on the evolution of plant centromeres. For instance, previous models in *Arabidopsis* have proposed an evolutionary cycle driven by the continuous invasion of centromere-targeted transposons alternating with the homogenization and expansion of satellite arrays[6]. Meanwhile, a recent study in *Rosa canina* has proposed an alternative evolutionary trajectory, highlighting the direct transition of ATHILA retrotransposons into centromeric satellite sequences[7]. Comparative analyses across *Solanaceae*[4], *Brassica*[8], and *Oryza*[9] have further confirmed that centromeric satellites are highly lineage-specific and undergo rapid turnover. Together, these single-base-resolution insights expand the current understanding of satellite DNA evolution, revealing a multifaceted evolutionary paradigm shaped by retrotransposon takeover, neocentromere emergence, and unconventional modes of DNA satellite integration.

Legume family (Fabaceae), comprising approximately 765 genera and more than 19,500 species,includes many important crops such as soybean, alfalfa, common bean, and licorice. Extensive efforts have been given to study the legume crops given their importance to food production, and the mechanisms of their ability to fix nitrogen. Over the past decade, the field of legume genomics has witnessed a transformative expansion, propelled by rapid advancements in sequencing technologies and computational assembly algorithms. Legumes, which hold immense ecological and agricultural significance due to their symbiotic nitrogen-fixing capabilities and high nutritional value, have been a primary focus of global genomic initiatives. This sustained effort has recently culminated in a landmark milestone: the establishment of reference genomes for 167 distinct

legume species, providing a robust phylogenetic framework for the clade[10]. Furthermore, the transition from fragmented drafts to complete T2T gap-free assemblies marks a new paradigm in plant genomics. Far beyond merely maintaining genome stability, centromeres serve as ideal systems for examining the dynamic and structural interplay between satellite DNA and retrotransposons, thereby providing an indispensable framework for elucidating evolutionary plasticity of complex legume centromeres in response to environmental extremes. Therefore, achieving gap-free sequences is pivotal for unraveling complex centromeric architectures.

Recent breakthroughs have successfully resolved the complete genomic architectures of agronomically vital crops, including soybean, alfalfa and common bean[11,12,13]. Crucially, these ultra-contiguous assemblies have unmasked the most complex and highly repetitive regions of the genome that were historically inaccessible such as centromeres and telomeres. By leveraging these T2T resources, recent studies have achieved resolution of centromeric landscapes, revealing highly dynamic, chromosome-specific architectures and intricate satellite repeat arrays in soybean and *Medicago* species[11,14,15]. Together, these studies have revealed a striking divergence in centromere architecture within the Fabaceae (legume) family. For example, the paleopolyploid soybean (*Glycine max*) centromeres have undergone dynamic satellite DNA expansion, loss, and repositioning. Notably, the divergence of centromeric satellites between annual and perennial *Glycine* species significantly outpaces genome-wide sequence variation, potentially contributing to the establishment of reproductive isolation. By contrast, the recently resolved telomere-to-telomere (T2T) genomes of *Medicago* revealed centromeres overwhelmingly dominated by CRM retrotransposons, with an exceptionally low abundance of satellite DNA[15]. Such extreme structural dichotomy, from highly enriched satellite arrays to retrotransposon-dominated domains, underscores a profound knowledge gap regarding the true architectural diversity of centromeric repeats across the entire legume family.

Here, we present the first family-wide pan-centromere dissection of Fabaceae by comparing 20 T2T or near-T2T legume genomes, including our newly-assembled T2T genomes of two *Glycyrrhiza* species. By identifying satellite families, quantifying array organization, and dissecting their interaction patterns with LTR retrotransposons, we revealed the dynamic mechanisms shaping centromere evolution across diverse species. Ultimately, this study not only filled a critical void in our panoramic understanding of legume centromere diversity, but also provided a novel theoretical

framework for deciphering the macroevolutionary mechanisms of plant satellite DNA.

Results

T2T Genome Assembly and Centromere Identification in Two *Glycyrrhiza* Species

Glycyrrhiza represents a highly representative medicinal and ecologically vital genus within the Fabaceae family, with complex repetitive landscapes and profound genomic plasticity. Previous chromosome-level assemblies of *Glycyrrhiza* species had gaps particularly in highly repetitive regions such as centromeres (Li et al., 2025). To elucidate the genomic structure and evolutionary dynamics of centromeres in *Glycyrrhiza* and across legumes, we first reconstructed telomere-to-telomere (T2T) reference genomes for *Glycyrrhiza* glabra and *Glycyrrhiza* inflata. By generating 23.38 Gb (~52.3×) of PacBio HiFi and 20.79 Gb (~46.5×) of Oxford Nanopore Technology (ONT) sequencing data (Table S1), we assembled a 463.54 Mb gap-free genome for *G. glabra* (contig N50: 56.94 Mb) and a 453.01 Mb assembly for *G. inflata* (contig N50: 59.59 Mb) (Table S1) with all eight centromeres were successfully identified in both species (Figure 1a; Figure S1). Both assemblies demonstrated high base-level accuracy and completeness, as further evidenced by quality values (QV) of 52.4 and 49.7 (Table S2), respectively, and BUSCO scores exceeding 99.2% (Table S3). Genome annotation of *G. glabra* and *G. inflata* predicted 30,371 and 31,039 protein-coding genes, with repetitive sequences occupying 63.15% and 62.75% of the genome, respectively (Table S4).

To robustly predict centromeres, we integrated tandem repeat profiling with 3D chromatin conformation. Using TRASH, we first predicted megabase-scale tandem satellite arrays comprising the putative centromeric cores in the *G. glabra* and *G. inflata* genomes, respectively (Figure S2&3). We then independently validated these predicted boundaries using interchromosomal Hi-C contact maps. In plant nuclei, centromeres typically underwnet spatial clustering, yielding distinct trans-interaction hotspots[18]. Notably, the extensive satellite arrays predicted by TRASH perfectly co-localized with these strong interchromosomal Hi-C interaction peaks[19]. This absolute concordance between satellite enrichment and 3D spatial clustering unambiguously defined the centromeric regions in both species. Centromeres accounted for 73.94 Mb (16.6% of the genome) in *G. glabra* and 69.12 Mb (15.6%) in *G. inflata*, and dominated by satellite repeats (Figure 1b), which

comprised 85.6% and 85.2% of their total lengths, respectively (Figure 1c&1g). Despite this satellite dominance, LTR retrotransposons interspersed among these arrays, constituting 1.5% and 3.1% of the centromeres. Although SIRE elements dominate centromeric region by total sequence length, these loci simultaneously harbored high copy numbers of recently inserted CRM elements (Figure S4-S7). While these two elements were comparably abundant in *G. glabra*, a 2.6-fold lineage-specific CRM expansion in G. inflata highlighted how recent mobilization has driven rapid divergence in centromere complexity.

To further characterize this structural divergence, we evaluated the detailed statistical metrics of the centromeres across both species. Overall, the centromeric landscape of *G. glabra* was defined by a gene-poor, repetitive heterochromatic state showing distinct structural and functional compartmentaliztion (Figure 1d). Using TRASH, we identified a 60-bp tandem repeat (CEN60) as the most abundant satellite in the *G. glabra* and *G. inflata* genome. The CEN60-defined core strictly excluded Gypsy/Copia retrotransposons which accumulated in the pericentromeric regions. This shared spatial organization aligned with compacted CEN60 heterochromatin governed by hierarchical DNA methylation (Figure 1f) featuring stable CG methylation alongside fluctuating CHG and localized CHH spikes. Such overarching structural blueprint was conserved in *G. inflata*, which mirrored the landscape with an identical satellite core (Figure 1h), spatial LTR exclusion (Figure 1i), and a comparable CG-dominated methylation network (Figure 1j). Furthermore, mapping of CEN60 variant frequencies across these genomic regions revealed a non-random distribution within the consensus repeats, peaking at positions 5 bp and 40 bp (Figure 1i&k).

Besides these global architectural signatures, individual chromosomes displayed distinct structural variations and local sequence dynamics between the two species. For instance, *G. glabra* Chr1 had a segmented architecture with significant boundaries, reflecting lineage-specific reorganization. This specific landscape was dominated by clustered Gly7 and Gly8 type monomers forming the structural backbone and driving sequence homogenization (Figure 1b). Additionally, a pairwise edit-distance matrix of consensus high-order repeat (HOR) units of about 12 Mb with 90 CEN60 copies across a localized region of *G. glabra* Chr1 centromere revealed a periodic block-diagonal structure, indicating HOR organization as homogenized subfamilies (Figure 1e). Similarly, Chr3 (Figure 1c) showed a centromeric interval distinctively dominated by satellite arrays

with depleted exon density and fluctuating seqence entropy. Despite strict functional conservation, these satellite-dominated centromeres underwent rapid structural turnover driven by localized sequence amplification. The abundant CEN60 monomers revealed nine distinct evolutionary clades (Gly1–Gly9) (Figure 2a&2b). Notably, while core consensus sequences were highly conserved with nucleotide polymorphisms limited mainly to positions 5 and 40, these CEN60 subfamilies displayed biased chromosomal distributions across the genome. Satellite composition showed partial variation among chromosomes, with some subfamilies within shared satellite types exhibiting chromosome-biased patterns. For example, the satellite arrays of Chr1 featured a mosaic structure of alternating Gly2 and Gly8 subfamilies, which formed a continuous and interwoven architecture (Figure 1b). Together, these patterns reveal chromosome-specific and independent evolutionary trajectories of centromeres in *Glycyrrhiza spp*.

Satellite-dominant centromere landscape in *Glycyrrhiza*

To further elucidate how these chromosome-specific patterns arose, we investigated the evolutionary dynamics of nine CEN60 subfamilies (Gly1-Gly9) in *G. glabra* and *G. inflata*. Monomer edit distance profiling of CEN60 subfamily revealed asynchronous chromosomal evolution (Figure S8). While *G. inflata* harbored substantial homogenized Gly1 arrays (Chr2/3/6), *G. glabra* exhibited localized divergence juxtaposed with recent expansion. This disparity was epitomized by Gly9 being highly mutated in *G. inflata* yet amplified as homogeneous repeats in *G. glabra*. These distinct profiles underscored independent trajectories of homogenization, mutation, and satellite turnover, demonstrating that a shared repeat library underwent rapid, independent remodeling driven by lineage-specific expansion-contraction dynamics.

Nevertheless, the organization of these repeats followed a unified pattern. For both monomeric and diad homogeneous regions (moHR and diHR), block size distributions displayed a consistent length-dependent decay across species, dominated by short repeat units with a progressive depletion of longer contiguous blocks (Figure S9). We resolved its HOR hierarchy and homogenization dynamics via constructing a structural nesting model (Figure 1e). Unlike the typical continuous ultra-long rice monomers9, *Glycyrrhiza* centromeres featured complex mixture of monomeric and dimeric units shaped by strict localized homogenization (Figure 2a). These monomers comprised 11.3-56.3% (*G. glabra*) and 10.7-52.6% (*G. inflata*) of total chromosomal monomeric content,

peaking on *G. glabra* Chr8 (56.4%) and *G. inflata* Chr7 (52.6%) (Figure S 10, Table S7). Monomer composition was species-specific: *G. inflata* was dominated by Gly9 (20.7%), Gly8 (17.9%), and Gly5 (13.1%), whereas *G. glabra* was enriched in Gly8 (27.5%) and Gly5 (19.2%) (Table S5). Genome-wide diHR profiling showed that non-random landscapes were dominated by elite combinations (Figure S11). In *G. glabra*, Gly7-Gly8 was the most frequent unit (15.75%), while Gly1-Gly5 maximized the sequence footprint (28.91%). Conversely, *G. inflata* reconfigured this architecture where Gly6-Gly5/Gly5-Gly6 comprised 39.8% of diHR regions, yet Gly7-Gly8 monopolized nearly half the total sequence footprint (Table S6). Strikingly, these elite arrays exhibited massive evolutionary shuffling across chromosomes. The Gly7-Gly8 and Gly6-Gly5 cohorts dominating *G. glabra* Chr1 and Chr8, respectively, were completely reallocated in *G.inflata*, where Gly6-Gly5 expansions shifted to Chr6 and Gly7-Gly8 monopolized Chr1. Ultimately, this dynamic interplay between highly conserved core dimers and hypervariable higher-order arrangements demonstrated how preferential tandem organization and localized expansions drove rapid centromeric evolution in *G. lycyrrhiza* (Figure S10-11).

Interestingly, this structural complexity of centromere was closely linked to its unique epigenetic characteristics. Using ONT sequencing reads, we analyzed DNA methylation in these well-defined centromeric regions. As expected, the centromeres exhibited high methylation levels in CG, CHG and CHH sequences for both species (Figure 1f,j & Figure S12-13). However, specific centromeric components displayed complex epigenetic partitioning between satellite arrays and invaded transposons. Notably, methylation levels of CRM elements were significantly lower than those in the surrounding CEN60 satellite arrays, whereas SIRE elements exhibited a hypermethylated state relative to these satellites (Figure S13).

A Framework for Centromere Type and Evolution in Fabaceae

The completely sequenced T2T genomes of two *Glycyrrhiza* species yielded high-resolution maps characterizing the complex satellite arrays that defined their centromere structures at the microevolutionary level. We expanded this analysis across 20 legume species using publicly available high-quality genomic data (Table S8) which exhibited variable genome sizes (0.29 Gb to 5.97 Gb) and chromosome numbers (n= 6 to 20). Notably, majority of these assemblies are gapless in centromeres with only four containing a few gaps. Using a smililar predictive methodology, a total of

251 candidate satellite arrays were successfully identified within the predicted centromere boundaries across these 21 genomes (Table S9). We investigated whether the hyper-variability and rapid turnover and satellite-driven evolutionary dynamics observed in *Glycyrrhiza* centromeres were shared across the Fabaceae family. Thus, we constructed a comprehensive comparative framework encompassing 251 fully resolved centromeres from 21 representative accessions (spanning 17 species, including G. max and M. sativa) to trace the macro-evolutionary trajectory of legume satellite DNA (Figure 3a). Global characterization of these centromeric repeat landscapes across the phylogeny revealed a profound structural diversity, partitioning legume centromeres into three distinct structural paradigms: satellite-dominant, LTR-dominant, and mixed architectures. Notably, the two *Glycyrrhiza* species uniquely represented the extreme satellite-dominant archetype, whereas the prevailing majority of other legumes possess LTR-dominant or mixed-type centromeres (Figure 3b; Table S10-11). Intriguingly, species assigned to mixed-type paradigm still exhibited prominent localized satellite enrichment (Figure 3c,d), highlighting the ubiquitous yet heterogeneous nature of satellite recruitment in legume centromeres (Table S10).

Evolutionary Diversification of Centromeric Satellites across Fabaceae

Despite compositional divergence, strong enrichment of satellite DNA remained a conserved centromeric paradigm across Fabaceae species. Chromosomal profiling indicated this expansion was a derived Fabaceae feature, arising independently in distantly related lineages (Figure S14). Thus, rather than strict vertical descent, shared selective pressures for chromosome segregation drove independent centromere evolution. Accordingly, comparative phylogenetics of *Glycine* and *Medicago* revealed profound intergeneric heterogeneity in satellite trajectories, highlighting highly-dynamic, lineage-specific centromere evolution (Figure 3e,f). Phylogenetically, *Glycine* centromeric satellites lacked species-specific clades, instead displaying a admixed reticulate topology (Figure 3e), suggesting the interplay of complex evolutionary forces. Interestingly, satellite DNA in *Medicago* exhibited perfect species-specific phylogenetic clustering, revealing rapid and efficient concerted evolution. Driven by active homogenization, intra-specific variations were swiftly purged, while inter-species divergence fostered the fixation of lineage-specific mutations. Independent amplification of distinct variants subsequently shaped the highly diversified satellite DNA architectures characterizing extant species. Furthermore, motif analysis of CEN60 satellites in

Glycyrrhiza revealed sustained sequence conservation throughout evolution (Figure 4c), indicating that essential structural and functional requirements severely restrict sequence divergence. Consequently, the turnover rate of this satellite family remained low, maintaining a stable evolutionary state.

Building upon these structural insights, mapping the TRASH-defined conserved blocks onto a two-dimensional distance-divergence landscape revealed that the evolution of centromeric satellite sequences was governed by a dynamic tension between centripetal homogenization and centrifugal expansion. Although the pervasive clustering of these conserved blocks established local homogenization as the default evolutionary mode, we observed striking spatial heterogeneity across the 21 genomes analyzed (Figure S15). At one extreme, *Cercis chinensis* exhibited a stringently constrained, 'hyper-compact' configuration, wherein satellite amplification was strictly confined to the immediate centromeric core. In contrast, the G. max (Gmax_WM82) accession displayed an aggressively expanded and sequence-dense evolutionary landscape, where complex evolutionary interplay enabled satellites to maintain high sequence fidelity manifested as low divergence over genomic distances. Collectively, these block-distance profiles delineated an evolutionary continuum, illuminating how centromeric arrays were sculpted by the antagonistic forces of core spatial constraint and lineage-specific genomic proliferation.

Furthermore, we analyzed pairwise similarities between the two most abundant satellite sequences in each centromere of these legume species. Pairwise satellite comparisons revealed a cross-species nested architecture where longer arrays enveloped shorter ancestral cores (Figure S16). This shared ancestry was evidenced by 46.7% homology between *Glycyrrhiza* CEN60 and *P. vulgaris* CEN540, alongside pan-legume conserved motifs perfectly aligning with the *Glycyrrhiza* Gly-Box (Figure S16). Ultimately, despite broad structural macro-diversity, *Glycyrrhiza* centromeres were not isolated anomalies but represented the extreme expansion of an ancient, core motif dynamically preserved throughout Fabaceae evolution (Figure S17).

Centromeric Co-localization of solo-LTRs and Satellite DNA

Based on multiple sequence alignments of consensus satellite sequences across diverse legume species, we identified conserved sequence motifs shared among these taxa, suggesting that these

elements may derive from the common ancestral unit within centromeric satellite DNA. This observation prompted us to investigate the evolutionary origin of these conserved elements. While the widespread presence of shared motifs across diverse legume species suggested a possible common ancestral origin, the mechanisms driving their genomic propagation and structural expansion remained unclear. To address this, we systematically explored potential sequence sources capable of generating such conserved repeat modules, including diverse classes of transposable elements and repetitive sequence derivatives. Among candidate elements, solo-LTR emerged as the principal genomic feature associated with these motif-bearing arrays, exhibiting sequence and spatial distributions that closely parallelled centromeric satellite organization. The convergence of these features raised the possibility that solo-LTR may serve as the ancestral templates for the birth and expansion of certain legume satellite DNA.

To identify and compare solo-LTR landscapes across legumes, we developed LTRtrace, an automated pipeline for intact LTR discovery, library construction, lineage classification and genome-wide solo-LTR annotation. Specifically, LTRtrace automates the extraction of the intact LTR retrotransposons to construct a comprehensive reference library, executes precise linage classification (Materials and Methods). We applied LTRtrace to 17 legume species to profile solo-LTR abundance, composition and spatial distribution (Figure S18). This donor-derivative relationship was further validated by assessing the co-localization of solo-LTR with centromeric satellite arrays. To investigate the spatial relationship among solo-LTRs, complete LTRs and satellite repeats, we analyzed the co-localized patterns of these soloLTRs sequence types within the centromere (CEN) and peri-centromere (PeriCEN) regions in 17 accessions (Figure S19). The results indicated a clear co-localization preference between solo-LTR and satellite DNA sequences within the centromere region (Figure S20, S21). Specifically, within the centromere region, a significant proportion of solo-LTR co-localized with satellite DNA (Table S13). For example, in *G. inflata*, approximately 47.0% (272/581) of solo-LTR co-localized with satellite sequences. In contrast, the frequency of such co-localized was markedly lower in the pericentromere regions of almost all species. Notable examples include *C. chinensis* (Cchi), *Delonix regia* (Dreg) and *P. vulgari* (Pvul), where, despite the presence of solo-LTR in the core centromere, no co-localization of solo-LTR and satellite DNA was observed in the pericentromeric region (Table S13). Interestingly, there were significant differences in the number of co-localization between solo-LTR and satellite DNA across

different species, with *C. arietinum* (Cari) having the highest number (1,758 in the CEN region and 26 in the PeriCEN region), whilst species such as *M. truncatula* (Mtru_A17) showed very few.

Genome-wide distribution and centromeric truncation of solo-LTR

We further compared solo-LTR distribution within the centromeric and chromosome arms, revealing strong spatial heterogeneity (Figure S22&23), with arm densities of 25.96-73.54 counts/Mb (Table S12) being markedly higher than in the centromeric region in most species. Notably, *V. radiata* reaches 73.54 counts/Mb on arms versus only 20.44 counts/Mb in its centromere. while *M. truncatula* displayed the most extreme disparity, with chromosome-arms density reaching 47.13 counts/Mb against just 1.72 counts/Mb in the centromere (>27-fold difference). G. max (Gmax_Jack) dsiplayed an unusually high overall load (25.33 copies/Mb in its centromere), consistent with its species-specific transposon history (Table S12). Genomic coverage (percentage of length) occupied by solo-LTR sequences indicated their enrichment in chromosome arms. For instance, solo-LTR coverage of G. max (Gmax_ZH13) in the arms (4.77%) was significantly higher than the centromere (2.80%), repesenting an approximately two-fold difference across the genome (Table S12). This disparity was even more pronounced for solo-LTR counts per Mb, with 37.90 in the arms versus 25.13 in the centromeres of Gmax_ZH13, and a difference of up to 34.72 in *M.truncatula* (Mtru A17). Notably, solo-LTR in centromeric regions displayed a characteristic truncation. Consistent with eariler observations, solo-LTR density in Mtru A17 was significantly higher in arms than in the centromeric region, whereas approximately 50% of all solo-LTR were co-localized with satellites which likely contributed to this marked disparity. Given the statistically significant association between solo-LTR and satellites (Table S13-14), we hypothesized that the satellites are likely derived from these solo-LTRs.

Sequence conservation and compositional continuity between solo-LTR and satellite DNA

Through in-depth analysis of the CEN60 satellite family in the centromere of *Glycyrrhiza*, we identified a highly conserved core motif, designated as the ‘Gly-Box’. Systematic sequence alignment of centromeric satellite consensus sequences from representative legume species against the Gly-Box (Table S16) revealed strong sequence homology across Fabaceae. Despite the extensive evolutionary divergence among these species, a substantial number of conserved nucleotide sites

persist within their centromeric satellites. Collinearity analysis uncovered distinct segmented collinear blocks between the Gly-Box and the satellite consensus sequences of each species. We observed an enrichment of solo-LTR in peri-/centromeric regions of these legumes, with an average sequence similarity to host satellite monomers significantly exceeding 40% (Figure 5b).

Multidimensional evidence elucidated the evolutionary trajectory of legume centromeric satellite originating from transposon remnants. Departing from the conventional paradigm that treats satellites and TEs as discrete entities, we applied a k-mer frequency-based feature extraction (k=3) to perform unsupervised clustering of centromeric solo-LTRs and mature satellites across legumes(Figure 4a, Figure S24). Principal component analysis (PCA) revealed that the first two components explain 9.1% to 33% of sequence variatio. Centromeric and pericentromeric solo-LTR did not segregate from satellite sequences but instead overlapped significantly, delineating a continuous evolutionary trajectory (Figure 4b, Figure S24). Hierarchical clustering revealed that solo-LTR nested within satellite clades and profound homogeneity in their nucleotide composition preferences supported by a strong positive correlation (R>0.5).

SatLTR as an evolutionary intermediate in centromeric satellite formation

We identified a distinctive ‘intermediate evolutionary structure’ in legume centromeres, where in a solo-LTR encapsulates contiguously satellite repeats. For instance, in *Cicer arietinum* (Figure 4f), a complete solo-LTR precisely nested an array of eight tandem satellites. Serving as genomic ‘molecular fossil’, these chimeric structures captured the critical initial stages of centromeric repeat origination. Importantly, this LTR-satellite encapsulation was not species-specific, as we observed analogous arrangement patterns across multiple representative legumes. High-dimensional k-mer clustering and structural homology analyses revealed profound structural continuity (Figure S24). Despite lacking internal coding regions, these solo-LTR exhibited strong sequence affinity with derived satellite. Given their intermediate evolutionary role and capacity to seed satellite arrays, we designated them ‘satLTR’ (Figure S25). Cross-species divergence suggested that these satLTR originated from diverse transposon superfamilies undergoing independent evolutionary and tandemisation. Collectively, our analysis demonstrated that legumes centromeric satellite did not arise *de novo*; rather, they evolved from truncated genomic satLTR. Subsequent intense local tandem amplification and homogenization ultimately shaped the modern satellite-dominated centromeric

architectures. Specifically, persistent transposon remnants (solo-LTR) within legume centromeres evaded functional decay to act as evolutionary ‘seed sequences’. They provided a structural microenvironment that initiated the localized tandem arraying of short repeats via cooperative or block-wise amplification. Continuous sequence homogenization of these nascent arrays ultimately drove their evolution into the massive satellite landscapes observed today. We therefore demonstrated that legume centromeric satellite (e.g., in liquorice, soybean, and chickpea) originated from fragmented genomic satLTR rather *de novo*. Through intense localized amplification and homogenization, these ultimately shaped the modern, satellite-dominated centromeric architecture .

An Evolutionary Model for Centromere Repeat Transition in Fabaceae

Plant centromeres were highly dynamic loci shaped by a continuous interplay between satellite array turnover and retrotransposon activity. Comparative genomic studies across diverse taxa, including *Arabidopsis* and rice, revealed that centromeric satellites underwent cyclical patterns of local expansion, contraction, and homogenization. Concurrently, the frequent insertion, recombination, and elimination of centromeric retrotransposons (e.g.,ATHILA elements) act as primary drivers of structural divergence. This dynamic was prominently exemplified in the closely related *Rosaceae* family, where nascent tandem repeats had recently emerged directly from paracentric LTR elements. Collectively, this ongoing evolutionary ‘arms race’ between retrotransposon invasion and cyclical satellite homogenization continuously reshapes centromeric architectures across species.

By resolving high-resolution centromere maps for two *Glycyrrhiza* and 19 close relatives (Figure 2a &3b), we established a comprehensive evolutionary framework detailing the origin and amplification of centromeric satellite. While legume centromeres exhibited extensive intra- and interspecific structural diversity, satellite arrays consistently defined their core region. Notably, spatial analysis of solo-LTR in specific genome (e.g. Mtru_A17) revealed a distinct distribution:despite an overall depletion of solo-LTR in the centromeric cores compared to chromosomal arms, their relative abundance was evenly partitioned between centromeric and pericentromeric domains. We proposed this localized depletion reflects adaptive genomic pressures that drove the conversion of ancestral solo-LTR into functional satellites. Supported by spatial colocalization, sequence homology, and k-mer/PCA profiling (Figure 4), these data provided compelling evidence for a continuous evolutionary trajectory from fragmented solo-LTR, through

satLTR intermediates, to modern satellite arrays (Figure S24).

We proposed a unified model wherein suppressed recombination drove the initial accumulation of intact LTR retrotransposons within centromeric cores, which subsequent intramolecular recombination, unequal crossover, and truncation reduce to structurally simplified solo-LTR. A distinct subset-term satLTR, retained intrinsic sequence motifs that act as ‘evolutionary seeds’. Localized tandem and block amplification, coupled with continuous homogenization, progressively expanded these satLTR into continuous satellite arrays. Crucially, nested configurations where multiple satellite units remained encapsulated within a single satLTR, provided a molecular ‘fossil record’ capturing early array genesis. Cross-species comparisons across Fabaceae demonstrated that while this evolutionary trajectory was broadly conserved,lineage-specific divergence dictated the unique centromeric or pericentromeric enrichment of specific LTR families (e.g., CRM, SIRE, and Tekay) . Ultimately, our model illustrated that centromeric architecture was jointly dictated by the dynamic interplay of retrotransposon invasion, tandem repeat expansion and sequence homogenization.

While derived from representative legumes, our satLTR-driven model established a robust framework for decoding the highly dynamic evolution of complex centromeres. Beyond the Fabaceae family ,this paradigm provided a crucial conceptual foundation for broader genomic investigations into plant centromeric ,evolutionary transposons-satellite interplay, and the global of repetitive landscapes.

LTRtrace and LTRtrace-Search enable solo-LTR annotation and cross-species exploration

To systematically characterize LTR-RTs-derived sequences across chromosomal compartments, we developed LTRtrace, an automated annotation pipeline for intact long terminal repeat retrotransposons (LTR-RTs) reconstruction and compartment-resolved solo-LTR analysis (Figure S26). LTRtrace enabled genome-wide identification of intact LTR-RTs and solo-LTR, assigned solo-LTR to centromeric, pericentromeric and chromosome-arm regions, and generated standardized annotations and visualization outputs. By separating solo-LTR signals according to chromosomal context, LTRtrace provided a unified pipeline for comparing LTR-RTs dynamics across heterochromatic and euchromatic regions, thereby facilitating cross-species analyses of centromere-associated transposable element.

LTRtrace-Search provided, to our knowledge, the first cross-species interactive resource that links centromere and pericentromere-associated solo-LTR annotations with a graphical analysis interface (Figure S27). By integrating solo-LTR annotations, nonredundant intact LTR references and compartment-resolved genomic coordinates across 17 high-quality legume genomes, the platform enabled zero-command-line exploration of centromere-specific LTR-derived sequences. Unlike conventional transposable-element annotation pipelines, which primarily focus on genome-wide element detection, LTRtrace-Search was centered on solo-LTR associated with centromeric and pericentromeric regions. This design allowed users to distinguish centromere-enriched solo-LTR families from elements distributed along chromosome arms and to identify conserved LTR superfamilies with potential utility as centromere-targeting sequence resources.

Discussion

Based on a comparative analysis of representative legume species, we showed that satellite DNA was a major determinant of centromere structure in legumes and may represent a core configuration shaped by long-term centromere evolution. In particular, the satellite-dominated centromeres observed in *Glycyrrhiza* provides a high-resolution example of this organization model within the legume family. Although legume centromeres exhibited diverse structural patterns, including LTR-enriched, mixed, and satellite-dominated types, multidimensional evidence consistently points to satellite DNA as a key repetitive element underlying the stabilization of the centromeric core. This pattern reflected both broadly shared principles of centromere organization and lineage-specific features of legume evolution.

First, extensive retention of ancestral satellite polymorphisms in extant *Glycine* species may have generated substantial intraspecific variation. During long-term evolution, incomplete elimination and homogenization of these satellite variants increased intra-lineage diversity, suggesting that concerted evolution and sequence homogenization in *Glycine* lagged behind mutation accumulation and species divergence; consequently, intraspecific variation overlaps with, or even exceeds, interspecific variation. In contrast, these sequences appeared to have undergone long-term purifying selection in *Glycyrrhiza*, indicating a low turnover rate and a relatively stable balance between amplification and recombination. Together, satellite DNA evolution differed markedly among these three Fabaceae genera. These differences reflected lineage-specific evolutionary

histories shaped by speciation, co-evolutionary dynamics, and chromosomal functional constraints. Specifically, *Glycine* was characterized by polymorphism retention, *Medicago* by rapid homogenization and differentiation, and *Glycyrrhiza* by functional conservation.

Arabidopsis thaliana centromeres were dominated by CEN180 satellite repeats, yet they were periodically periodic invaded by ATHILA transposons and subsequently reshaped through satellite homogenization, revealing a dynamic interplay between satellites and transposons; In rice, functional centromeres were jointly composed of CEN155 satellites and centromere-associated transposons, forming locally homogenized and multilayered nested structures 9. In contrast, although legumes showed more pronounced phylogenetic differences in repeat composition, their long-term evolutionary trajectory appeared to converge on a similar outcome: stably satellite dominance within the centromeric core.

A key conclusions of this study was that legume centromeres likely did not arise *de novo*, but may have originated from solo-LTR remnants that accumulated in the centromeric and peri-centromeric regions and subsequently underwent truncation and structural transformation. Solo-LTR and satellites exhibited a markedly stronger spatial association within the centromeric core than in the PeriCEN region, suggesting that their co-occurrence was unlikely to be random and may instead be shaped by the distinctive chromatin environment and recombination-suppressive context of the centromeric. At the genome-wide scale, solo-LTR density was generally higher on chromosome arms than in centromeric regions; however, centromeric solo-LTR more frequently exhibit truncation, chimerism, and direct association with satellites. In some species, despite low overall solo-LTR density in centromeric regions, the fraction of satellites-associated solo-LTR was markedly elevated, suggesting that their presence in the centromeric core reflects selective retention and structural conversion rather than simple accumulation. This evolutionary pattern was conceptually consistent with the cyclical process in *Arabidopsis*, in which ATHILA invasion was followed by satellite-mediated homogenization, implying that transposon-derived sequences in centromeres were not merely passive remnants but may serve as precursors or participants in subsequent satellite remodeling.

Sequence-level analysis further corroborated this hypothesis. Despite extensive diversification across legume species, consensus of centromeric satellite families retained conserved nucleotide sites, pointing toward a shared a ancestral origin. For instance, the conserved core motif within the

Glycyrrhiza CEN60 family shared detectable homology with satellite consensus sequences from multiple legume taxa, indicating that these centromeric satellites retained deeply conserved signatures despite rapid evolutionary turnover. Crucially, centromeric solo-LTR displayed significant sequence similarity not only to the host satellite monomer but also to the terminal repeat of intact LTR transposons.This dual homology suggested that these solo-LTR were neither typical LTR remnants nor fully mature satellites, but rather represented a transitional evolutionary intermediate bridging the two.

K-mer composition analyses further substantiated this finding. Principal component analysis (PCA) revealed that solo-LTR from the CEN and PeriCEN regions do not cleanly separated from mature satellites: rather,they displayed partial Co-localized and continuous trajectory in the projection space, reflecting clear compositional continuity. Hierarchical clustering corroborated this by uncovering nested relationships and significant positive correlation between specific LTRs and satellites. Crucially,we identified characteristic evolutionary intermediates across multiple legume species, wherein a single solo-LTR fragment encompass or intersperse with contiguous satellite units. These chimeric arrays served as ‘molecular fossils’ of the satellite genesis, capturing the critical early stages where short repeat motifs emerge at truncated LTR boundaries or internal regions prior to expanding into tandem arrays. This trajectory paralleled observations in rice where Cent155 satellites exhibited chromosome-specific homogenization and multi-layered organization suggesting that localized amplification coupled with sequence homogenization represented a broadly conserved mechanism paradigm for the emergence and stabilization of plant centromeric satellites.

The spatial differentiation of solo-LTR families further revealed that the transition from transposon remnants to satellites does not follow a uniform trajectory across the Fabaceae. Notably, the CRM lineage exhibited the most pronounced enrichment within the centromeric core-consistent with the functional attributes of centromeric Co-localized retrotransposons; whereas families such as SIRE, ATHILA, and Tekay were more broadly dispersed across pericentromeric or wider chromosomal domains. The lineage-specific expansion of certain families suggested that diverging taxa exploit distinct transposon templates as primary substrates for satellite innovation. Consequently, while the generative mechanisms of legume centromeric satellites were conserved, the specific satLTR serving as ‘evolutionary seeds’ were highly specie or lineage-specific. Concordantly, prior cytogenetic studies in alfalfa revealed a multiplicity of satellite and minisatellite families within

these regions, indicating that there was no singular structural blueprint for Fabaceae centromeres; rather,diverse repeat elements can assembly into distinct configurations to fulfill shared centromeric requirements.

Overall, this study shifted our understanding of legume centromeres evolution from a paradigm of transposons-satellite coexistence to one where satellite sequences originate, at least in partly, from transposon remnants. In this sense, satellite DNA should be viewed not merely as a static scaffold for mature centromeres, but as a major evolutionary product of repetitive sequence remodeling in legumes. Consequently, we proposed that the central evolution axis of legmue centromeres comprised a continuous transition from solo-LTR to satLTR and ultimately to mature centromeric satellites, driven primarily by localized amplification and homogenization. This framework not only elucidated the composition and origin of legume centromeres but also established a new conceptual foundation for explaining transposons-satellite co-evolution, structural plasticity, and repeat sequence innovation in plant centromeres.

Methods and Materials

Whole-genome sequencing

We conducted whole genome DNA sequencing using the BGI, ONT and PacBio HiFi platforms. Genomic DNA was extracted from the callus tissues of *Glycyrrhiza glabra* and *Glycyrrhiza inflata* using the QIAGEN Genomic DNA Kit. For BGISeq, DNA was directly isolated using CTAB method, BGISeq DNA PCR-free library were constructed and sequenced on an BGI T7 platform. For Pacbio HiFi sequencing, a standard HiFi library was prepared according to the SMRTbell Express Template Prep Kit 2.0 manual (Pacific Biosciences, CA, USA). For ONT ultra-long sequencing, the standard library was prepared using the SQK-ULK001 kit, following the standard protocol. The purified library was sequenced using a PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK). We used Hi-C sequencing technology to assist scaffolding. The isolated cross-linked DNAs were purified, digested with Mbo I enzyme, blunt-end-repaired, and tagged with biotin. The biotin-containing DNA fragments were captured and PCR-enriched to construct a Hi-C library. The library was then sequenced on the BGI T7 platform according to the PE150 strategy.

Genome assembly

The project employs the ul mode of Hifiasm (v0.19.9-r616)[20] for the hybrid assembly of HiFi reads and ONT reads. *De novo* assembly of the raw ONT long reads was performed using NextDenovo[21] assembler (v2.2) with the default parameters read_cutoff =1 k, seed_cutoff=25,211, and NextPolish (v1.3.0)[22] was used six times [ONT long reads (two rounds) and short reads (four rounds)] to polish the initial draft assembled contigs. Because this genome had a high heterozygosity rate, we used purge_dups (v1.2.3)[23]to remove primary contigs that had both haplotypes rather than a contig and its associated haplotig.

Trimmomatic (v0.40)[24] was used to trim high-quality paired-end reads to remove low-quality bases and adapter sequences from the reads, and all filtered reads were aligned to contigs using a juicer (https://github.com/aidenlab/juicer, v3)25 to calculate the contact frequency. Then, 3ddna (v18.09.22)[25] was used with two iterative rounds for misjoin correction (-r1) using default parameters. The oriented scaffolds were utilized to generate the interaction matrices with a juicer to inspect and manually correct with Juicebox assembly tools (v1.11.08). The gap-free genome was used to fill the gaps in the genome generated by reads and contigs with TGS-GapCloser[26]. To correct potential SNV and SV, we followed the polishing pipeline of McCartney and colleagues[27].

Telomere identification

For telomere identification, plant telomeric sequences (CCCTAAA) were identified, and the expected 16 telomeres (8 chromosomes) were identified using the telomere pipeline developed by the VGP (https://github.com/VGP/vgp-assembly). Telomeric reads/contigs (HiFi, ONT reads,contigs) were manually identified, and telomeric sequences were patched for the telomere-missing chromosomes.

Assessment of assembly quality

Complementary methods were employed to evaluate the quality of the genome assembly. First, genome completeness was assessed based on conserved plant genes in the BUSCO (v3.0.2)[29] databases. Second, Illumina/BGI short-reads were mapped to the assembled genome using BWA (v0.7.8) to assess coverage rate and average depth. Third, TGS long-reads were mapped to the assembled genome using minimap (v2.21)[30] to assess coverage rate and average depth. Fourth, The LTR_retriever (v1.0.7)[31] package was used to evaluate genome quality.Fifth, Mequery (v1.3)[32] was used to assess the consensus quality (QV) value and completeness of the genome assembly.

Repeat annotation

The repetitive sequences, including tandem repeats and TEs were searched. First, we used Tandem Repeats Finder (TRF, v4.09.1) to annotate the tandem repeats using the following parameters: 2 7 7 80 10 50 2000. Then TEs were identified at both the DNA and protein levels using a combination of *de novo* and homology-based approaches. At the DNA level, LTR_FINDER (v1.0.7)[34] was first used to identify LTR-RTs and RepeatModeler (v2.0.1)[35] was utilized to construct a *de novo* repeat library, which comprised a repeat consensus database with classification information. We employed RepeatMasker (v4.1.2)[36] to search for similar TEs in the known Repbase TE library and *de novo* repeat library. At the protein level, RepeatProteinMask within the RepeatMasker package was used to search against the TE protein database using a WU-BLASTX engine.

Gene annotation

We used homologous, ab initio and transcriptome-assisted annotation to predict the structure of coding genes. For homologous annotation, Tblastn (v2.11.0+) was used to compare the related species to the reference genome. Then, the aligned sequences and their corresponding proteins were filtered and transmitted to the Exonerate (v2.4.0)[37] for accurate alignment. Augustus (v3.4.0)[38] and GlimmerHMM (v3.0.4)[39] were used for *de novo* annotation. For RNA-seq data, we use both *de novo* and genome-based transcriptome assemblies. RNA-seq alignments were produced using HiSat2 (v.2.2.1)[40] and then RNA-seq alignments were further assembled into transcripts with genome-guided assembler Stringtie (v.2.1.7)[41]. Additionally, the transcriptome was assembled *de novo* using Trinity (v2.8.5)[42]. We built a comprehensive transcriptome database using all transcripts from RNA-seq and Iso-seq according to PASA pipeline (v2.4.1). Maker (v3.01.03)[43] was used to integrate the predicted gene sets into a nonredundant, more complete and reliable gene set. Finally, PASA pipeline (v2.4.1) was used to update maker consensus predictions, adding UTR annotations and models for alternatively spliced isoforms.

Functional annotations

Gene functions were inferred according to the best match of the alignments to the National Center for Biotechnology Information (NCBI) Non-Redundant (NR), Kyoto Encyclopedia of Genes and Genomes (KEGG) database[44], Gene Ontology (GO)[45], TrEMBL and Swiss-Prot[46] protein databases using Diamond BLASTP (v2.0.7)[47] with an E-value threshold of 1E-5. The protein domains were

603 annotated using InterProScan (v5.50-84.0)[48] based on InterPro[49] protein databases.

604 **Annotation of non-coding RNA genes**

605 We used tRNAscan-SE (v2.0.9) algorithms[50] with default parameters to identify the genes associated
606 with tRNA, which is an adaptor molecule composed of RNA used in biology to bridge the
607 three-letter genetic code in messenger RNA (mRNA) with the twenty-letter code of amino acids in
608 proteins. RNAmmer (v1.2)[51] was used to predict rRNA sequences. snoRNAs are a class of small
609 RNA molecules that guide chemical modifications of other RNAs, mainly ribosomal RNAs, transfer
610 RNAs and small nuclear RNAs. MiRNAs and snRNAs were identified by Infernal (v1.1.2)[52]
611 software against the Rfam (v14.6) database[53] with default parameters.

612 **Satellite annotation and analysis**

613 For satellite repeat annotation we applied TRASH, which uses periodic k-mers to identify tandem
614 repeats[17],using TRASH (https://github.com/vlothec/TRASH). Genomic sequences were segmented
615 into 1-Kb windows to assess local k-mer counts and identify repetitive regions. Score of each
616 window was determined based on the proportion of repeated k-mers, with regions scoring above a
617 defined threshold considered to contain repeats. Tandem repeats within these windows were
618 characterized by the distances between identical k-mers, aiding in the identification of consensus
619 sequences. To pinpoint centromeric regions within each chromosome, all satellite repeats
620 identification as fixed-length from the TRASH results were extracted and mapped according to their
621 chromosomal positions. A fixed-length array within one's own genome, with a repeat unit spacing of
622 no exceed 150 Kb, is defined as a centromeric block. The largest contiguous block on each
623 chromosome is regarded as a candidate centromeric region.

624 **Structural organization analysis of satellite arrays**

625 We followed the methodology of the rice pan-centromere study[9] to analyze the structure of satellite
626 arrays using an incremental compression strategy. This involved iteratively merging identical
627 monomer units. Monomeric homogeneous regions (moHRs) were identified when more than five
628 consecutive satellites from the same family were present. Transposons (TEs) embedded within the
629 satellite arrays were masked.

630 Based on the moHR-compressed satellite arrays, we defined diad homogeneous regions (diHRs)
631 consisting of at least five diad repeat units. For example, in the diad-compressed centromere

sequence ‘Gly1Gly2Gly3Gly4Gly5’, the corresponding edge set (edge set) includes the pairs Gly1-Gly2{2}, Gly2-Gly3{2}, Gly3-Gly4{2}, and Gly4-Gly5{1}. The multiplicity (multiplicity) of the edge (Gly1, Gly2) is defined as the number of times monomer family Gly1 is immediately followed by monomer family Gly2 in each diad-compressed centromere. The traversal count represents the frequency with which a given cycle appears in each dimer-compressed centromere sequence.

Gly-box and motif analysis

To analyze the conservation of conserved motifs across multiple genomes, we selected satellite populations from three representative genera within the sample set and performed whole-genome scans on each. We first analyzed the data using MEME v.5.4.1 to obtain frequency matrices, and then constructed motif libraries for each corresponding genus. Subsequently, we compared the motif libraries of the three genera against the whole genome using the default parameters of FIMO, thereby identifying individual motifs that occurred and were identified across the entire genome.

Sequence Alignment and Identity Calculation

Pairwise sequence alignments between centromeric satellite consensus sequences and the reference (e.g.,CEN60) were performed using the MAFFT L-INS-i strategy (--localpair --maxiterate 1000). To account for the length heterogeneity inherent to satellite DNA, sequence identity was calculated using the PID3 metric. Specifically, identity is defined as the number of exactly matched base pairs divided by the original length of the shorter sequence. This approach effectively reflects true local homology while eliminating alignment gap biases caused by length variations.

Characterizing sequence variation and evolutionary trajectories in satellite DNA block-wise amplification

To systematically characterize large-scale amplification events within centromeric satellite arrays, we developed a three-step pipeline using TRASH-annotated centromeric repeats as input.

Step 1: Conserved block detection. Satellite monomers previously identified and classified into higher-order repeat (HOR) families by TRASH were extracted, and their multiple sequence alignments (MSA) from the TRASH HOR output were used directly. For each centromeric satellite family per chromosome, aligned monomer sequences were subjected to a sliding-window similarity search: for each monomer at position i, the Hamming distance was computed against downstream

monomers within a window of 1,000 positions. Pairs with Hamming distance ≤5 served as seeds, which were greedily extended bidirectionally while maintaining the distance threshold and consistent strand orientation. This identified two categories of conserved blocks: (i) tandem blocks — two blocks on the same strand co-directionally extended, and (ii) inverted blocks — two blocks on opposite strands extending toward each other. Each block pair was annotated with genomic coordinates, block size (number of monomers), and cumulative variant score. Each monomer also received a Repetition Score representing the total number of blocks in which it participates.

Step 2: Divergence-distance landscape. For each block pair, we computed the genomic distance between the two blocks (kbp) and the per-monomer divergence (cumulative variant score / block size). A point-density scatter plot of these two metrics was generated using two-dimensional kernel density estimation, revealing the global distribution of recent (closely spaced, low-divergence) versus ancient (widely separated, high-divergence) amplification events.

Automated LTR-RTs Reconstruction and solo-LTR Annotation Pipeline

LTRtrace (https://github.com/Liurxx/LTRtrace) was developed as an R-based command-line pipeline for genome-wide reconstruction of long terminal repeat retrotransposons (LTR-RTs) and compartment-resolved annotation of solo-LTR. The pipeline enable solo-LTRs to be consistently assigned to centromeric, pericentromeric and chromosome-arm regions.Intact LTR-RTs candidates were first identified *de novo* using two complementary tools, LTR_Finder (v.1.2) and LTRharvest (v.1.6.2). The resulting candidate sets were merged and refined with LTR_retriever (v.2.9.0) to generate a high-confidence intact LTR-RTs set, an LTR reference library and genome-wide repeat annotations. The curated intact LTR-RTs sequences were classified using TEsorter (v.1.3). Complete LTR elements were then extracted and clustered with CD-HIT (v.4.8.1) to construct a non-redundant LTR reference library. Alternatively, users can bypass this step by using the precompiled Fabaceae LTR libraries distributed with LTRtrace.

Solo-LTR candidates were detected using solo_finder.pl from LTR_retriever and classified by sequence similarity against the corresponding LTR reference library using BLASTn/BLAST+ (v.2.16.0). For each solo-LTR, the best-supported hit was used to assign its superfamily or family annotation. To improve classification sensitivity in repetitive heterochromatic regions, initially unclassified centromeric and pericentromeric solo-LTR were subjected to a relaxed BLAST rescue

search. Chromosome-arm solo-LTR were processed independently using the same primary and rescue-classification strategy, allowing compartment-specific LTR dynamics to be evaluated without cross-region interference. LTRtrace outputs standardized annotation tables, genomic coordinates, compartment-specific FASTA files and graphical summaries for centromeric, pericentromeric and chromosome-arm solo-LTR. Together, this workflow provides an end-to-end framework for reconstructing intact LTR-RTs, annotating solo-LTR and comparing LTR-RTs dynamics across distinct chromosomal compartments.

LTRtrace-Search platform and centromere -associated solo-LTR annotation

We developed LTRtrace-Search (https://github.com/Liurxx/LTRtrace-Search) as an interactive cross-species platform for the annotation, visualization and comparative analysis of centromere and pericentromere-associated solo-LTR. The platform integrates precomputed solo-LTR annotations, nonredundant full-length LTR retrotransposon references and genome-scale positional information into a graphical user interface, enabling users to query, compare and visualize solo-LTR distributions across species without command-line operation.

For the current release, high-quality genome assemblies from 17 legume species were used to construct a systematic centromere-associated LTR resource. solo-LTR and intact LTR retrotransposons were annotated and curated for each genome, followed by redundancy reduction of full-length LTR elements to generate species-level reference sets. Genomic locations were further assigned to three compartments: centromeric core regions, pericentromeric regions and chromosome arms. This three-part partitioning allowed centromere-enriched solo-LTR to be distinguished from broadly distributed or chromosome-arm-associated LTR-derived sequences.

LTRtrace-Search was specifically designed to focus on centromere-associated solo-LTR, a class of LTR-derived sequences that is not explicitly targeted by existing general-purpose transposable-element annotation tools. The platform supports cross-species comparison of solo-LTR abundance, family or superfamily composition, genomic distribution and compartment-specific enrichment. In particular, solo-LTR sequences located within centromeric core regions were retained as a sequence resource for downstream functional and synthetic-biology applications. Conserved solo-LTR superfamilies that are enriched in centromeric regions across species but absent or depleted from chromosome arms may provide candidate sequence modules for the design of

centromere-targeting vectors.

DNA methylation analysis using ONT data

To identify CG, CHG and CHH methylation sites, we used DeepSignal3 (https://github.com/PengNi/deepsignal3) to detect 5-methylcytosine methylation in ONT sequencing reads. Dorado was used for base calling, followed by sequence realignment. Subsequently, DeepSignal3 was utilized in conjunction with the corresponding model (plant_r1041_5khz_C_epoch4.ckpt) to predict methylation at CG, CHG and CHH sites. Finally, the call_freq function within the DeepSignal3 package was used to generate the methylation frequency for each CG, CHG and CHH site. Cytosines with fewer than five covering reads were excluded from the DNA methylation analysis. Finally, methylation levels were summarized in 50 kb non-overlapping windows.

Declarations

Data availability

Raw sequencing data generated in this study have been deposited in the NGDC BioProject database under accession PRJCA064152, with assembled genomes available at https://ngdc.cncb.ac.cn/.

Code availability

The source code of software for LTRtrace and LTRtrace-Search are publicly available on GitHub at https://github.com/Liurxx/LTRtrace and https://github.com/Liurxx/LTRtrace-Search , respectively.

Acknowledgement

This work was supported by the Key Research and Development Program of Xinjiang Uygur Autonomous Region (2024B02024, 2024B02024-2). The authors would like to thank the Bioinformatics Platform at Peking University Institute of Advanced Agricultural Sciences for providing the high-performance computing resources.

Author contributions

Bin Guo and Li Guo designed and supervised the project. Rongxing Liu was responsible for data curation, formal analysis, software, database construction, visualization, and writing original draft. Weikai Chen contributed to formal analysis and script development. Gang Zhou provided wild plant materials of *Glycyrrhiza glabra* and *Glycyrrhiza inflata*. Jie Feng was responsible for plant

cultivation and maintenance of licorice tissues/materials. Huan Wang, Ke Wang, and Rui Li contributed to methodology and formal analysis and provided important guidance and methodological support for data analysis. Yanping Fu and Wei He contributed to the interpretation and discussion of the results. Xizhen Yue, Jianing Wang, Tianzi Yao, Yongjie Qin, Yirui Zhao, He Wang, and Hong Tao. provided resources and assisted with material preparation during the early stages of the study.

Declaration of interests

The authors declare no conflict of interests.

Ethics declaration

not applicable.

Reference

1. Naish, M. and Henderson, I.R. (2024). The Structure, Function, and Evolution of Plant Centromeres. *Genome Res* 34, 161-78.

2. Chen, W., Wang, X., Sun, J., et al. (2024). Two Telomere-to-Telomere Gapless Genomes Reveal Insights into Capsicum Evolution and Capsaicinoid Biosynthesis. *Nat Commun* 15, 4295.

3. Chen, W., Yan, M., Chen, S., et al. (2024). The Complete Genome Assembly of Nicotiana benthamiana Reveals the Genetic and Epigenetic Landscape of Centromeres. *Nat Plants* 10, 1928-1943

4. Chen, W., Chen, S., Wang, J., et al. (2026). Gapless and Phased Genome Assembly Reveals Post-Allopolyploidization Subgenomic Diversification of Tobacco Centromeres. *Genome Biol*, in press.

5. Guo, L., Wang, X., Ayhan, D.H., et al. (2025). Super Pangenome of Vitis Empowers Identification of Downy Mildew Resistance Genes for Grapevine Improvement. *Nat Genet* 57, 741-753.

6. Wlodzimierz, P., Rabanal, F.A., Burns, R., et al. (2023a). Cycles of Satellite and Transposon Evolution in Arabidopsis Centromeres. *Nature* 618, 557-65.

7. Herklotz, V., Zhang, M., Nascimento, T., et al. (2025). Bimodal Centromeres in Pentaploid

774 Dogroses Shed Light on Their Unique Meiosis. *Nature* 643, 148-57.

775 8. Chen, W., Wang, J., et al. (2025). Pancentromere Landscape and Dynamic Evolution in Brassica
776 Plants. *Nat Plants*, 11, 2240–2253.

777 9. Xie, L., Huang, Y., Huang, W., et al. (2025). Genetic Diversity and Evolution of Rice
778 Centromeres. *Nat Genet*, 57, 2808-2818.

779 10. Shi, Y., Zhang, C., Yang, W., et al. (2026). Toward Telomere-to-Telomere Genomics in Fabaceae:
780 Unlocking Comparative and Functional Insights into Symbiotic Nitrogen Fixation. *Cell Genom*
781 6, 101246.

782 11. Hou, J., Wei, Y., Liu, H., et al. (2026). Centromere Evolution in Annual and Perennial Soybeans
783 and Its Implication for Hybridization in Cultivated Species. *Genome Biol* 27, 90.

784 12. Sod, B., Jiang, X., He, F., et al. (2026). The Haplotype-Resolved and near-Telomere-to-Telomere
785 Genome Assembly for the Autotetraploid Alfalfa. *Plant Commun* 7, 101691.

786 13. Zhao, B., Zhang, H., Zhao, Q., et al. (2025). Gap-Free Genome Assembly and Metabolomics
787 Analysis of Common Bean Provide Insights into Genomic Characteristics and Metabolic
788 Determinants of Seed Coat Pigmentation. *J Genet Genomics* 52, 852-55.

789 14. Liu, Y., Yi, C., Fan, C., et al. (2023). Pan-Centromere Reveals Widespread Centromere
790 Repositioning of Soybean Genomes. *Proc Natl Acad Sci* USA 120, e2310177120.

791 15. Shen, L., Yi, C., Liu, Y., Han, F., and Feng, J. (2025). Two Complete Telomere-to-Telomere
792 Genome Assemblies of Medicago Reveal the Landscape and Evolution of Its Centromeres. *Mol*
793 *Plant* 18, 1409-12.

794 16. Li, Y., Xia, C., Luo, M., et al. (2025). Comparative Genomics of Three Medicinal *Glycyrrhiza*
795 Species Unveiled Novel Candidates for the Production of Important Bioactive Compounds.
796 *Plant J* 122, e70223.

797 17. Wlodzimierz, P., Hong, M., and Henderson, I.R. (2023b). TRASH: Tandem Repeat Annotation
798 and Structural Hierarchy. *Bioinformatics* 39, btad308.

799 18. Grob, S., Schmid, M.W., and Grossniklaus, U. (2014). Hi-C Analysis in Arabidopsis Identifies

800 the KNOT, a Structure with Similarities to the Flamenco Locus of Drosophila. *Mol Cell* 55,
801 678-693.

802 19. Dong, P., Tu, X., Chu, P.-Y., et al. (2017). 3D Chromatin Architecture of Large Plant Genomes
803 Determined by Local A/B Compartments. *Mol Plant* 10, 1497-1509.

804 20. Cheng, H., Jarvis, E.D., Fedrigo, O., et al. (2022). Haplotype-Resolved Assembly of Diploid
805 Genomes without Parental Data. *Nat Biotechnol* 40, 1332-1335.

806 21. Hu, J., Wang, Z., Sun, Z., et al. (2024). An Efficient Error Correction and Accurate Assembly
807 Tool for Noisy Long Reads. *Genome Biol* 25, 107.

808 22. Hu, J., Fan, J., Sun, Z., et al. (2020). A Fast and Efficient Genome Polishing Tool for Long-Read
809 Assembly. *Bioinformatics* 36, 2253-2255.

810 23. Guan, D., McCarthy, S.A., Wood, J., et al. (2020). Identifying and Removing Haplotypic
811 Duplication in Primary Genome Assemblies. *Bioinformatics* 36, 2896-2898.

812 24. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A Flexible Trimmer for Illumina
813 Sequence Data. *Bioinformatics* 30, 2114-2120.

814 25. Durand, N.C., Shamim, M.S., Machol, I., et al. (2016). Juicer Provides a One-Click System for
815 Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* 3, 95-98.

816 26. Xu, M., Guo, L., Gu, S., et al. (2020). TGS-GapCloser: A Fast and Accurate Gap Closer for
817 Large Genomes with Low Coverage of Error-Prone Long Reads. *Gigascience* 9, giaa094.

818 27. Mc Cartney, A.M., Shafin, K., Alonge, M., et al. (2022). Chasing Perfection: Validation and
819 Polishing Strategies for Telomere-to-Telomere Genome Assemblies. *Nat Methods* 19, 687-695.

820 28. Benson, G. (1999). Tandem Repeats Finder: A Program to Analyze DNA Sequences. *Nucleic*
821 *Acids Res* 27, 573-580.

822 29. Simão, F.A., Waterhouse, R.M., Ioannidis, P., et al. (2015). BUSCO: Assessing Genome
823 Assembly and Annotation Completeness with Single-Copy Orthologs. *Bioinformatics* 31,
824 3210-3212.

825 30. Li, H. (2018). Minimap2: Pairwise Alignment for Nucleotide Sequences. *Bioinformatics* 34,

826 3094-3100.

827 31. Ou, S., Chen, J., and Jiang, N. (2018). Assessing Genome Assembly Quality Using the LTR
828 Assembly Index (LAI). *Nucleic Acids Res* 46, e126.

829 32. Rhie, A., Walenz, B.P., Koren, S., et al. (2020). Merqury: Reference-Free Quality, Completeness,
830 and Phasing Assessment for Genome Assemblies. *Genome Biol* 21, 245.

831 33. Benson, G. (1999). Tandem Repeats Finder: A Program to Analyze DNA Sequences. *Nucleic*
832 *Acids Res* 27, 573-580.

833 34. Xu, Z. and Wang, H. (2007). LTR_FINDER: An Efficient Tool for the Prediction of Full-Length
834 LTR Retrotransposons. *Nucleic Acids Res* 35, W265-W268.

835 35. Flynn, J.M., Hubley, R., Goubert, C., et al. (2020). RepeatModeler2 for Automated Genomic
836 Discovery of Transposable Element Families. *Proc Natl Acad Sci USA* 117, 9451-9457.

837 36. Tarailo-Graovac, M. and Chen, N. (2009). Using RepeatMasker to Identify Repetitive Elements
838 in Genomic Sequences. Curr Protoc *Bioinformatics* 25, 4.10.1-4.10.14.

839 37. Slater, G.S.C. and Birney, E. (2005). Automated Generation of Heuristics for Biological
840 Sequence Comparison. *BMC Bioinformatics* 6, 31.

841 38. Stanke, M., Steinkamp, R., Waack, S., et al. (2004). AUGUSTUS: A Web Server for Gene
842 Finding in Eukaryotes. *Nucleic Acids Res* 32, W309-W312.

843 39. Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM: Two Open
844 Source Ab Initio Eukaryotic Gene-Finders. *Bioinformatics* 20, 2878-2879.

845 40. Kim, D., Paggi, J.M., Park, C., et al. (2019). Graph-Based Genome Alignment and Genotyping
846 with HISAT2 and HISAT-Genotype. *Nat Biotechnol* 37, 907-915.

847 41. Kovaka, S., Zimin, A.V., Pertea, G.M., et al. (2019). Transcriptome Assembly from Long-Read
848 RNA-Seq Alignments with StringTie2. *Genome Biol* 20, 278.

849 42. Grabherr, M.G., Haas, B.J., Yassour, M., et al. (2011). Full-Length Transcriptome Assembly
850 from RNA-Seq Data without a Reference Genome. *Nat Biotechnol* 29, 644-652.

851 43. Cantarel, B.L., Korf, I., Robb, S.M.C., et al. (2008). MAKER: An Easy-to-Use Annotation
852 Pipeline Designed for Emerging Model Organism Genomes. *Genome Res* 18, 188-196.

853 44. Kanehisa, M., Goto, S., Sato, Y., et al. (2012). KEGG for Integration and Interpretation of
854 Large-Scale Molecular Data Sets. *Nucleic Acids Res* 40, D109-D114.

855 45. Ashburner, M., Ball, C.A., Blake, J.A., et al. (2000). Gene Ontology: Tool for the Unification of
856 Biology. *Nat Genet* 25, 25-29.

857 46. Boeckmann, B. (2003). The SWISS-PROT Protein Knowledgebase and Its Supplement
858 TrEMBL in 2003. *Nucleic Acids Res* 31, 365-370.

859 47. Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive Protein Alignments at Tree-of-Life
860 Scale Using DIAMOND. *Nat Methods* 18, 366-368.

861 48. Jones, P., Binns, D., Chang, H.-Y., et al. (2014). InterProScan 5: Genome-Scale Protein Function
862 Classification. *Bioinformatics* 30, 1236-1240.

863 49. Mitchell, A., Chang, H.-Y., Daugherty, L., et al. (2015). The InterPro Protein Families Database:
864 The Classification Resource after 15 Years. *Nucleic Acids Res* 43, D213-D221.

865 50. Chan, P.P., Lin, B.Y., Mak, A.J., et al. (2021). tRNAscan-SE 2.0: Improved Detection and
866 Functional Classification of Transfer RNA Genes. *Nucleic Acids Res* 49, 9077-9096.

867 51. Lagesen, K., Hallin, P., Rødland, E.A., et al. (2007). RNAmmer: Consistent and Rapid
868 Annotation of Ribosomal RNA Genes. *Nucleic Acids Res* 35, 3100-3108.

869 52. Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. (2009). Infernal 1.0: Inference of RNA Alignments.
870 *Bioinformatics* 25, 1335-1337.

871 53. Kalvari, I., Nawrocki, E.P., Ontiveros-Palacios, N., et al. (2021). Rfam 14: Expanded Coverage
872 of Metagenomic, Viral and microRNA Families. *Nucleic Acids Res* 49, D192-D200.

Figure Legends

Figure 1 | High genetic diversity in the *G. glabra* and *G. inflata* centromere.

(A) Circos plot illustrating genomic features of the *G. glabra.* Tracks from outer to inner represent: chromosomes, densities of Gypsy, Copia, GC, satellite, gene, CG, CHG and CHH methylation contexts. **(B)** The bottom panel shows the frequency of different monomer families in CEN60 as displayed in Glygla for CEN1. The two middle track diagrams illustrate the sequence identity between the monomer families and the consensus sequence. The top figure displays the StainedGlass sequence identity heatmap. **(C)** To represent satellite centromere architecture, chromosome 3 of *G. glabra* (Glygla) is plotted showing entropy rate (black), and the proportion of sequence that was CEN60 satellite (red), exon (green), and Gypsy (red), and Copia (blue). **(D)** Data as in C, but analyzed in scaled windows between the center of CEN60 arrays and the telomeres. **(F)** Quantification of CEN60 DNA methylation plotted across the centromeres in Glygla. **(H-J)** As for C-F, but analyzing Glyinf. **(E)** Heatmap of a representative region within Glygla CEN1, shaded according to pairwise variants between CEN60. **(K)** Frequency of CEN60 sequence variants across the consensus repeat. The colour-coded DNA sequence is shown below (red, A; dark red, T; navy, G; azure, C).

Figure 2 | Overview of the centromeric genetic and epigenetic variation between two *Glycyrrhiza* genomes. (A) Complete assembly of centromeres from two *Glycyrrhiza* genomes, Glygla and Glyinf, reveals both small and large-scale variation in centromere sequence, structure and epigenetic landscape. **(B)** A maximum-likelihood phylogenetic tree of representative CEN60 satellite sequences. Nine superfamilies (Gly1 to Gly9) are clustered and colored. **(C)** The number of copies of biased monomers in different chromosomes varies in the replication of dimers. Dimers of the same type form distinct blocks, with multiple such blocks recurring. Variable copies of both types of monomers are present, resulting in a pattern of X_nY_m or X_mY_n.

Figure 3 | Comprehensive analyses of centromere landscapes in diverse Fabaceae genomes.

(A) Dated phylogenetic relationships and genome sizes of the analyzed Fabaceae species. These species represent major Fabaceae species and external group, including *Papilionoideae*, *Cercidoideae*, *Detarioideae*, *Myricaceae*, and *Morella rubra*. Genome sizes (Mb) were estimated using available reference genomes. **(B) Left:** Bar plots showing the proportion of the centromeric arrays that are composed of satellites (pink), transposons (blue), or other sequences (grey), according to major taxonomic groups. Middle: As for a, but showing the size of the centromere satellite regions. **Right:** As for a, but showing LTR identity values for intact LTR retrotransposons (green) in the centromere region. **(C-D)** as in Fig. 1C, but analysing *Medicago sativa* (tetraploid) and *Medicago truncatula* (R108). **(E-F)** A maximum-likelihood phylogenetic tree of representative *Glycine* and *Medicago* satellite sequences. **(G-I)** Data as in c, but analysed in scaled windows between the center

of satellite arrays and the telomeres.

Figure 4 | Sequence similarity between solo-LTR and satellites in major Fabaceae species.

(A) The red annotation bars represent solo-LTR on the centromere, whilst the orange annotation bars represent solo-LTR near the centromere, moving towards the black dots (satellites). From each cluster, 500 sequences were randomly selected to compare the frequency vectors in the corresponding dimensions (k=3). **(B)** Data as in A, but analysed used PCA for cluster analysis **(C)** Gly-Box: Identification of conserved motifs using the CEN60 population of *Glycyrrhiza*. **(D-E)** Sequence similarity analysis was performed using the Gly-Box sequence and sequences from soybeans and chickpeas. **(F)** A similarity analysis was performed on satellites within the solo-LTR region, using chickpeas as a representative example.

Figure 5 | The diagram illustrates the evolutionary pattern in Fabaceae.

(A) Cladogram representation of select species. To the right the results of FIMO analysis are shown for each species' centromeric regions satellite examining the presence of the indicated motif. **(B) Right:** Comparison of sequence similarity in the consensus sequence of the satellite element of selected species. **Left:** Conserved sites identified by sequence alignment between the consensus sequence of the satellite element of the selected species and the Gly-Box. **Bottom:** Sequence alignment results between the Gly-Box motif and motifs identified in the complete consensus satellite sequences of the selected species. **(C)** The evolutionary origin and amplification patterns of satellite DNA in Fabaceae.

Figure 1

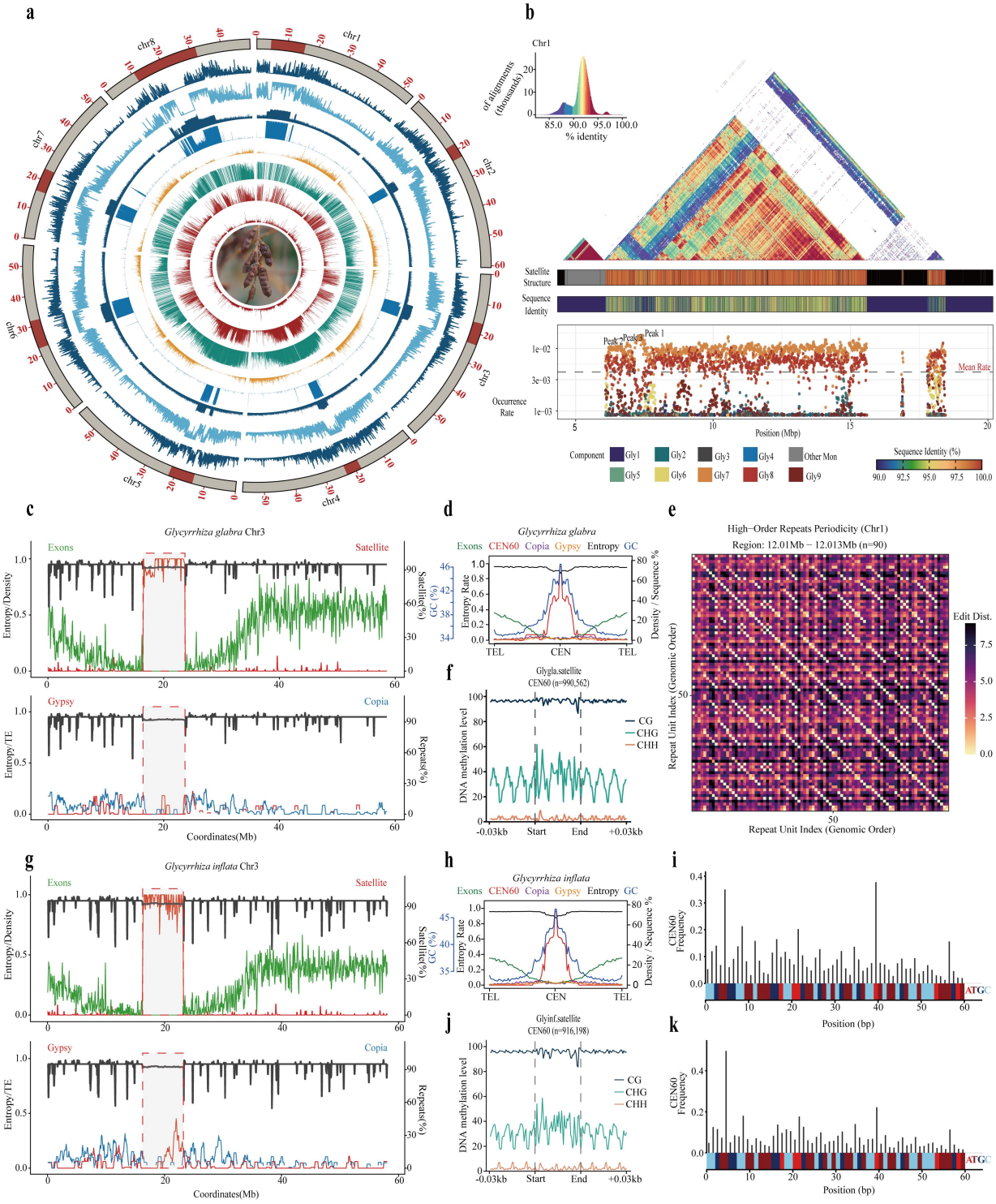
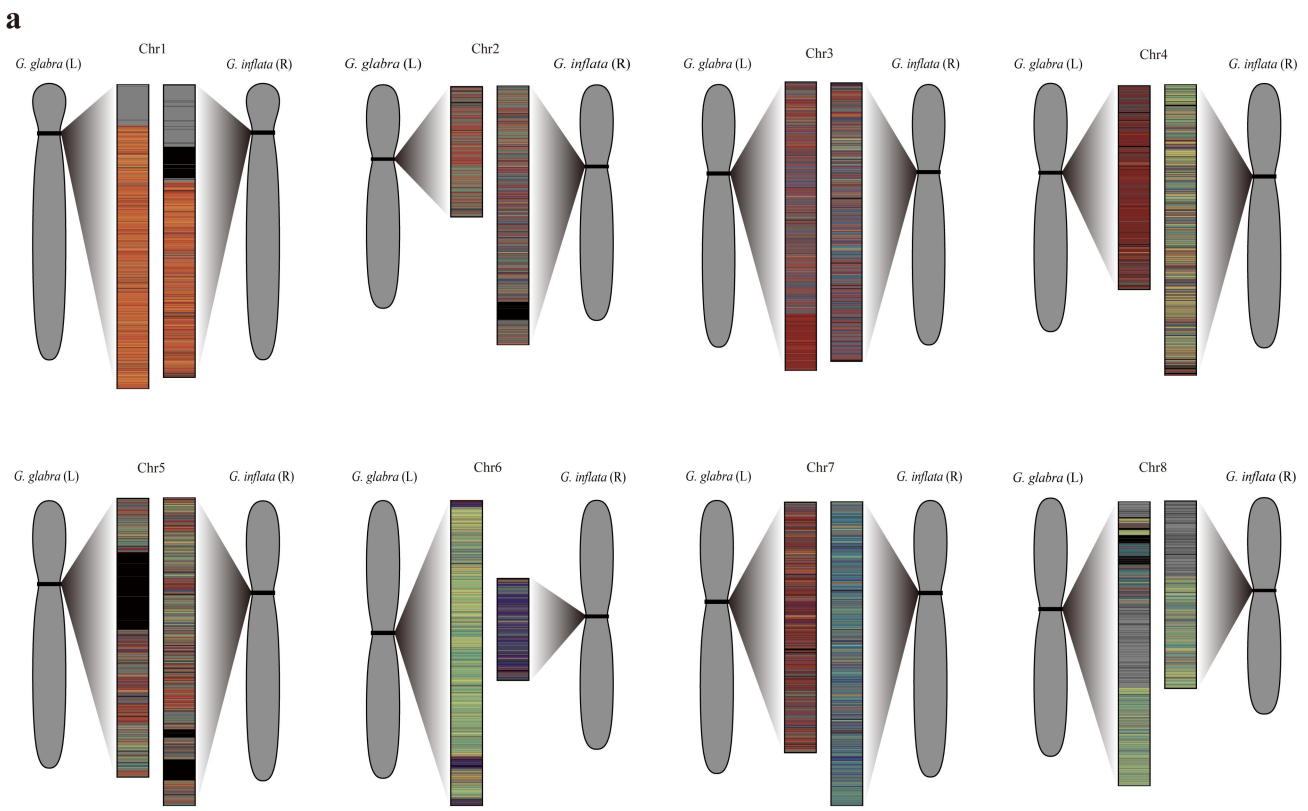

a
b
c
d
e
f
g
h
i
j
k
Chr1
% identity
Satellite Structure
Sequence Identity
Peak 1
Peak 2
Peak 3
Mean Rate
Occurrence Rate
Position (Mbp)
Component
Gly1
Gly2
Gly3
Gly4
Other Mon
Gly5
Gly6
Gly7
Gly8
Gly9
Sequence Identity (%)
Glycyrrhiza glabra Chr3
Exons
Satellite
Gypsy
Copia
Entropy/Density
Entropy/TE
Satellite(%)
Repeats(%)
Coordinates(Mb)
Glycyrrhiza glabra
CEN60
Entropy
GC
Entropy Rate
GC (%)
Density / Sequence %
TEL
CEN
Glygla.satellite
CEN60 (n=990,562)
DNA methylation level
CG
CHG
CHH
-0.03kb
Start
End
+0.03kb
High−Order Repeats Periodicity (Chr1)
Region: 12.01Mb − 12.013Mb (n=90)
Repeat Unit Index (Genomic Order)
Edit Dist.
Glycyrrhiza inflata Chr3
Glycyrrhiza inflata
Glyinf.satellite
CEN60 (n=916,198)
CEN60 Frequency
Position (bp)
ATGC

Figure 2


a
Chr1
G. glabra (L)
G. inflata (R)
Chr2
G. glabra (L)
G. inflata (R)
Chr3
G. glabra (L)
G. inflata (R)
Chr4
G. glabra (L)
G. inflata (R)
Chr5
G. glabra (L)
G. inflata (R)
Chr6
G. glabra (L)
G. inflata (R)
Chr7
G. glabra (L)
G. inflata (R)
Chr8
G. glabra (L)
G. inflata (R)

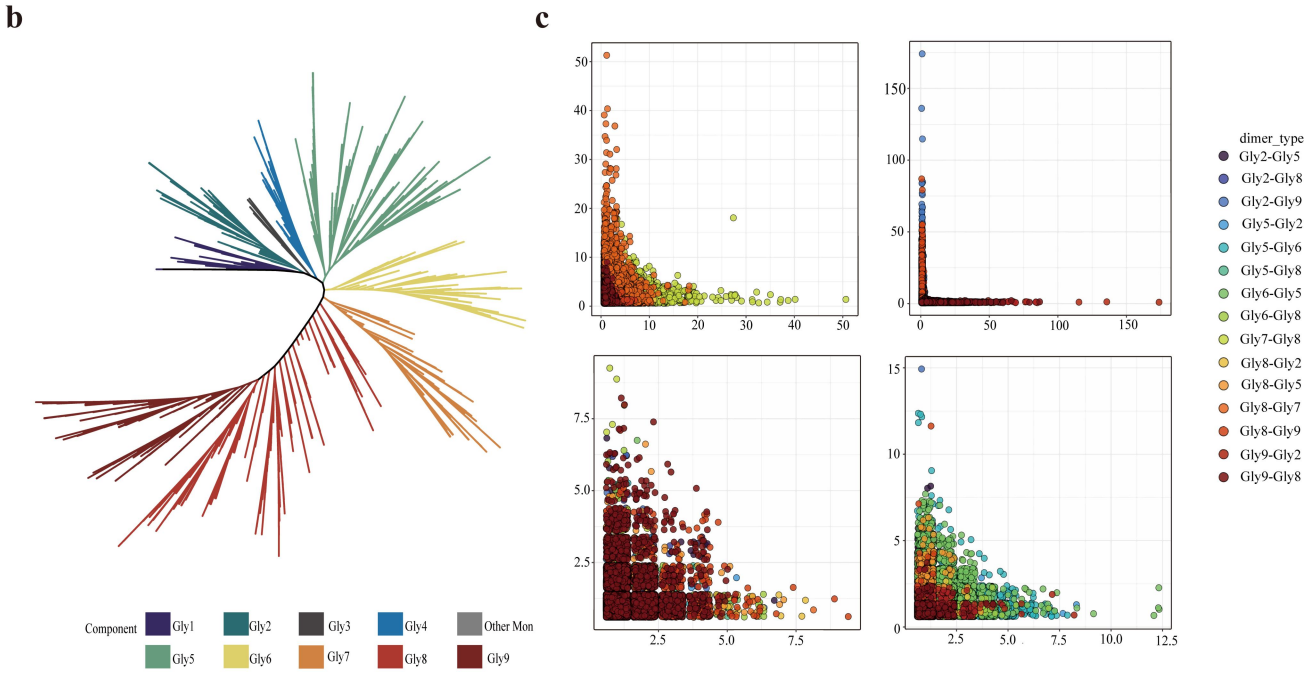

b
Component
Gly1
Gly2
Gly3
Gly4
Other Mon
Gly5
Gly6
Gly7
Gly8
Gly9
c
dimer_type
Gly2-Gly5
Gly2-Gly8
Gly2-Gly9
Gly5-Gly2
Gly5-Gly6
Gly5-Gly8
Gly6-Gly5
Gly6-Gly8
Gly7-Gly8
Gly8-Gly2
Gly8-Gly5
Gly8-Gly7
Gly8-Gly9
Gly9-Gly2
Gly9-Gly8

Figure 3

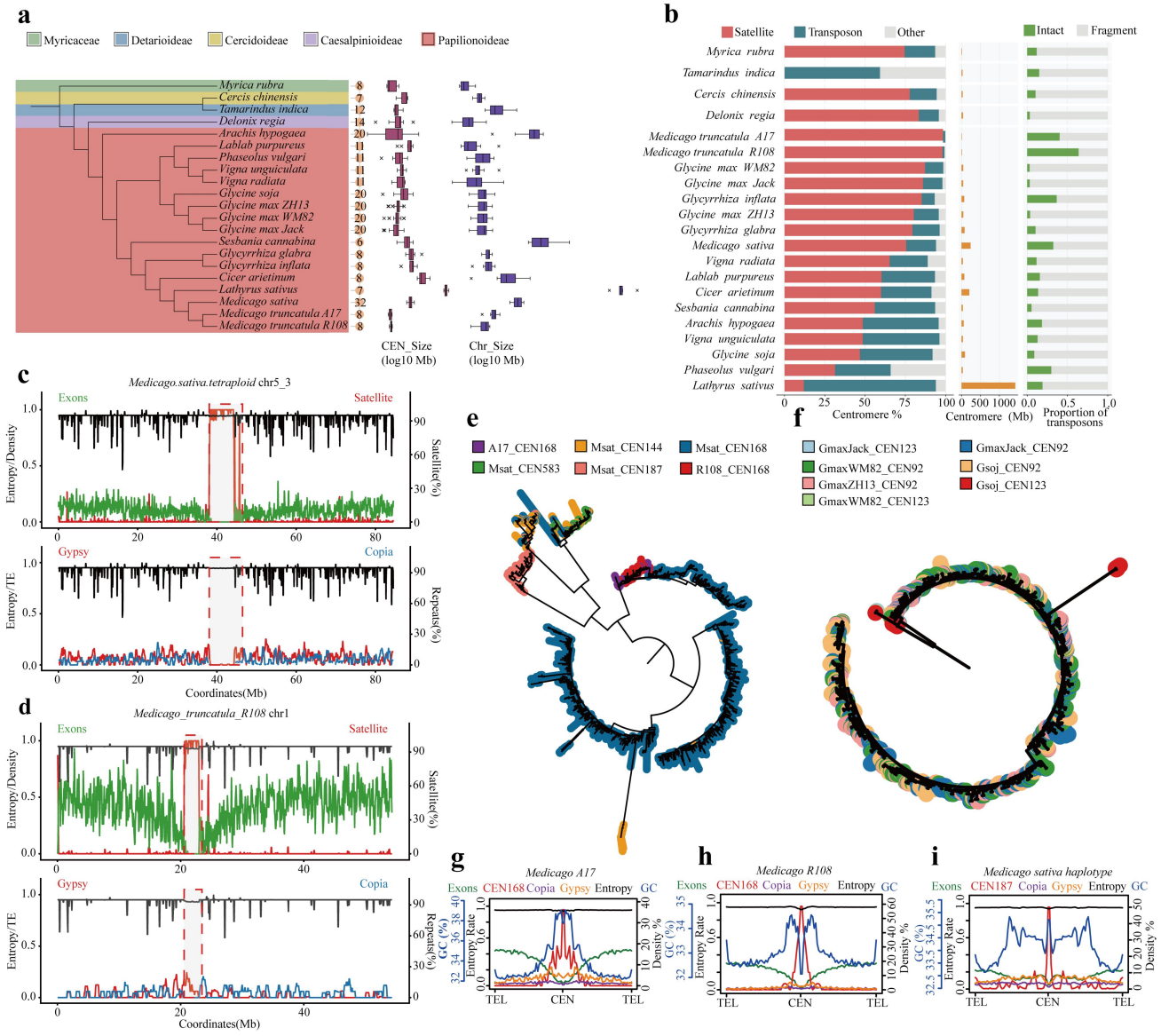

a
Myricaceae
Detarioideae
Cercidoideae
Caesalpinioideae
Papilionoideae
Myrica rubra
Cercis chinensis
Tamarindus indica
Delonix regia
Arachis hypogaea
Lablab purpureus
Phaseolus vulgari
Vigna unguiculata
Vigna radiata
Glycine soja
Glycine max ZH13
Glycine max WM82
Glycine max Jack
Sesbania cannabina
Glycyrrhiza glabra
Glycyrrhiza inflata
Cicer arietinum
Lathyrus sativus
Medicago sativa
Medicago truncatula A17
Medicago truncatula R108
CEN_Size (log10 Mb)
Chr_Size (log10 Mb)
b
Satellite
Transposon
Other
Intact
Fragment
Centromere %
Centromere (Mb)
Proportion of transposons
c
Medicago.sativa.tetraploid chr5_3
Exons
Satellite
Entropy/Density
Satellite(%)
Gypsy
Copia
Entropy/TE
Repeats(%)
Coordinates(Mb)
d
Medicago_truncatula_R108 chr1
e
A17_CEN168
Msat_CEN144
Msat_CEN168
Msat_CEN583
Msat_CEN187
R108_CEN168
f
GmaxJack_CEN123
GmaxJack_CEN92
GmaxWM82_CEN92
Gsoj_CEN92
GmaxZH13_CEN92
Gsoj_CEN123
GmaxWM82_CEN123
g
Medicago A17
Exons CEN168 Copia Gypsy Entropy GC
Entropy Rate
GC (%)
Density %
TEL
CEN
h
Medicago R108
i
Medicago sativa haplotype
Exons CEN187 Copia Gypsy Entropy GC

Figure 4

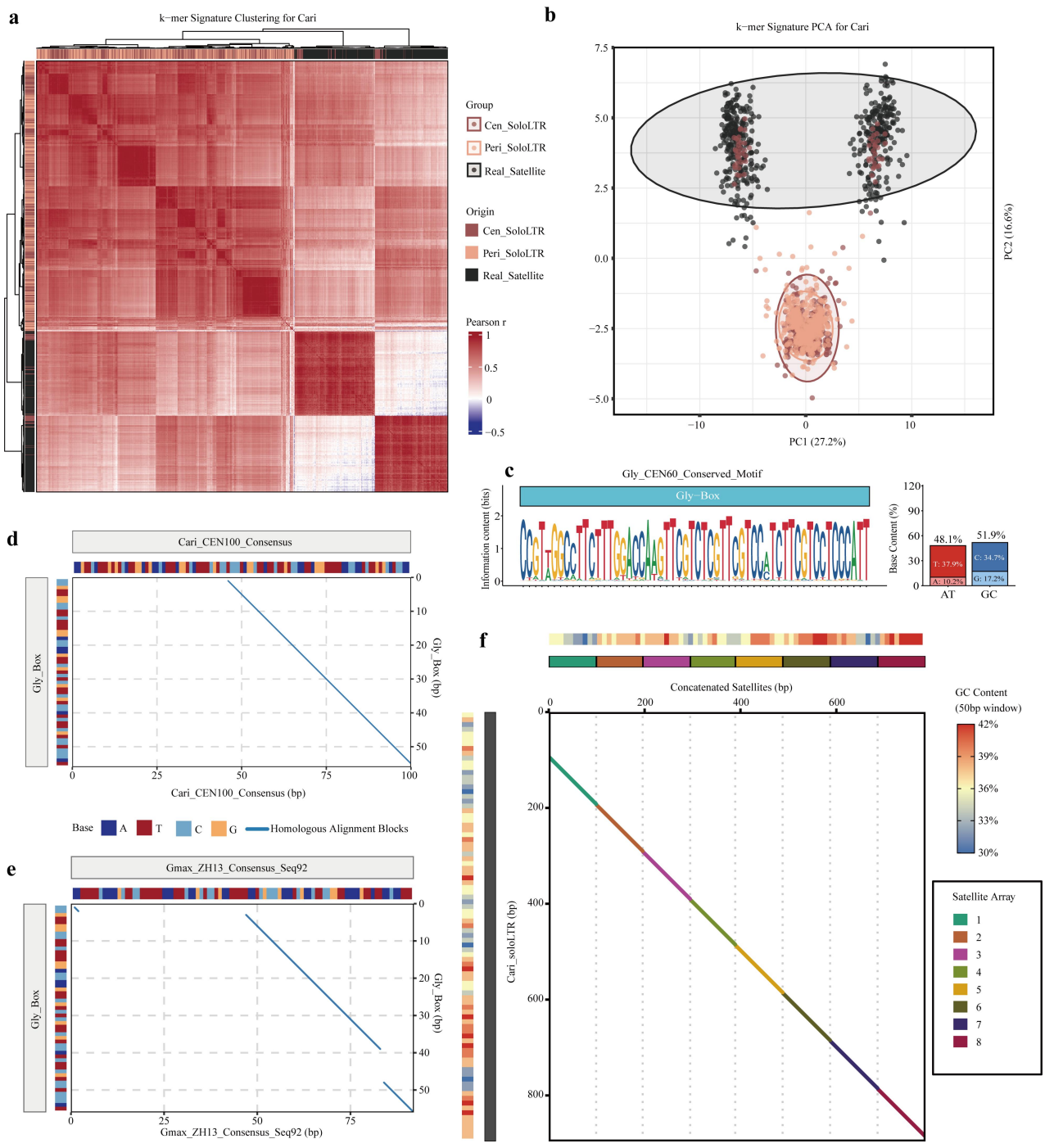
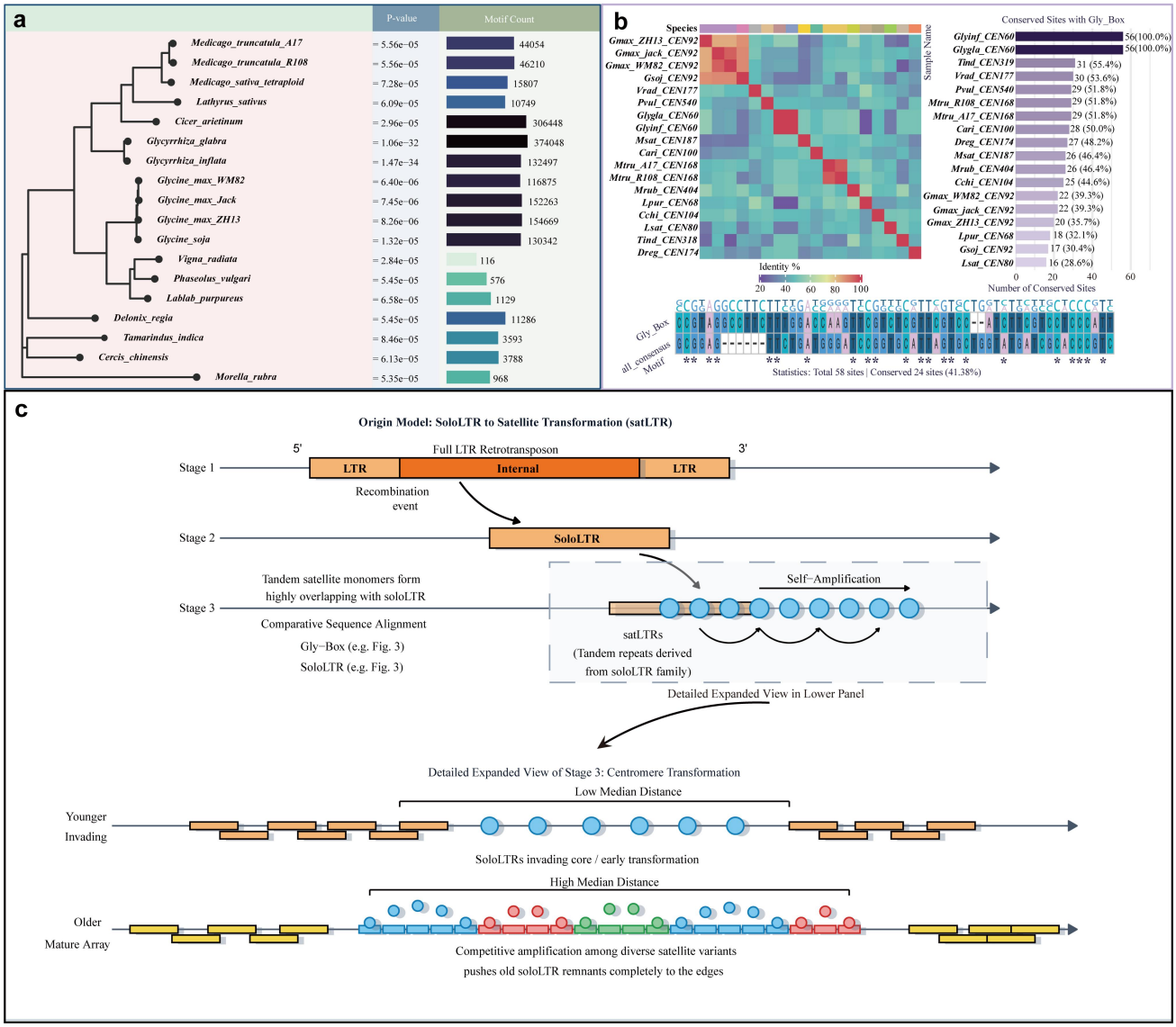

a
k−mer Signature Clustering for Cari
Group
Cen_SoloLTR
Peri_SoloLTR
Real_Satellite
Origin
Cen_SoloLTR
Peri_SoloLTR
Real_Satellite
Pearson r
1
0.5
0
−0.5
b
k−mer Signature PCA for Cari
PC1 (27.2%)
PC2 (16.6%)
c
Gly_CEN60_Conserved_Motif
Gly−Box
Information content (bits)
Base Content (%)
48.1%
51.9%
T: 37.9%
A: 10.2%
C: 34.7%
G: 17.2%
AT
GC
d
Cari_CEN100_Consensus
Gly_Box
Gly_Box (bp)
Cari_CEN100_Consensus (bp)
Base
A
T
C
G
Homologous Alignment Blocks
e
Gmax_ZH13_Consensus_Seq92
Gly_Box
Gly_Box (bp)
Gmax_ZH13_Consensus_Seq92 (bp)
f
Concatenated Satellites (bp)
Cari_soloLTR (bp)
GC Content
(50bp window)
42%
39%
36%
33%
30%
Satellite Array
1
2
3
4
5
6
7
8

Figure 5


a
P-value
Motif Count
Medicago_truncatula_A17
Medicago_truncatula_R108
Medicago_sativa_tetraploid
Lathyrus_sativus
Cicer_arietinum
Glycyrrhiza_glabra
Glycyrrhiza_inflata
Glycine_max_WM82
Glycine_max_Jack
Glycine_max_ZH13
Glycine_soja
Vigna_radiata
Phaseolus_vulgari
Lablab_purpureus
Delonix_regia
Tamarindus_indica
Cercis_chinensis
Morella_rubra
= 5.56e−05
= 5.56e−05
= 7.28e−05
= 6.09e−05
= 2.96e−05
= 1.06e−32
= 1.47e−34
= 6.40e−06
= 7.45e−06
= 8.26e−06
= 1.32e−05
= 2.84e−05
= 5.45e−05
= 6.58e−05
= 5.45e−05
= 8.46e−05
= 6.13e−05
= 5.35e−05
44054
46210
15807
10749
306448
374048
132497
116875
152263
154669
130342
116
576
1129
11286
3593
3788
968
b
Species
Sample Name
Conserved Sites with Gly_Box
Identity %
Number of Conserved Sites
Gly_Box
all_consensus
Motif
Statistics: Total 58 sites | Conserved 24 sites (41.38%)
c
Origin Model: SoloLTR to Satellite Transformation (satLTR)
5'
3'
Full LTR Retrotransposon
Stage 1
LTR
Internal
LTR
Recombination event
Stage 2
SoloLTR
Tandem satellite monomers form highly overlapping with soloLTR
Stage 3
Comparative Sequence Alignment
Gly−Box (e.g. Fig. 3)
SoloLTR (e.g. Fig. 3)
Self−Amplification
satLTRs
(Tandem repeats derived from soloLTR family)
Detailed Expanded View in Lower Panel
Detailed Expanded View of Stage 3: Centromere Transformation
Low Median Distance
Younger
Invading
SoloLTRs invading core / early transformation
High Median Distance
Older
Mature Array
Competitive amplification among diverse satellite variants
pushes old soloLTR remnants completely to the edges

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable.xlsx
- SupplementaryFigure.pdf