Genomics 118 (2026) 111210

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Research Paper

Haplotype-resolved genome assembly of *Musella lasiocarpa* reveals the critical role of structural variations in chromosomal and genome evolution

Qing Liu [a,b,c,*,1], Dongli Cui [a,b,d,1], Yaqi Tian [a,b,d], Yehan Wang [a,b,d], Mathieu Rouard [e], John Seymour Heslop-Harrison [a,b,f,*], Trude Schwarzacher [a,b,f,*], Ziwei Wang [a,b,*,2]

[a] *Guangdong Provincial Key Laboratory of Applied Botany / Key Laboratory of National Forestry and Grassland Administration on Plant Conservation and Utilization in Southern China, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China*
[b] *South China National Botanical Garden, Guangzhou 510650, China*
[c] *Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China*
[d] *College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China*
[e] *Biodiversity International, Parc Scientifique Agropolis II, 34397 Montpellier Cedex 5, France*
[f] *University of Leicester, Institute for Environmental Futures, Department of Genetics and Genome Biology, Leicester LE1 7RH, UK*

ARTICLE INFO

Keywords:
Musella lasiocarpa
Telomere-to-telomere genome assembly
Centromere
Repetitive DNA
Haplotypes
Structural variation
Cold adaptation

ABSTRACT

Musella lasiocarpa (MLA, 2*n* = 18, Musaceae) is an endangered species native to south-western China. We assembled its haplotype-resolved, telomere-to-telomere genomes with a genome size of 503.6 Mb consisting of 52.8% repetitive DNA. A 134 bp tandem repeat, Mlcen, was identified at all centromeres, and telomere sequences were present at 30 of 36 assembled pseudo-chromosome ends. The distal gene-rich regions display high synteny, whereas retrotransposon polymorphisms between haplotypes occurred throughout chromosomes, contributing to diversity. Phylogenetic analysis shows MLA diverged from *Ensete* 42 million years ago, and together they share a common ancestor with *Musa*. Among 35,312 protein-coding genes, 14 up-regulated and 34 down-regulated transcription factors were identified under cold treatment. This high-quality genomic resource advances our understanding of MLA chromosomal evolution characterized by structural variations, repetitive DNA dynamics, and cold-responsive genes at both haplotype and species levels; and enables genome-assisted improvement of more resilient crops such as bananas and *Ensete*.

1. Introduction

Musella lasiocarpa (Franch.) C.Y. Wu ex H.W. Li (MLA, 2*n* = 2*x* = 18, family Musaceae), is an endangered species [1] endemic to Yunnan and Sichuan provinces of China (Fig. 1A-D). Commonly known as Dì Yǒng Jīn Lián, MLA is valued both as an ornamental plant and a feed crop. It has a feature of striking bracts, vibrant yellow in var. *lasiocarpa* and orange-red in var. *rubribracteata* (Fig. 1B-D). In addition, its pseudo-stems are used as a starch source for animals in Yunnan [2]. In the family Musaceae (order Zingiberales), MLA and its sister group *Ensete* (*x* = 9) are closely related to bananas (*Musa spp.*; *x* = 7, 9, 10, and 11) [3]. MLA can be distinguished from *Ensete* and *Musa* by its erect, compact rosette-shaped inflorescences, and by chloroplast genome analyses [4]. Notably, the species demonstrates higher cold tolerance than other members of the Musaceae; it can withstand the severe winter conditions prevalent in Sichuan [5].

More than ten genomes of species within Musaceae [6] have been assembled using long-read sequencing technologies (Oxford Nanopore Technologies, ONT, and Pacific Biosciences, PacBio), and with Hi-C chromatin conformation capture. Chromosome-scale assemblies are available for *Musa acuminata* [7], *M. schizocarpa* [8], *M. balbisiana* [9], *Ensete ventricosum* (GCA_029747655.1), *M. textilis* [10], *E. glaucum* [11], *M. beccarii* [12], diploid *M. acuminata* [13,14], and *M. velutina* and *M. ornata* [15] and in comparative studies of wild and cultivated *Musa*

* Corresponding authors at: Guangdong Provincial Key Laboratory of Applied Botany / Key Laboratory of National Forestry and Grassland Administration on Plant Conservation and Utilization in Southern China, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China.
E-mail addresses: liuqing@scib.ac.cn (Q. Liu), phh4@le.ac.uk (J.S. Heslop-Harrison), ts32@le.ac.uk (T. Schwarzacher), wangziwei@sgu.edu.cn (Z. Wang).
[1] These authors contributed equally to this work.
[2] Present address: School of Biology and Agriculture, Shaoguan University, Shaoguan 512005, China

https://doi.org/10.1016/j.ygeno.2026.111210
Received 29 October 2025; Received in revised form 17 December 2025; Accepted 26 January 2026
Available online 27 January 2026

species [16]. The haploid genome sizes of these species range from 428 Mb [15] to 603 Mb [17]. A recent publication by Zhao et al. [18] reported the genome of *Musella lasiocarpa* but it does not have an in-depth analysis of repetitive sequences, haplotype resolved assemblies and structural variations (SVs) as we report. Our genome assembly, annotation and raw data for *M. lasiocarpa* are public on NCBI and the Banana Genome Hub (https://banana-genome-hub.southgreen.fr/) with tools for comparative analysis in the Musaceae [6].

Haplotype-resolved, telomere-to-telomere genome assemblies provide valuable insights into chromosome structural variations and allelic differences within a species. These have been generated from *Musa acuminata* accessions (a diploid; [14]), the triploid banana cultivar AAA 'Baxijiao' from the Cavendish cultivar group [13], two additional AAA genome cultivars from the Cavendish and Gros Michel cultivar groups [19], and two AAB genome triploid hybrids representing the Plantain and Silk cultivar groups [20]. The haplotypes reveal the chromosomal contributions of ancestral species to the genomes of cultivated bananas within the genus *Musa* [16]. Repetitive DNAs constitute a dynamic and rapidly evolving component that differentiates haplotypes within the same species. Analyses of chromosomal rearrangements provide insight into the complex genomic organization of repetitive sequences, extending beyond the implications of chromosome number variation (*x*

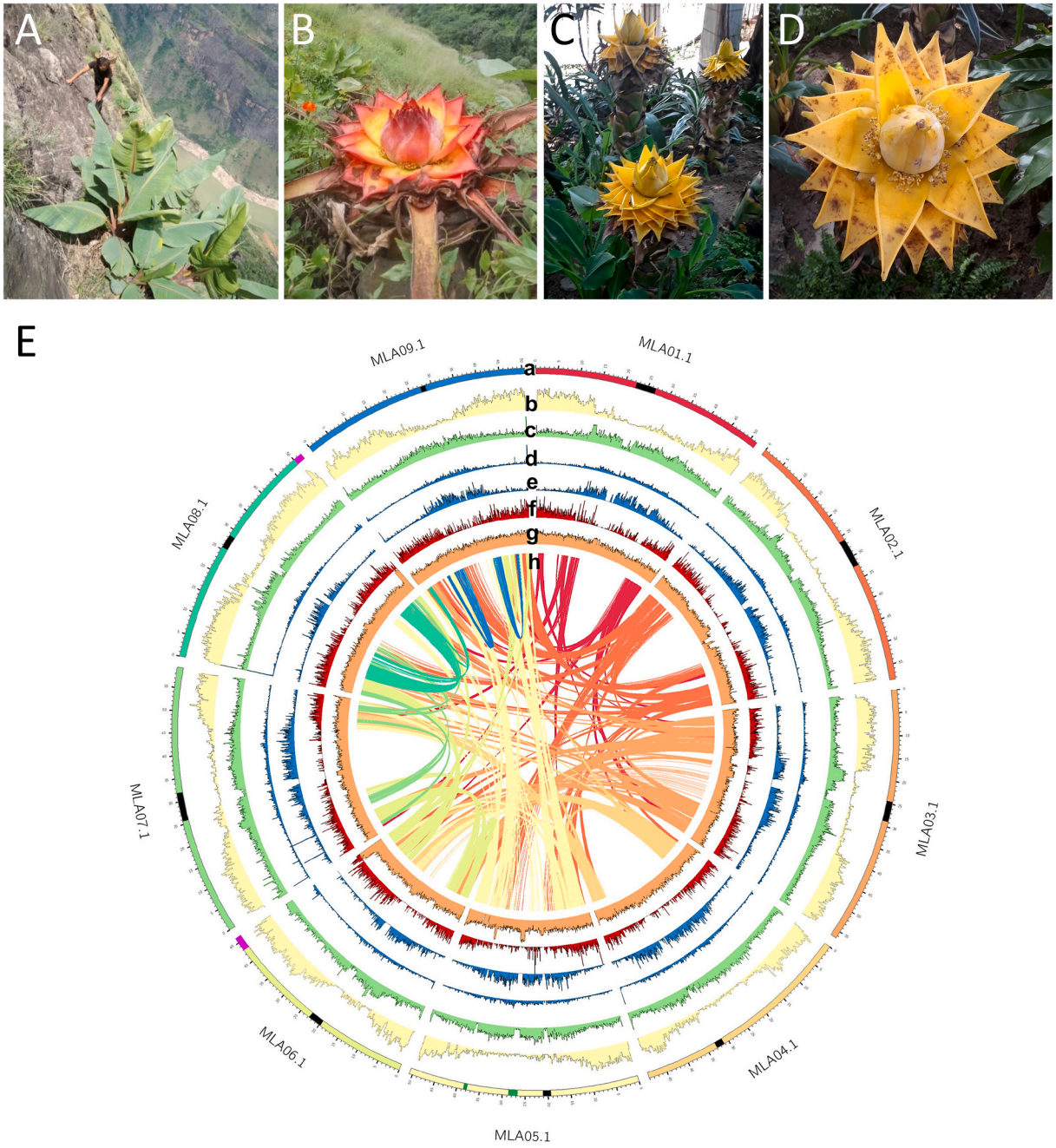


Fig. 1. Plant morphology, habitat and genomic features of *Musella lasiocarpa* (MLA). **A**: Cliff habitat (Nanhua County, Chuxiong Yi Autonomous Prefecture, Yunnan, China). **B**: *M. lasiocarpa* var. *rubibracteata* in the wild. **C-D**: *M. lasiocarpa* var. *lasiocarpa* in the greenhouse of South China National Botanical Garden. **E**: MLAh1 (chromosomes designated haplotype 1) genomic features. (**a**) Nine chromosomes (scale: 5.0 Mb) with green, pink, and black boxes representing 5S (MLA05.1), 45S (MLA06.1 and MLA08.1) rDNA and centromere positions; (**b**) Gene density; (**c**) Repeat density; (**d**) *Copia* density; (**e**) *Gypsy* density; (**f**) DNA transposon density; (**g**) Simple repeat density; (**h**) Syntenic genomic blocks, connected with curves. b–h: 10,000 bp bins. For MLAh2 genomic features see Fig. S1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

= 9 or 11) observed among three genera of Musaceae [6,21,22]. Haplotype assemblies are able to elucidate the relationships between repetitive components in species such as, for example, lemon [23], melon [24], and jujube [25]. Moreover, it is also essential to investigate structural variations, including retroelement insertion polymorphisms and large chromosomal rearrangements (inversions, translocations, duplications, and deletions).

Cold stress is a key factor restricting the geographical distribution and productivity of Musaceae species [26,27]. Among them, *Musella lasiocarpa* exhibits remarkable tolerance to low temperatures. By analyzing the transcriptome data, the candidate genes and regulatory pathways associated with cold tolerance in MLA can be identified. Overall, our study presents the first haplotype-resolved, near-telomere-to-telomere genome assembly including intensive manual curation (as suggested by [28]) of *M. lasiocarpa*, characterized by structural variations, repetitive DNA dynamics, and cold-responsive genes. The high-quality genome assembly serves as a valuable genomic resource for breeding programs aimed at developing resilient crops, including bananas and *Ensete* species.

2. Results

2.1. De novo haplotype-resolved genome assembly

We generated a *de novo* haplotype resolved, chromosome-level genome assembly of *Musella lasiocarpa* var. *lasiocarpa* (MLA). Using long-single-molecule PacBio HiFi and ONT reads with 62 × to 204 × sequencing depth, along with Hi-C and Illumina data (Table S1), we assembled nine pseudo-chromosomes for each haplotype (Figs. 1E and S1), yielding genome sizes of 500.05 Mb for MLAh1 and 498.37 Mb for MLAh2 with individual chromosomes ranging from 40.91 Mb to 59.46 Mb in length (Tables 1 and S2A). Chromosome identity and orientation were determined based on their homology to the *Ensete glaucum* (EGL) genome assembly ([11]; itself based on homology to *Musa*). Because the parental origin of each chromosome is unknown, the pseudo-chromosomes with fewer contigs and higher N50 values were designed as MLAh1, which was used as the reference genome for subsequent analyses (Tables 1 and S2A).

The haploid genome size of MLA was initially estimated to be 436 Mb using the 21-mer distribution visualized by GenomeScope v.2.0 [29], also showing a heterozygosity of 1.24% (Fig. S2A, Table S2B). Based on sequence coverage distribution and mapped nucleotide (backmap.pl v.0.5; [30]) results, the genome size was refined to 503.58 Mb (Table S2B), which is slightly smaller than the 535 Mb reported by Zhao et al. [18].

2.2. Quality assessment of genome assembly

Our haplotype-resolved genome assemblies captured 99% of the MLA genome (Tables 1 and S2), organized into nine pairs of gapless pseudo-chromosomes. Genome completeness and the predicted protein sequences were assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO) [31], yielding scores greater than 98% (Table S3), which were higher than those reported for other related monocotyledonous assemblies (Fig. S2B). In the Hi-C interaction heatmap, signal intensity along the diagonal, representing interactions between neighboring sequences, was stronger than that of non-diagonal positions (Fig. S3), indicating that the haploid assemblies are complete, and no structural variations were caused by assembly algorithms. Telomeric repeats (TTTAGGG)n were detected at 30 of the 36 chromosome ends (Table S4). A 134 bp centromeric repeat (Mlcen) was detected at all centromere positions (Figs. 1E and S1; Table S5) and showed high sequence homology to Egcen from *E. glaucum* [11]. Additionally, 5S and 45S rDNA arrays were localized on chromosomes MLA05, MLA06, and MLA08 (Table S6). These results collectively provide high-quality, haplotype-resolved, telomere-to-telomere assemblies of the MLA genome.

2.3. Gene annotation and expression analysis

2.3.1. Gene prediction and functional annotation and ontology

The protein-coding genes (PCGs) of the MLA haplotypes were annotated using an integrated strategy that combined RNA-seq-assisted and *ab initio* prediction with homology-based annotation. In total, 35,387 and 35,237 PCGs were predicted in MLAh1 and MLAh2, and about 94% of these genes were functionally annotated using seven major protein databases (NCBI NR, [32]; InterPro, [33]; GO, [34]; KOG, [35]; KEGG, [36]; TrEMBL and SwissProt, [37]) (Tables 2 and S3C). The number of annotated genes represents roughly a 3% increase compared to the 34,391 genes reported by Zhao et al. [18] in a previous assembly and annotation. Predicted gene density was lower near centromeric regions and gradually increased toward the chromosome ends (Figs. 1E and S1).

Table 1
Summary of genome assembly statistics and functional annotation of MLA haplotypes.

Genome assembly	MLAh1 (Reference)	MLAh2
Estimated genome size by backmap [a]	503.58 Mb	
Assembly size	500,050,725 bp	498,373,817 bp
Percentage of genome assembled	99.30%	98.97%
GC ratio	40%	41%
Total number of contigs	365	295
Contig N50[b]	49,080,785 bp	18,553,585 bp
Total number of scaffolds	357	274
Scaffold N50	51,603,189 bp	53,045,455 bp
Anchored into pseudo-chromosomes	474,465,207 bp	462,245,547 bp
Number of pseudo-chromosomes	9	9
Quality of genome assembly		
Hi-C mounting rate	94.88%	92.75%
BUSCO completeness of assembly	98.51%	98.82%
Functional annotation		
Number of predicted genes[c]	35,387	35,237
BUSCO completeness of annotation	97.96%	97.83%
Number and proportion (%) of functionally annotated genes	33,232 (93.91%)	33,172 (94.14%)

[a] MLAh1 genome assembly analyzed by backmap pl v.0.5 [30] see Table S2.
[b] Longer N50 assigned to haplotype h1.
[c] Gene number predicted by *ab initio*, transcriptome, and homolog-based structure prediction.

Table 2
Summary of *Musella lasiocarpa* genome assembly—gene prediction and repetitive DNA proportions.

a. Protein-coding genes (PCGs)	
Number of PCGs	35,387 (MLAh1) 35,237 (MLAh2)
Average length per gene	4498.07 bp
Average number of exons per gene	5.20
Average length of exon per gene	237.02 bp
Average length of intron per gene	777.77 bp
Average length of proteins coded	410.75 aa

b. Repetitive DNA proportion by RepeatMasker (Average of MLAh1 and MLAh2)	
Total	52.81%
Class I retroelements	34.85%
LTR retroelements	34.40%
Copia	14.52%
Gypsy	11.92%
Unknown or mixture	7.87%
Non-LTR LINES	0.46%
Class II DNA transposons	9.01%
MITEs	3.69%
Unknown interspersed repeats	3.98%
Simple repeats and low complexity	1.29%

c. Tandem repeat proportion in Illumina reads	
Total	5.68%
45S rDNA	2.58%
5S rDNA	0.27%
Centromeric sequence Mlcen	2.33%
Microsatellites (<8 bp motif)[a]	0.50%

[a] Data summary from EDTA [84].

Gene family analysis of Musaceae species (*Musa acuminata*, *M. textilis*, *M. beccarii*, *Musella lasiocarpa*, *Ensete ventricosum*, and *E. glaucum*) clustered 195,076 genes (89.4%) of the total 218,084 into 31,235 ortho-groups (Table S7). The remaining 23,008 genes (10.6%) were unassigned, representing putative singletons with no detectable orthologs in the other species (Table S7). Among the six species, 18,453 ortho-groups (59.1%) identified by UpSetR [38] were shared, corresponding to 22,670 genes (72.6%) that constitute the core or pseudo-core gene sets (Fig. 2A), assuming that orthogroups absent in only one species are likely due to annotation artefacts, while 212 orthogroups (0.7%) were specific to *Musella*, with no homologs identified in *Musa* or *Ensete* (Fig. 2A). Synonymous substitution rate (Ks) analysis showed a peak at ~0.53 in both intra and interspecific comparisons involving MLA, *E. glaucum*, and *M. acuminata* (Figs. 2B and S4), supporting the occurrence of the α/β whole genome duplication (WGD) events during the evolution of the Musaceae genome. Additionally, a Ks peak around 0.75 for the MLA-*Zingiber officinale* comparison [39] supports the γ WGD event associated with the Zingiberales lineage.

2.3.2. Gene family expansion and contraction

Protein sequences from MLA and 13 other monocotyledonous species were clustered into 33,044 gene families. In the MLA genome, 16,595 orthologous genes were identified — fewer than in the eight other Musaceae species but greater than in the other Zingiberales species (Table S8). Among these, 994 single-copy orthologs were shared across 14 species while 155 orthologs were unique to MLA (Fig. S5A, left; Table S9). Among 154 gene families (consisting of 1,618 genes) showing significant expansion, and 162 (with 346 genes) showing significant contraction (Fig. 2C; Table S8B, C) in MLA, the enrichment analysis of the expanded gene families indicated 55 genes with the GO term "photosynthesis/light reaction" (Fig. S5A, right; Table S10). Phylogenetic reconstruction and divergence time estimation based on the shared single-copy orthologs indicated that *Musella* and *Ensete* form sister taxa that diverged approximately 42.0 million years ago (Mya), and that both lineages diverged from *Musa* around 57.0 Mya (Fig. 2C).

2.3.3. Differential gene expression (DEG) under cold treatment

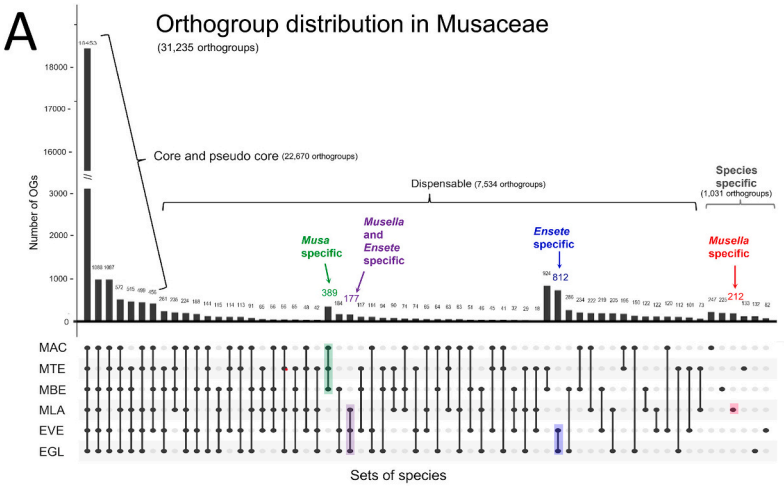
When subject to a 10 °C temperature drop from 25 °C (normal conditions) to 15 °C (cold treatment) for 48 h, *M. lasiocarpa* var. *lasiocarpa* (yellow bracts) and var. *rubribracteata* (red bracts) showed significant transcriptomic responses. Approximately 10% of genes were differentially expressed, with 3151 differentially expressed genes (DEGs) in yellow bracts and 3,726 in red bracts, showing at least a two-fold change in expression ($P < 0.05$) between normal- and cold-treated samples (Fig. 2D, upper left). Of these, 1,470 genes were down-regulated in both varieties, whereas only 88 genes were up-regulated (Fig. 2D, upper right; Table S11). KEGG pathway and GO enrichment analyses of DEGs revealed that the up-regulated genes were enriched in pathways related to environmental adaptation and DNA repair and recombination (Fig. 2D, bottom left; Table S12A). Conversely, the down-regulated genes were predominantly involved in fundamental cellular processes, transcription, metabolism and particularly photosynthesis (Fig. 2D, bottom right; Table S12B). Notably, genes associated with environmental adaptation showed significant expansion in *M. lasiocarpa* (Fig. S5A, right; Table S10).

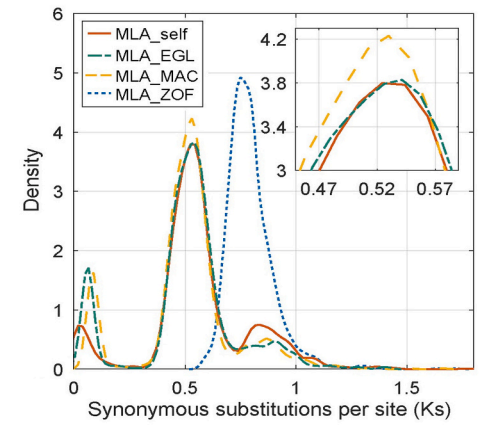
Among 1,470 down-regulated genes and 88 up-regulated genes in both MLA varieties, the 48 differentially expressed transcription factors (DETFs) were identified. Of these, 14 DETFs were up-regulated and 34 were down-regulated, all of which are involved in cold acclimation (Fig. 2E, right; Table S13). Previous studies [40,41] reported that those 14 up-regulated genes belong to 10 transcription factor (TF) families, while the 34 down-regulated genes belong to 13 TF families (Table S13). These TFs are associated with the major cold stress-responsive pathways previously characterized in grasses and Arabidopsis as well as *Musa* [26,27]. The DEGs are distributed across all chromosomes except MLA01, with notable clustering on MLA02, MLA05, and MLA06. However, they are not linked to any structural variations identified in the MLA genome (Fig. S6).

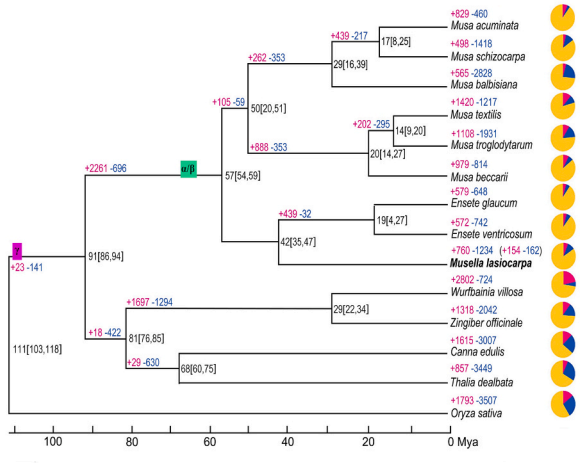
2.4. Synteny and structural variations between haplotypes

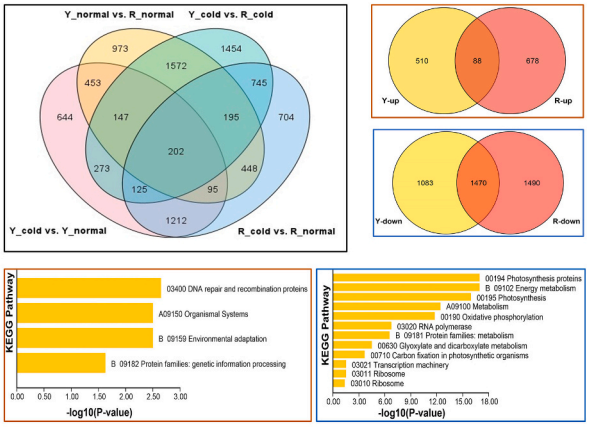
Pairwise whole-genome comparisons revealed largely conserved synteny between the two haplotypes. Structural variations including inversions, translocations, and duplications, were abundant in the polymorphic proximal pericentromeric regions (Fig. 3; Table S14), and some large SVs were present in gene-rich regions of chromosome arms. Chromosomes MLA01, MLA03, and MLA05 exhibited the highest proportion of rearrangements, whereas MLA02, MLA07, and MLA09 maintained the greatest synteny. Synteny was most pronounced at the chromosome ends as evidenced by the SyRI collinearity plots [42] by dense, parallel grey lines in the distal regions (Fig. 3) that are gene-rich (Fig. 1E). In contrast, synteny was increasingly disrupted near the centromeres, due to accumulation of repetitive sequences causing SVs. Furthermore, the 45S rDNA arrays on MLA06 and MLA08 displayed substantial differences between the two haplotypes (Fig. 3).

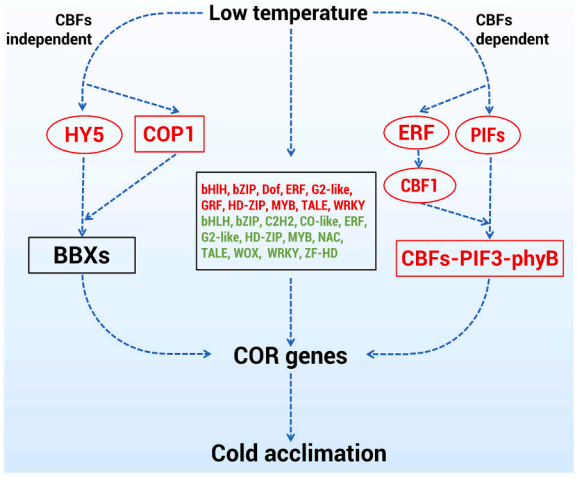
Detailed chromosome sequence dotplots together with SyRI analysis (Fig. S7) showed that each chromosome contained between two to seven regions, ranging 0.1 to 2.2 Mb, that harbored major structural rearrangements, such as inversions, translocations, and duplications including tandem and mixed repetitive sequences (Table S14). Interestingly, some SVs included genes or coding regions (*e.g.*, Ml01i, Ml04ii, Ml05iii, Ml09i; Fig. S7A, D, E, J), while others either lacked genes entirely (Ml06ii; Fig. S7F), or contained gaps of uncharacterized sequence (Ml02ii; Fig. S7B). Several tandem arrays contributed to the polymorphisms observed between haplotypes, and many of the arrays were rich in transposable elements (*e.g.*, Ml02iii, Ml03ii, Ml06iii, Ml08iii; Fig. S7B, C, F, I). For instance, a 600 kb insertion (Ml09i; Fig. S7J) consisted of eight copies of a 12 kb monomer, which included a gene (polygalacturonase), fragments of chloroplast DNA, and some

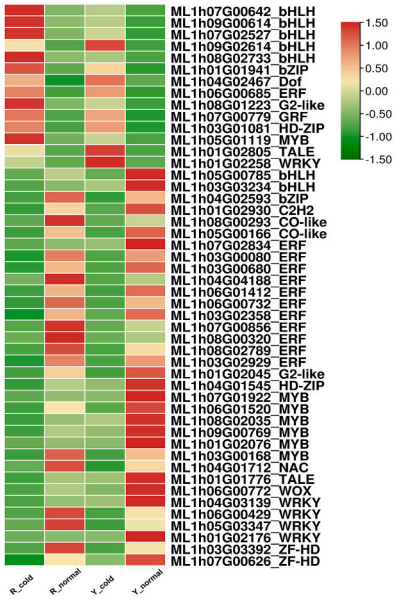












(caption on next page)

Fig. 2. Comparative analysis of gene families between *Musella lasiocarpa* (MLA), five Musaceae species, and other monocotyledonous species, and identification of differentially expressed genes (DEGs) of MLA varieties under cold treatment. **A**: UpSetR diagram of shared orthogroups in MLA and five Musaceae species (MAC: *Musa acuminata*; MTE: *M. textilis*; MBE: *M. beccarii;* EVE: *Ensete ventricosum*; EGL: *E. glaucum*). The shared orthogroups (OGs; ≥ two sequences / OG) are listed for each species set. **B**: Synonymous substitutions per site (Ks) plot showing whole genome duplication (WGD) of MLA_self (MLAh1) and divergence events between MLA-EGL and MLA-MAC (~0.53 inset), and the divergence of MLA-ZOF (~0.75; ZOF, *Zingiber officinalis*). **C**: Phylogenetic relationships and timescale of 14 monocotyledonous species inferred on the basis of 994 single-copy orthologous groups by PAML using *Oryza sativa* as outgroup species. Black numbers after branch nodes represent divergence times (Mya, million years ago) with confidence intervals of each node in square brackets, and pink (+) and blue (−) numbers near branches and species names represent expansion (pink) and contraction (blue) orthologous gene families. The green and pink boxes indicate α/β and γ WGD events, respectively. Pie diagrams on the right show the proportion of gene families undergoing expansion (pink), contraction (blue) or unchanged (yellow) (Database in Table S8D). **D**: Venn diagrams and KEGG enrichment of DEGs of MLA varieties under cold treatment. **Top left** Venn diagrams of 3,151 DEGs in var. *lasiocarpa* (Y) and 3,726 DEGs in var. *rubibracteata* (R). **Top right** Venn diagrams (Table S11) of the 88 up-regulated (Upper orange square) and 1,470 down-regulated (Bottom blue square) DEGs shared by MLA varieties. **Bottom left** (Table S12A) shows KEGG enrichment analysis of up-regulated DEG proteins shared by MLA varieties. **Bottom right** (Table S12B) shows KEGG enrichment analysis of down-regulated DEG proteins shared by MLA varieties. **E**: Transcription factor (TF) identification under cold treatment. **Left**: Model of TF function under cold treatment, with the black boxes denoting downstream genes inferred from literature [40,41]. HY5, COP1, and PIF genes were determined based on KEGG enrichment analysis. The central box contains identified TFs, with ovals representing TFs and squares representing non-TF genes. Red words represent up-regulated, and green words represent down-regulated genes. Dashed arrows represent the regulatory pathways await to be proven. Abbreviations: CBFs, C-repeat binding factors; COP1, constitutive photomorphogenic1 protein, HY5, elongated hypocotyl 5; BBXs, B-box domain protein; ERF, ethylene response factors; PIFs, phytochrome-interacting factors; phyB, phytochrome B; COR, cold-responsive genes. **Right**: Expression levels of 14 up-regulated and 34 down-regulated DEGs in MLA varieties under cold treatment (Table S13). The relative expression levels were generated by TBtools [76] with row scale normalization. Green bars to −1.5 represent the relative gene expression level below the average, red bars to 1.5 represent gene expression level above the average. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

retroelement domains (APR, aspartic protease), along with additional other retroelement fragments located outside the repetitive monomer motif (Figs. 4A and S7J). The transfer of organellar genes to the nucleus has been well documented, and the presence of their fragments within these repeat monomers is noteworthy [43].

Prominent inversions (2–5 Mb) were detected outside the centromeric regions of MLA03, MLA04, and MLA08 (Fig. S7C, D, I). ONT reads spanning the inversion breakpoints confirmed these real SVs were not caused by assembly artefacts (Fig. S8). Analysis of breakpoints using dotplots revealed small inverted repeats at the sites, indicative of LTR retroelement activities, implying that haplotype diversity in *Musella* is primarily driven by LTR retrotransposon amplification and insertion (Fig. S8Aii-iv, Bii-iv, Cii-iv).

2.4.1. Haplotype SNPs, short InDels, and heterozygosity

Analysis of single nucleotide polymorphisms (SNPs) identified 3,352,140 variants, accounting for 0.67% of the genome with 563,104 insertions and deletions (InDels <50 bp) within syntenic blocks (Fig. 4B, top). Most SNPs and InDels were located in intergenic regions (43.82%), followed by upstream and downstream regulatory regions and introns (Fig. 4B, bottom). Only a small fraction (4.81%) occurred in exonic regions, frameshift mutations or alterations resulting in gain or loss of stop codons were rare (0.70–1.26%).

2.5. Repeat analysis

In MLA, 263,543,055 bp (52.81%) of repetitive sequences were identified by RepeatMasker (Tables 2 and S15), consistent with the RepeatExplorer2 result (51.08%; Fig. S9B) and similar to other banana genomes (52.62% in *M. acuminata*, [7]; 55.02% in *E. glaucum*, [11]). The most abundant class of repeats were long terminal repeat retroelements (LTRs), which accounted for 34.40% of the MLA haplotypes (Table 2). In addition, multiple short tandem arrays of various repeat monomers, including microsatellites (simple sequence repeats, SSRs), were detected in both *Musa* and *Ensete* assemblies using RepeatExplorer2 [44].

2.5.1. Retrotransposons

The most abundant classes of retrotransposons were *Copia* (14.5%) and *Gypsy* (11.9%) (Tables 2 and S15), with proportions slightly lower than those reported for *E. glaucum* [11]. In *Musa acuminata*, *Copia* elements were notably more prevalent (29%) than *Gypsy* elements, which accounted for 11% and 19% in D'Hont et al. [45] and Martin et al. [16], respectively. The chromosome distribution of both *Copia* and *Gypsy* elements showed higher densities in proximal regions, contrasting with the distal enrichment of genes (Figs. 1E, S1, and S9A), a pattern consistent with observations across other Musaceae species. Analysis of LTR retroelement insertion times in two haplotypes indicated two major episodes of LTR activity: a recent peak around 0.46 Mya and older peak around 3 Mya (Fig. 4C). In comparison, *Ensete glaucum* exhibits a predominant insertion peak between 3.5 and 5.5 Mya with a secondary peak at 0.5 Mya [11], whereas *M. balbisiana* and *M. acuminata* display major peaks at 0.5 Mya and 1.5 Mya, respectively [9]. These results suggest that retroelement insertion timings are species-specific within Musaceae, reflecting at least two distinct bursts of retroelement activity and genome instability (Fig. 4C).

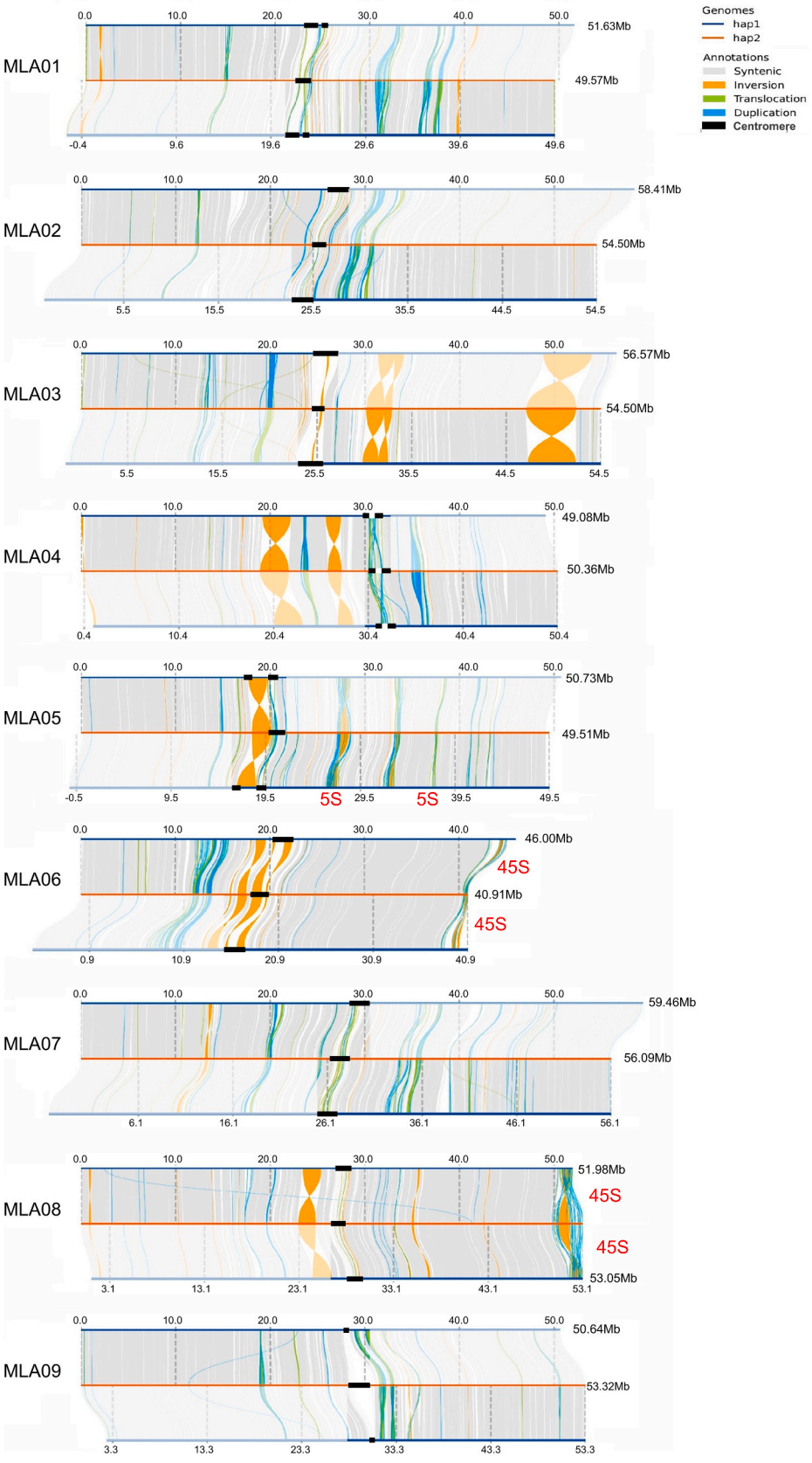
Our haplotype assemblies enabled a detailed examination of Presence Absence Variations (PAVs) between homologous chromosome pairs (Fig. 4D). Genomic variations >50 bp were distributed relatively uniformly across the chromosomes, with localized peaks on MLA02 (Fig. 4D, red density maps), that correspond to regions identified in the synteny plots (Fig. 3). Extraction of polymorphisms between haplotypes that were annotated as transposable elements (Fig. 4D, blue density maps) also showed an overall uniform distribution across nine chromosomes, with additional enrichment in centromeric regions where transposable elements were more abundant (Fig. 1E).

2.5.2. Telomeres

MLAh1 and MLAh2 each contained arrays of 160–6,000 copies of the telomeric monomer (TTTAGGG/CCCTAAA), located at 30 of 36 pseudochromosome ends (Table S4). Telomeric sequences were absent from the right ends of MLA06 and MLA08, both of which terminate with 45S rDNA arrays, as well as from the left ends of MLA05.1 and MLA08.2 (Table S6A). Consistent with previously published genome assemblies, telomeres were not assembled on chromosomes bearing 45S sites. Additionally, no ONT reads >50 kb were detected that contained both 45S rDNA loci and telomeric monomers. Fluorescence *in situ* hybridization (FISH), however, detected the telomeric $(TTTAGGG)_7$ at the ends of all chromosomes, appearing as double dots of variable intensity, including at sites distal to the nucleolar organizing regions (NORs; Figs. 4E and S10A). We propose that the extension of the transcribed rDNA array within the interphase nucleus may cause DNA breakage during extraction. Since the chromosomal segment distal to the NOR appears to be very small based on FISH results (Fig. S10A, B), such segments containing the telomeres may have been lost during sequencing library preparation and, consequently, could not be assembled.

2.5.3. Centromeres

A 134 bp centromeric monomer (Egcen) from *E. glaucum* was used to identify a homologous centromeric repeat, Mlcen, in MLA. Mlcen



(caption on next page)

Fig. 3. Comparison maps of MLAh1 (yellow line) and MLAh2 (blue line) by SyRI [42] plots. Syntenic regions, inversions, translocations, and duplications are indicated by grey, orange, green, and blue curves, respectively; Black boxes represent centromeres. The 5S and 45S rDNA loci are indicated by red words. Since the largest InDels occurred around centromeres, while synteny was highest along distal arm regions, except for the more structural variations (SVs) on MLA06 left ends and on MLA08 right ends due to 45S rDNAs. To present all SVs, MLAh1 (yellow line) was placed in the central position with MLAh2 (blue line) above and below; the alignments started from the left side for the above pairs, and from the right side for the below pairs in each plot to show clearly synteny along chromosome arms. Statistics of all SVs are summarized in Table S14. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

accounts for approximately 2.33% of the MLA genome (86,889 copies; Tables 2 and S5) and shares 98% sequence identity with Egcen. Notably, Mlcen was not detected in the centromeric region of *Musa* chromosomes [11]. Cluster analysis of Illumina sequencing reads using RepeatExplorer2 [44], identified Mlcen as the first cluster representing about 2.3% of the genome, and displaying a characteristic 'star-shaped' pattern typical of tandem satellite repeats (Fig. 4F). In addition, *k*-mer analysis using TAREAN program [44] confirmed Mlcen as a high confidence satellite repeat). Mlcen is a GC-rich sequence (46.6%) that contains part of the functional CENP-B box as well as 11-mer motif, suggesting low sequence variability within its 134 bp monomer. Arrays of Mlcen were detected on nine chromosome assemblies (Table S5B), and FISH revealed strong, though variably intense signals at centromeric regions (Figs. 4E, 5A, and S10A, B). Based on these Mlcen arrays, centromeric locations were defined within the assemblies (Figs. 1E, S1), and chromosome arm lengths were estimated (Table S5C, D).

Comparison of MLA haplotypes reveals differences in centromeric organization. The Mlcen arrays range from 1.0 to 10.6 Mb, with the Mlcen region of MLA09.1 containing less than half the copy number of MLA09.2 (Table S5B). In some chromosomes, the Mlcen arrays were interrupted by genes (MLA01), LTR retroelements (MLA04), or inversions MLA05 (Fig. 3). The regions surrounding these arrays show variable structural features: some harbor genes and few variations on MLA03 and MLA08 (Fig. S7C, I), while others exhibit extensive inversions, translocations, and duplications, such as on MLA05, MLA06, MLA07 (Fig. S7E-H).

2.5.4. rDNA

In the MLA assemblies, two 45S rDNA loci were identified at the distal ends of the right arm of MLA06 and MLA08, and a single 5S rDNA locus was located in the middle of the right arm of MLA05 (Figs. 1E and S1), consistent with FISH results (Figs. 5A-D and S10C-F). The 5S rDNA monomer was 658 bp in length, comprising a 119 bp of gene (typical for all plants) and a 539 bp the intergenic spacer (IGS), with 2,069 copies representing 0.27% of MLA genome (Tables 2 and S6). The 5S rDNA locus on MLA05 was organized into two major arrays separated by roughly 10 Mb unrelated sequences (Fig. 5E, F). FISH revealed the organization of 5S rDNA on ml05, exhibiting a discontinuous signal along extended prometaphase chromosomes (Fig. 5A, right). This observation supports the presence of two distinct 5S rDNA arrays on ml05, analogous to the fragmented loci on EGL05 and the paired 5S rDNA sites reported in *Musa spp.* [11,46].

FISH analysis revealed that 45S rDNA sites at the ends of chromosome ml06 and ml08 co-localized with the tandem repeat MuTR (GenBank: AM905888, Teo and Schwarzacher unpubl. data) (Fig. 5B, D). Sequence analysis confirmed that each 45S rDNA monomer typically contains four copies of a degenerate MuTR tandem repeat. The arrangement on MLA06 and MLA08 follows the structure: 18S rRNA gene-ITS1–5.8S rRNA gene-ITS2-26S rRNA gene-NTSL-STR-MuTR-NTSR (Fig. 5G). The 45S rDNA monomer in MLA is 8,347 bp in length (Table S6B), slightly longer than in *Musa acuminata* (7,553 bp; [47]) but shorter than in *E. glaucum* (9,984 bp) [11]. The arrays are occasionally interrupted by LTR retroelement insertions (Fig. 5H), indicating a complex organization of the 45S rDNA loci (Figs. S11 and S12). In total, MLA contains 1,465 copies of 45S rDNA monomer, accounting for 2.58% of the genome (Tables 2 and S6B). This pattern differs from *Musa spp.* (1, 2, or 4 sites) and EGL (1 site) [11,46,48].

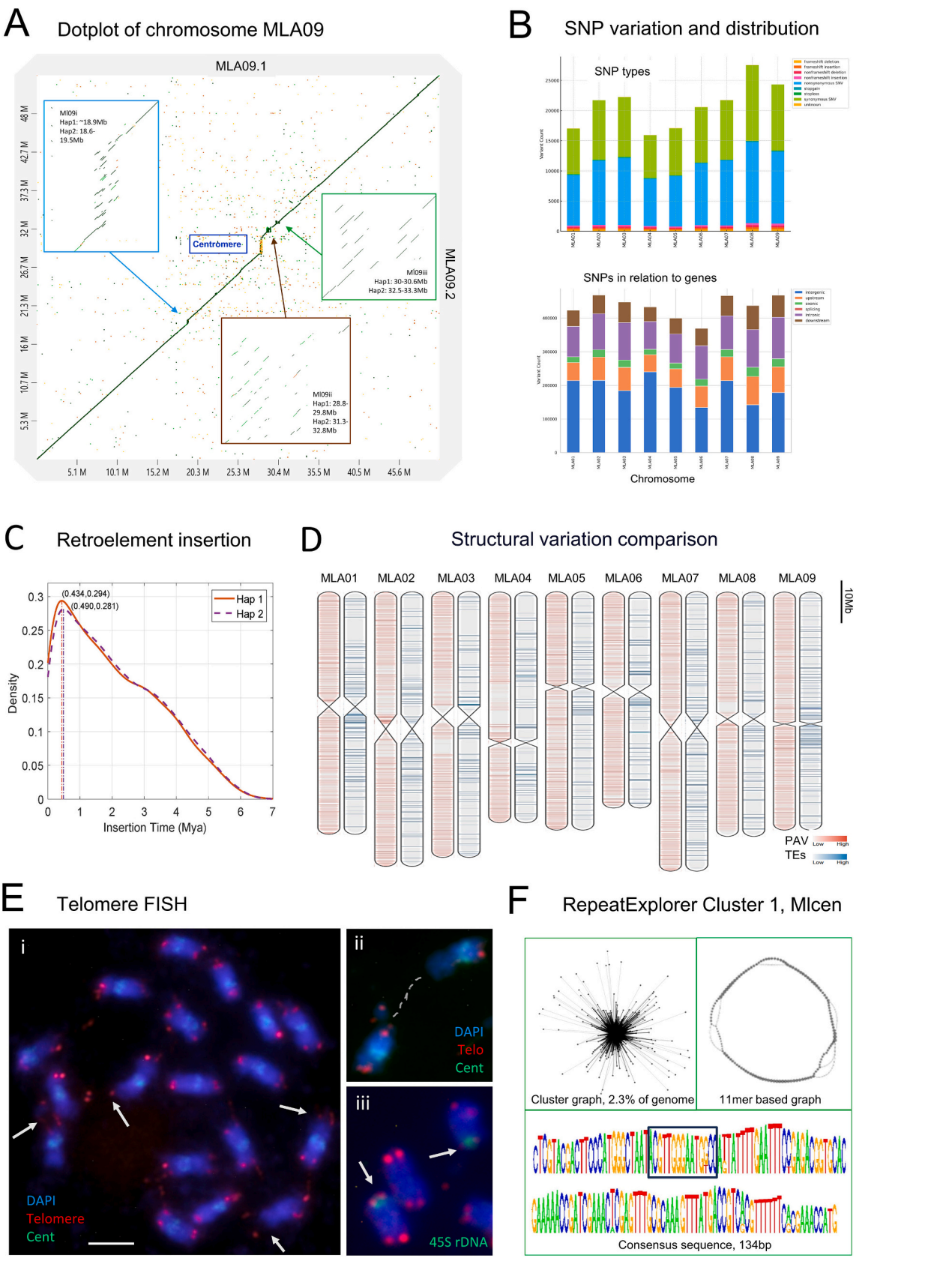
2.6. Synteny and karyotype evolution

Syntenic analyses were performed to compare the genomic structures of *Musella lasiocarpa* ($x = 9$, MLAh1) with other Musaceae species (Fig. 6), including *E. glaucum* (EGL, $x = 9$, [11]), *Musa acuminata* v.4 (MAC, $x = 11$) and *M. beccarii*, (MBE, $x = 9$, [12]). The assemblies showed similar numbers of gene orthogroups (Fig. 2A). Using MCScan [49], the number of syntenic genes of MLA-MAC (19,683) was higher than MLA-EGL (17,529) and EGL-MBE (14,722). The highest number of syntenic genes were present between MBE-MAC (21,214) in Musaceae (Tables S7B).

There was a high level of similarity between MLA-EGL (Fig. 6A). Six chromosome pairs–EGL02/MLA02, EGL03/MLA03, EGL05/MLA05 (containing 5S rDNA arrays), EGL07/MLA07, EGL08/MLA08, and EGL09/MLA09 shared both gene order and content. This was supported by parallel lines in SynViso [50] and dotplot results (Fig. S13). However, synteny was less conserved in pericentromeric regions rich in retroelements and repetitive DNAs (Fig. 1E), despite the presence of the shared centromeric repeats (Mlcen and Egcen). The remaining three chromosomes showed a small number of SVs: MLA01 consisted of the intercalary region of EGL01, with additional segments derived from translocations and internal inversions involving EGL04 and EGL07. MLA04 contained three large segments from EGL01, EGL04, and EGL06, along with smaller parts from EGL07. MLA06 had segments originating from EGL04 and EGL06. Among these, MLA07 showcased the highest similarity to EGL07 in terms of gene content (Fig. 6A).

The positions of rDNA loci underwent substantial shifts during species radiation, often accompanied by a depletion of PCGs on chromosome regions immediately adjacent to 45S rDNA loci in Musaceae [11,12,51]. A comparison of the karyotypes of MLA and EGL revealed changes in both individual chromosome sizes and centromere positions, particularly in chromosomes 01 and 04 (Fig. 6B). Notably, the chromosome arm carrying the 45S rDNAs in EGL06 showed depletion of genes [11], while the 45S rDNA bearing arms in MLA06 and MLA08 maintained gene density similar to other chromosome arms (Fig. 6A). Furthermore, the 45SrDNA genes in MLA are located within a different genomic context compared to EGL: on MLA06, they are on the opposite arm relative to EGL06, and on MLA08, they are in region where EGL had no 45S rDNA sequences. As mentioned previously, both 45S rDNA sites in MLA are positioned at the end of the genome assemblies; however, the arrays are not completely assembled, as indicated by the absence of the telomere sequence TTTAGGG that is observed as signals by FISH (Figs. 4E and S10B).

A larger number of chromosomal rearrangements were observed between MLA ($x = 9$) with MAC ($x = 11$) (Fig. 6C), beyond the consequence of chromosome number differences. Except for MLA05 (entirely syntenic with MAC05), all chromosomes showed multiple rearrangements. The inversions and translocations mainly involved large segments except for MLA04, which is composed of regions derived from MAC03, MAC04, MAC06 and MAC10. MLA09 includes all of MAC11 chromosome and part of MAC04, indicating possible centromere gain or loss. An interesting scenario is found for EGL-MLA-MBE ($x = 9$); in comparison with the $x = 9$ EGL and MLA, no single chromosome remained syntenic in their entirety and showed extensive rearrangements. Hence, we postulate that MBE ($x = 9$) possesses a derived karyotype, as evidenced by fewer rearrangements revealed in MBE-MAC synteny (Fig. 6D).



(caption on next page)

Fig. 4. MLA structural variation comparison and repeat analyses. **A**: Structural variations of MLA09 dotplot showing Ml09i, Ml09ii and Ml09iii and the centromere repeats. **B**: SNP variation between MLA haplotypes. Variation types are shown on the top, and SNP positions within genes are shown on the bottom. **C**: Retroelement insertion time with a peak value of 0.46 Mya for MLA haplotypes. **D**: MLA haplotype structural variation comparison. Left density maps represent the distribution of Presence-Absence Variations (PAVs, red), right density maps represent retroelement variations (TEs, blue) with 0.10 Mb bins. **E**: Fluorescence *in situ* hybridization (FISH) on MLA metaphase chromosomes (blue with DAPI): (i) and (ii) Centromeric Mlcen probe (green) and telomere probe (red). Arrows represent the weak nucleolar organizer regions (NORs), and the broken satellite repeats are indicated by dotted line; (iii) 45S rDNA probe (green) and telomere probe (red). For complete images of (ii) and (iii), see Fig. S10, where enhancements showing that the telomere signals are present on all chromosome ends including those with 45S rDNA, bar = 5 μm. **F**: RepeatExplorer cluster 1 (CL1) graphs and consensus sequence of 134 bp Mlcen in MLAh1. **Top left**: 'Star-shaped' tandem CL1; **Top right**: CL1 graph from TAREAN pipeline [44]; **Bottom**: The consensus sequence of 134 bp Mlcen monomer variation. The black rectangle representing the functional CENP-B box motif. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. Discussion

3.1. Haplotype assembly

MLA haplotype assembly is critical for revealing the nature of polymorphisms within a single non-inbred plant, whose heterozygous genome is rich in repetitive sequences (Figs. 1, 3, and S1). Studies of allelic differences between genes have formed the foundation of genetics for decades; however, consensus and mosaic assemblies often omit important allelic variations that influence key traits. In contrast, haplotype assemblies of polymorphic species show allelic variation across all genes within a single accession ([13,52]). The existing genome assemblies constitute valuable resources for allelic studies, including genes associated with disease resistance [18], secondary metabolites [53], and cold tolerance in the Musaceae. Our haplotype assemblies further enhance this understanding by revealing structural variation including inversions, translocations, sequence amplifications, and duplicatons, between haplotypes.

3.2. Genes and cold adaptation

Cold is a major abiotic factor that markedly affects banana production [26]. The plants' response to cold is complex, involving both rapid "early" (< 6 h) and slow "late" (> 24 h) responses [27] with genes either up- or down-regulated. In our cold treatment experiments, we identified genes and transcription factors known to play roles in cold tolerance of MLA (Fig. 2E). The C-repeat/DREB binding factor (CBF)-dependent responsive pathway (involving ERF, Ethylene Response Factors, Fig. 2D) has been identified as important for survival of MLA, *Musa* [27], grasses [41], and Arabidopsis [40]. These genes, along with other stress-related transcription factors (such as WRKY, zinc-finger, and basic helix-loop-helix HLH, with the PIF phytochrome subfamily; or lectin receptor-like kinases, [54]), contribute to the differential expression in cold-treated MLA (Fig. 2D, bottom) as well as in previous studies [40,41]. These TFs are candidate genes to support breeding strategies for cold-resilient crops in Musaceae [27,55].

Genome assemblies and transcriptome data are proving useful for revealing the regulation of secondary metabolite production and the variation. Zhao et al. [18] show the *O*-methyltransferases involved in phytoalexin biosynthesis, providing valuable genetic resources related to anti-fungal compounds. Flavonoid biosynthesis is critical for plant adaptation to extreme environments [53]. As examples, the assembly and transcriptome analyses of the *Phanera championii* [56] reveal how alleles among haplotypes influence flavonoid biosynthesis. Cui et al. [53] show that in MLA genotypes, cold stress modulates the expression of flavonoid biosynthesis genes and TFs, altering the relative abundance of different anthocyanins and their biosynthetic pathways.

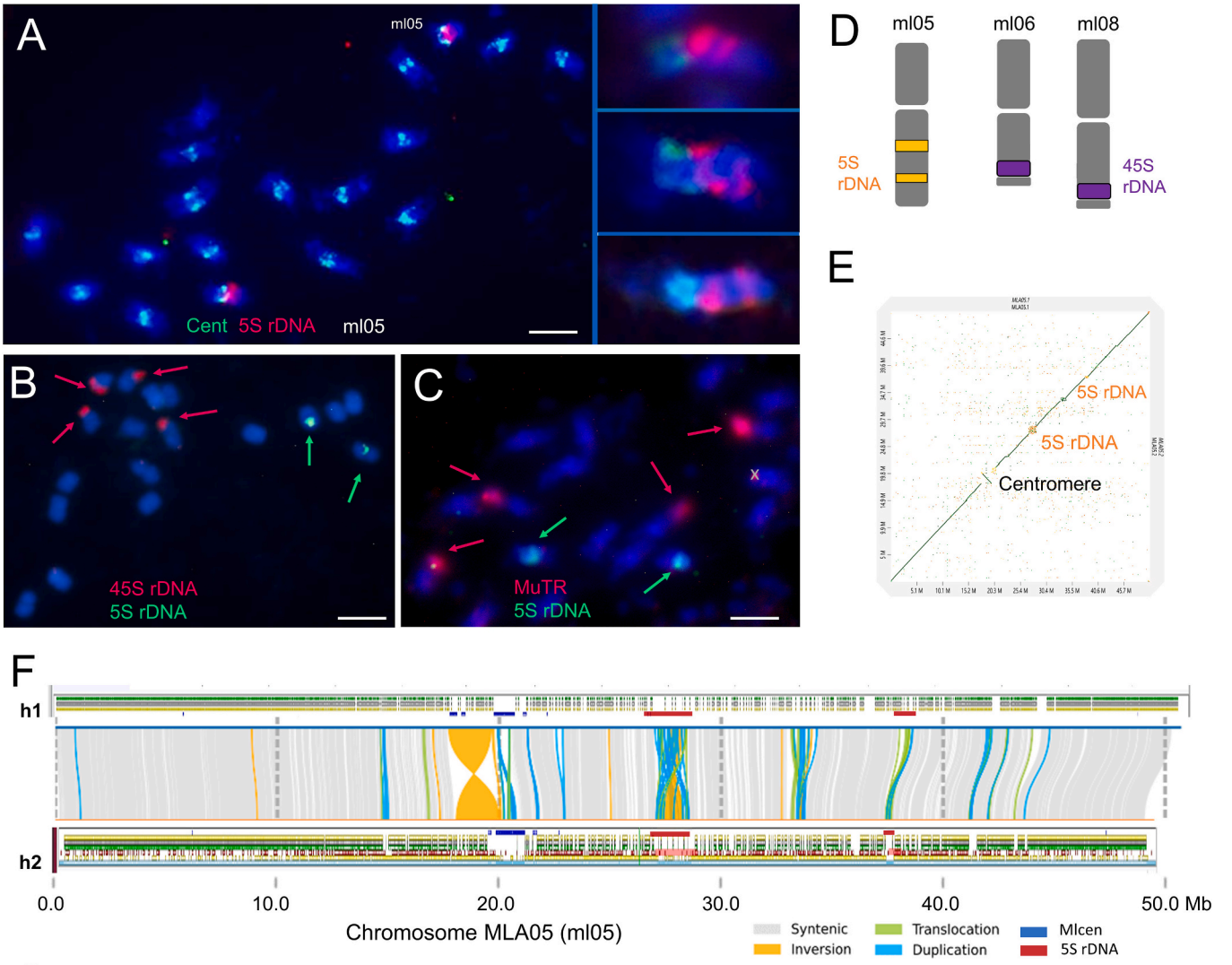
3.3. Haplotype SNPs and heterozygosity

The haplotype-based SNP comparison revealed a heterozygosity level of 0.78%, which is lower than the 1.24% inferred by GenomeScope (Fig. S2A). This estimate falls within a two-fold range of values reported for other Musaceae species by the similar approach, *e.g.*, *Musa acuminata* (0.59%, [14]), *M. acuminata* ssp. *banksii* (0.02–0.34%, [57]), *Ensete glaucum* (0.164%, [11]), and *E. ventricosum* (0.73%, [58]). Our result aligned with the expected analysis differences, as GenomeScope often over-estimates heterozygosity due to the presence of pseudogene and retroelement fragments in *k*-mer profiles. SNPs occurred throughout the genic and intergenic regions of the genome (Fig. 4D), and potential inbreeding events in some lineages. The reduced density of SNPs in exonic region likely reflects stronger evolutionary constraints that limit the tolerance of variation in protein-coding sequences. Within these exonic SNPs, the proportions of synonymous and non-synonymous variants were nearly equal (Fig. 4B), indicating a balance between mutations that preserve protein function.

Structural variations between haplotypes exhibited a range of differences (Fig. 3) in broad, repeat-rich, pericentromeric regions (Figs. 1E, S1), as elucidated in Musaceae genomes [6,16,59]. Polymorphisms were confirmed by long-single-molecule ONT reads (*e.g.*, Fig. S8). The broad distribution of retroelements between haplotypes (Fig. 4D) is consistent with the polymorphisms identified among *Musa* accessions, identified through Inter-Retroelement Amplified Polymorphisms (IRAPs) [60]. This consistency demonstrates the reliability of the polymorphisms for variety identification [61]. Our genome assemblies will serve as a useful resource for future studies aimed at understanding the impact of SVs, including retroelement insertions, on gene expression.

In *Musa*, phased haplotype assemblies have been generated from diploid [14], triploid *M. acuminata* [13,19], and *Musa* AAB triploid hybrids [20]. In wild diploid *M. acuminata*, 47 translocations (cumulative size 2.7 Mb), 23 inversions (11.3 Mb) and 53 duplications (1.33 Mb) were detected. These values are substantially lower than in MLA, where 459 translocations (7.4 Mb), 140 inversions (23 Mb) and 1,108 duplications (13.5 Mb) were observed. In *Musella*, the average recombination rate was 1.2 events per Mb, representing 26–37 recombination events per chromosome arm (Table S5C) per meiosis (Table S14). Interestingly, the relatively large inversion fragments (8,045,525 bp) of MLA03 did not alter the recombination rate, unlike MLA01, which displayed the highest number and sizes of translocations and duplications. These inversions likely suppress crossover formation in the affected chromosome regions (as shown by [62]). In triploid *M. acuminata*, the synteny plots revealed additional rearrangements, particularly in haplotype BXJ2 ([13]: their Fig. 1f). Overall, comparisons of multiple *Musa* genomes indicate that cultivated *Musa* are genetically derived from multiple wild progenitors [16]. In contrast, in diploid *Musella,* no inter-chromosomal exchanges or gaps flanking inversions were observed. Both phased assemblies were continuous (Fig. S8), as confirmed by single-molecular reads (Figs. S7 and S9). This suggests that homologous exchanges involving loss of segments between chromosomes as seen in MLA are more likely associated with speciation events [16].

Three haplotype-resolved assemblies of the apple genome, each approximately 650 Mb in size, provide a useful comparator [52]. As observed in Musaceae, gene density was highest near the chromosome ends (their Fig. 3a circle 6 of [52]), while transposable elements tended to be in proximal regions (also their circle 6). SVs were distributed across chromosomes (their circle 2), including approximately ten large inversions of multiple-megabase size (their Fig. 3c). In apple, 4.76% of SNPs were located in the coding sequences, similar to the 4.81% in *Musella*. However, unlike MLA, the distribution of SNPs and InDels in the apple genome was uneven, with certain chromosomes or large segments



(caption on next page)

Fig. 5. Distribution and structure of 5S and 45S ribosomal DNA (rDNA) in MLA genome. **A-C**: FISH on MLA metapahse chromosomes (blue with DAPI). **A**: 5S rDNA probe (red) on chromosome ml05 and Mlcen probe (green). Complete metaphase on the left and enlarged pro-metaphase chromosomes on the right showing the double 5S rDNA site. **B**: 5S (green) and 45S (red) rDNA probes. **C**: MuTR (red, arrows) and 5S rDNA (green, arrows) probes. Bar = 5 μm. **D**: Diagram showing locations of 5S rDNAs (orange squares) on ml05 and 45S rDNA loci (purple squares) on ml06 and ml08. **E**: Dotplot of MLA05.1 and MLA05.2 showing two 5S rDNA loci are located on the right arm of ml05 with the 10 Mb separation. **F**: Structural variations between MLA05.1 and MLA05.2. Syntenic regions, inversions, translocations, duplications, Mlcen, and 5S rDNA are indicated by grey, yellow, green, blue, dark blue, and red boxes. Bars above and below chromosome represent distributions of genes, coding regions (green, yellow grey), and LTR-retroelements (light red). **G**: The 45S rDNA monomer and 1,465 copy arrays. The left monomer consists of tandem repetitive units of 18S–5.8S-28S rRNA and non-transcribed spacer (NTS) regions (NTSL-STR-MuTR-NTSR). **H**: Dotplot of 45S rDNA arrays (8,347 bp) against MLAh1 (Left top). The MuTR monomers are denoted by grey dotted squares (Left up and bottom). Right blue square: Self dotplot of the inserted LTRs (6,100 bp) within 45S rDNA arrays. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

exhibiting very few variations (their Fig. 3a circles 3 and 4 in [17]). This pattern raises the possibility that uniparental disomy, potentially arising from meiotic from non-disjunction in inbred parents (a phenomenon best known in animals), may have occurred in this system.

The breeding system and natural pollinators of *M. lasiocarpa* remains poorly understood. In the horticultural trade and possibly in the wild, the species primarily propagates through vegetative side-suckers [63], with fruit production being rare in cultivation. Female sterility promotes clonal growth [64], restricts gene flow, and helps maintain heterozygosity. Additionally, SVs in the genome can disrupt meiosis, leading to reduced fertility [64] and inhibits recombination within the species, complicating efforts to produce interspecific or intergeneric hybrids or generate desirable recombinant genotypes.

3.4. Repeat analysis and polymorphisms

Variation in the copy number of tandem repeats, including rDNA arrays, is commonly observed between homologous chromosomes, both in species and within a single individual [65]. In *M. lasiocarpa*, the 45S rDNA arrays on MLA06 and MLA08 exhibited pronounced differences between haplotypes (Figs. 3 and 5F). Small tandem arrays also contribute to haplotypic polymorphisms; however, many of these arrays were degenerate and interspersed with other sequences including chloroplast fragments, genes, and retroelements. These mixed arrays tend to occur in genomic regions that are rich in either genes or retroelements (Figs. 3 and S7). The polymorphisms located in and around centromeric regions, which were generally depleted in genes, are likely generated through uneven crossing-over. Such SVs have been reported in various species. For instance, in poplar, a ~ 200 kb tandem-repeat array is inserted at the centromere of Chr04A relative to Chr04T [66]. In plants, chromosome pairing during meiosis is conventionally initiated at the telomeres, with recombination occurring more frequently in distal regions [67]. Consequently, the highly syntenic MLA chromosome ends likely facilitate accurate homologous pairing and recombination during meiosis. Conversely, the high density of repeats, frequent SVs and inversions around centromeres (Figs. 1E and 3), are expected to suppress meiotic recombination in these pericentromeric regions.

3.5. Evolution, diversification, and synteny in Musaceae

Our genome assembly confirmed *Musella* to be a distinct genus within Musaceae, exhibiting a closer relationship to *Ensete* than to *Musa*. *Musella* can be distinguished from *Musa* [68] and *Ensete* [69] by its erect, compact rosette inflorescences, and its genome assembly provides an opportunity to identify the genes underlying these unique traits. Syntenic comparison revealed extensive rearrangements in EGL and MLA relative to *Musa acuminata* ($x = 11$) (Fig. 6D), and even more so compared to *M. beccarii*, which we propose as a derived $x = 9$ species. A previous study by D'Hont et al. [45] showed that *Musa* underwent three rounds of WGD events. In this study, we also detected evidence for the α and β WGDs at the Cretaceous-Paleocene boundary (Fig. 2B), as well as the more ancient γ WGD event at ~100.0 Mya [45] and *E. glaucum* [11], with no further additional MLA-specific WGD events. Dating based on single-copy orthologs indicated that the *Musella* and *Ensete* genera diverged approximately 42 Mya, while they split from *Musa* around 57 Mya (Fig. 2C). These estimates align with findings from recent nuclear and chloroplast genome studies [4,70].

4. Conclusion

Our publicly available, high-quality, haplotype-resolved, telomere-to-telomere (T2T) genome assembly of MLA, generated through the integration of long- and short-read sequencing technologies, provides detailed insights into SVs between MLA haplotypes and the role of repetitive sequences in generating diversity. In general, the detailed analysis of the nature of haplotype variations suggests the mechanisms for recombination-lead generation of polymorphisms, often involving repetitive DNA, including between genotypes of inbred lines. In addition to gene-level differences, these SVs likely contribute to phenotypic variation, limit recombination within the species, and influence the formation of wide hybrids. Such insights are critical for effectively utilizing crop genetic resources (in Musaceae, [71], and more widely) to enhance the agronomically important traits through breeding and crop improvement.

5. Materials and methods

Additional information about material and methods is provided in supplementary file (Data S1).

5.1. Collection of plant material and extraction of DNA

Fresh leaves of *Musella lasiocarpa* var. *lasiocarpa* and var. *rubribracteata* were collected from the South China Botanical Garden (originally sourced from Nanhua city, Chuxiong Yi Autonomous Prefecture, Yunnan, China). Genomic DNA was extracted using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany), and its quality was assessed through gel electrophoresis, a NanoDrop One UV–Vis spectrophotometer and a Qubit 3.0 fluorometer (Thermo Fisher Scientific, USA).

5.2. Library preparation and whole genome sequencing

Multiple sequencing libraries were generated using different technologies: the PacBio HiFi library was prepared with the SMRTbell Express Template Prep Kit 2.0, the Oxford Nanopore library was created using the SQK-LSK109 kit on PromethION, and the Hi-C library [72] was prepared with modified protocols on the Illumina NovaSeq 6000 platform (Illumina, USA). For the Hi-C library preparation, cells were cross-linked with 2% formaldehyde, digested using *Dpn*II (New England Biolabs), and biotinylated with biotin-14-dCTP, and generating 2 × 150 bp paired-end Illumina reads. For ONT transcriptome sequencing, RNA was extracted using the TRNzol Universal Kit (Tiangen). Quality control of the sequencing data was performed by fastp v.0.23.3 [73].

5.3. Transcriptome sequencing for cold response

MLA varieties were initially cultivated at 25 °C and then subjected to cold treatment of 15 °C for 48 h. RNA was extracted using the Plant RNA Kit (OMEGA-R6827) and sequenced on the DNBSEQ-T7 platform (BGI, Shenzhen, China) with three biological replicates. DEGs were identified

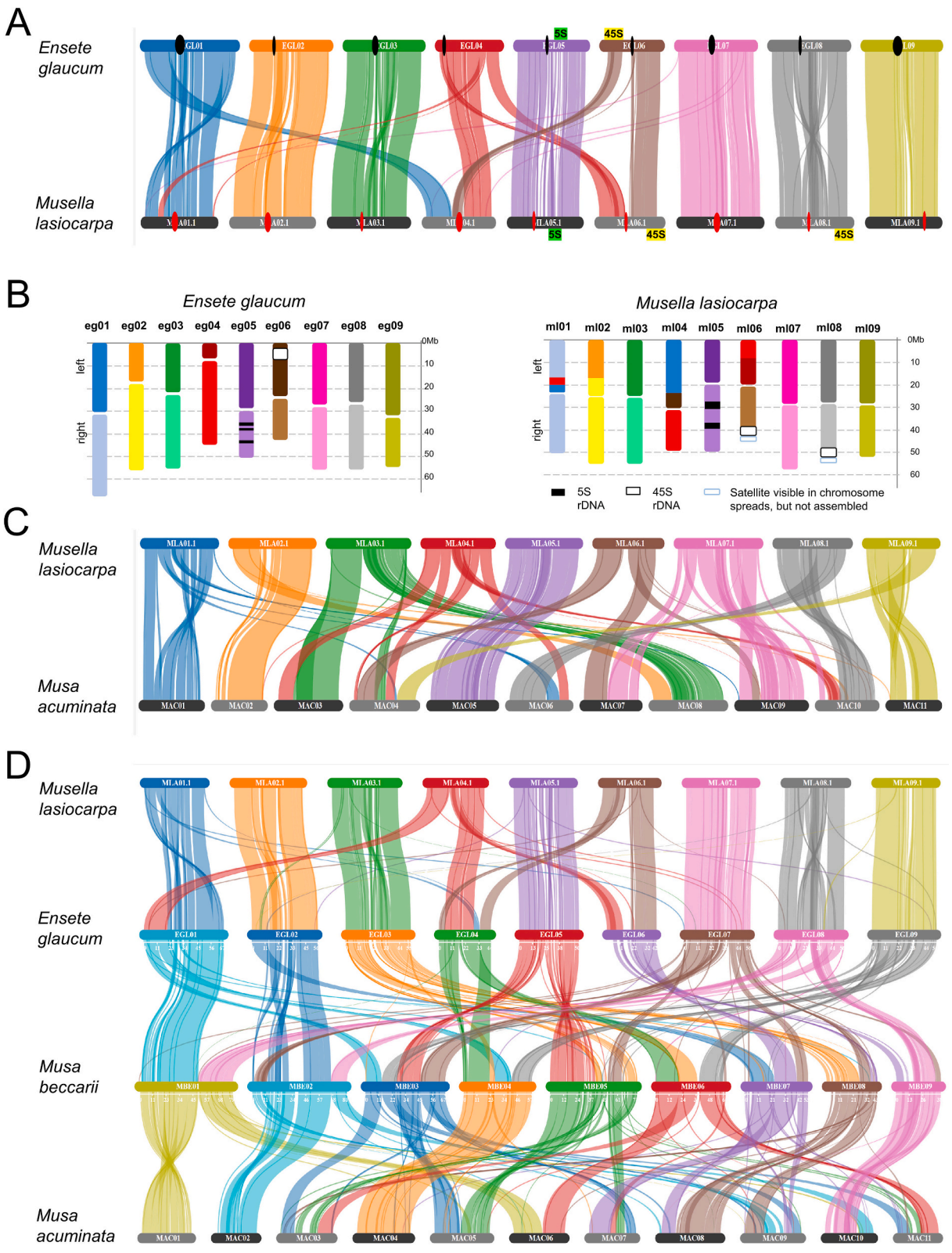


Fig. 6. Genome synteny plots of MLA and other Musaceae species. **A**: Synteny plot of MLA-*Ensete glaucum* (EGL). Centromeres are indicated by red and black dots, and 5S and 45S rDNA sites are indicated by green and yellow boxes, respectively. **B**: Synteny blocks on EGL adopted from Wang et al. [11] and MLA karyotypes (Data from Table S5D). Syntenic regions on short and long arms are indicated by darker and lighter colors, and 5S, 45S rDNAs, and satellites are indicated by black, white, and blue boxes, respectively. **C**: Synteny plot of MLA-*Musa acuminata* (MAC). Four MAC chromosomes (MAC01/02/05/11) correspond to parts of four MLA chromosomes (MLA01/02/05/09), while the remaining seven chromosomes are split into multiple fragments, constituting nine MLA chromosomes. The syntenic regions of MLA-MAC (Fig. 6C) are few than those of MLA-EGL (Fig. 6A). **D**: Synteny plot of Musaceae showing extensive rearrangements between EGL-MBE and MBE-MAC ($x = 11$). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

using DESeq2 v.1.54.0 [74] with thresholds of fold-change >2 and the adjusted *P*-value (padj) < 0.05. TFs associated with cold adaptation were identified using PlantTFDB [75] with default setting (without exclusive selection of "best hit in *Arabidopsis thaliana*"). A heatmap of gene expression levels was generated by TBtools [76], applying row scale normalization.

5.4. k-mer analysis and genome assembly

The genome size was estimated using GenomeScope v.2.0 [29] with 21-mers generated by Jellyfish v.2.3.0 [77]. The phased genome assembly was generated by combining PacBio HiFi reads, HiC reads, and ONT reads [filtered using Filtlong v.0.2.4 [78] and assembled with Hifiasm v.0.16.1 [79]]. The resulting genome assemblies were then manually refined by Juicer v. 1.5 [80] and Juicebox [81].

5.5. Genomic evaluation and annotation

Among other methods, the quality of the assembly was assessed using BUSCO v.5.5.0 [82]. Repeats were annotated with RepeatModeler [83] and EDTA v.2.1.0 [84], and LTR elements were classified using TEsorter [85]. Genome masking was then performed with RepeatMasker (http://www.repeatmasker.org). Gene prediction and functional annotation were carried out on the soft-masked genomes by integrating *ab initio*, homology-based, and transcriptome-assisted prediction approaches. *Ab initio* prediction was performed by Augustus v.3.5.0 [86] and SNAP [87]. The homologous protein sequences were aligned by Tblastn v.2.7.1 [88] and GeMoMa [86], the alignment results were used to predict gene structures. For transcriptome prediction, RNA-seq data and ONT read alignment were performed by STAR v.2.7 [89], StringTie v.2.7 [90] and PASA [91]. All evidence was weighted and merged with EVidenceModeler [92] to produce the final non-redundant gene set (Full details in Data S1).

Gene functions were annotated by Diamond v.2.0.11.149 [93], based on comparisons against major databases including NCBI NR [32], InterPro [33], GO [34], KOG [35], KEGG [36], TrEMBL, and SwissProt [37]. Conserved sequences, motifs, and domains of proteins were identified using InterProScan v.5.68 [94] and Hmmscan v3.3.2 [95] based on comparisons with the InterPro and Pfam databases [96]. The MLA genome features were visualized using Circos [97].

5.6. Gene family and genome synteny analysis

Gene families were analyzed using OrthoFinder v.2.4.0 [98], and divergence times were estimated with PAML v.4.9j [99]. R package UpSetR [38] was used for the orthogroup distribution visualization. WGD events were analyzed using WGDI v.0.6.5 [100]. Structural variations were detected with Mummer v.4.0.0 [101] and SyRI v.1.6 [42]. To examine genomic synteny among MLA and EGL/MAC/MBE, MCScanX [49] was employed, and resulting synteny blocks were visualized using SynVisio [50].

5.7. FISH analysis

Root tip preparation and FISH were carried out following Schwarzacher and Heslop-Harrison [102]. Details of probe mixtures were provided in Table S16. Images were acquired using Nikon Eclipse 80i microscope (Nikon, Japan) equipped with UV-2 A, FITC, and TRITC filters. Image overlays were created with Nikon NIS-elements software and adjusted in Photoshop v.25.6.0.

CRediT authorship contribution statement

Qing Liu: Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition, Data curation, Conceptualization. **Dongli Cui:** Writing – review & editing, Writing – original draft, Methodology. **Yaqi Tian:** Writing – original draft, Resources, Methodology, Investigation. **Yehan Wang:** Writing – original draft, Resources, Methodology, Investigation. **Mathieu Rouard:** Validation, Software, Investigation, Formal analysis, Data curation. **John Seymour Heslop-Harrison:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis. **Trude Schwarzacher:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis. **Ziwei Wang:** Validation, Software, Investigation, Formal analysis, Data curation.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32370402, 32070359), Guangdong Flagship Project of Basic and Applied Research (2023B0303050001), Chinese Academy of Sciences (CAS) President's International Fellowship Initiative (2024PVA0028), Compilation Project for Flora of Nanling Grasses (Forestry Administration of Guangdong Province), UK Research and Innovation (UKRI) *via* the Engineering and Physical Sciences Research Council (EPSRC; EP/Y00597X/1-project RP13W471907), and UK Biotechnology and Biological Sciences Research Council (BB/R022828/1). We would like to thank Jing Li for MLA sample collection, Benagen Technology Company Ltd. (Wuhan, China) for Illumina, Nanopore transcriptome sequencing and Beijing Genomics Institute-Shenzhen (BGI, Shenzhen) for PacBio HiFi sequencing and T2T genome assembly.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2026.111210.

Data availability

The telomere-to-telomere genome assemblies of *Musella lasiocarpa* have been deposited in NCBI (PRJNA1117752 for MLAh1; PRJNA1117751 for MLAh2) and the website of https://banana-genome-hub.southgreen.fr/node/50/16444789 [6]. The raw sequence data including Illumina, PacBio HiFi, and ONT genome sequences, as well as RNAseq data, can be accessed in the NCBI SRA library (SAMN41579419), and the China National Center for Bioinformation at https://ngdc.cncb.ac.cn/gsa/browse/CRA014572.

References

[1] J. Plummer, R. Allen, S. Kallow, Musella lasiocarpa., The IUCN Red List of Threatened Species, e.T98249468A, 2022.
[2] C.L. Long, S. Ahmed, X.Y. Wang, Y.T. Liu, B. Long, C.Y. Yang, Y.N. Shi, X.Y. Li, R. Guo, Why *Musella lasiocarpa* (Musaceae) is used in Southwest China to feed pigs, Econ. Bot. 62 (2008) 182–186.
[3] J.S. Heslop-Harrison, T. Schwarzacher, Domestication, genomics and the future for banana, Ann. Bot. 100 (2007) 1073–1084.
[4] N. Fu, M. Ji, M. Rouard, H.F. Yan, X.J. Ge, Comparative plastome analysis of Musaceae and new insights into phylogenetic relationships, BMC Genomics 23 (2022) 223.
[5] H. Ma, Q.J. Pan, L. Wang, Z.H. Li, Y.M. Wan, X.X. Liu, *Musella lasiocarpa* var. *rubribracteata* (Musaceae), a new variety from Sichuan, China, Novon 21 (2011) 349–353.
[6] G. Droc, G. Martin, V. Guignon, M. Summo, G. Sempéré, E. Durant, A. Soriano, F. B. Baurens, A. Cenci, C. Breton, et al., The banana genome hub, a community database for genomics in the Musaceae, Hortic. Res. 9 (2022) uhac221.
[7] C. Belser, F.C. Baurens, B. Noel, G. Martin, C. Cruaud, B. Istace, N. Yahiaoui, K. Labadie, E. Hřibová, J. Doležel, et al., Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing, Commun. Biol. 4 (2021) 1047.
[8] C. Belser, B. Istace, E. Denis, M. Dubarry, F.C. Baurens, C. Falentin, M. Genete, W. Berrabah, A.M. Chèvre, R. Delourme, et al., Chromosome-scale assemblies of

plant genomes using nanopore long reads and optical maps, Nat. Plants 4 (2018) 879–887.
[9] Z. Wang, H.X. Miao, J.H. Liu, B.Y. Xu, X.M. Yao, C.Y. Xu, S.C. Zhao, X.D. Fang, C. H. Jia, J.Y. Wang, et al., *Musa balbisiana* genome reveals subgenome evolution and functional divergence, Nat. Plants 5 (2019) 810–821.
[10] L.C. Galvez, R.B.L. Koh, C.F.C. Barbosa, J.C. Asunto, J.L. Catalla, R.G. Atienza, K. T. Costales, V.M. Aquino, D.P. Zhang, Sequencing and *de novo* assembly of abaca (*Musa textilis* Née) var. *abuab* genome, Genes 12 (2021) 1202.
[11] Z. Wang, M. Rouard, M.K. Biswas, G. Droc, D.L. Cui, N. Roux, F.C. Baurens, X. J. Ge, T. Schwarzacher, J.S. Heslop-Harrison, et al., A chromosome-level reference genome of *Ensete glaucum* gives insight into diversity and chromosomal and repetitive sequence evolution in the Musaceae, GigaScience 11 (2022) giac027.
[12] Z.F. Wang, M. Rouard, G. Droc, J.S. Heslop-Harrison, X.J. Ge, Genome assembly of *Musa beccarii* shows extensive chromosomal rearrangements and genome expansion during evolution of Musaceae genomes, GigaScience 12 (2023) giad005.
[13] H.R. Huang, X. Liu, R. Arshad, X. Wang, M.W. Li, Y.F. Zhou, X.J. Ge, Telomere-to-telomere haplotype-resolved reference genome reveals subgenome divergence and disease resistance in triploid Cavendish banana, Hortic. Res. 10 (2023) uhad153.
[14] X. Liu, R. Arshad, X. Wang, W.M. Li, Y.F. Zhou, X.J. Ge, H.R. Huang, The phased telomere-to-telomere reference genome of *Musa acuminata*, a main contributor to banana cultivars, Sci. Data 10 (2023) 631.
[15] T.W. Xiao, X. Liu, N. Fu, T.J. Liu, Z.F. Wang, X.J. Ge, H.R. Huang, Chromosome-level genome assemblies of *Musa ornata* and *Musa velutina* provide insights into pericarp dehiscence and anthocyanin biosynthesis in banana, Hortic. Res. 11 (2024) uhae079.
[16] G. Martin, B. Istace, F.C. Baurens, C. Belser, C. Hervouet, K. Labadie, C. Cruaud, B. Noel, G. Chantal, F. Salmon, et al., Unravelling genomic drivers of speciation in *Musa* through genome assemblies of wild banana ancestors, Nat. Commun. 16 (2025) 961.
[17] Z. Li, J. Wang, Y. Fu, Y.L. Jing, B.L. Huang, Y. Chen, Q.L. Wang, X.B. Wang, C. Y. Meng, Q.Q. Yang, et al., The *Musa troglodytarum* L. genome provides insights into the mechanism of non-climacteric behaviour and enrichment of carotenoids, BMC Biol. 20 (2022) 186.
[18] W.L. Zhao, J.Z. Wu, M. Tian, S. Xu, S.Y. Hu, Z.Y. Wei, G.Y. Lin, L. Tang, R. Y. Wang, B.Y. Feng, et al., Characterization of O-methyltransferases in the biosynthesis of phenylphenalenone phytoalexins based on the telomere-to-telomere gapless genome of *Musella lasiocarpa*, Hortic. Res. 11 (2024) uhae042.
[19] X.X. Li, S. Yu, Z.H. Cheng, X.J. Chang, Y.Z. Yun, M.W. Jiang, X.Q. Chen, X. H. Wen, H. Li, W.J. Zhu, et al., Origin and evolution of the triploid cultivated banana genome, Nat. Genet. 56 (2024) 136–142.
[20] Z.W. Xie, Y.Y. Zheng, W.D. He, F.C. Bi, Y.Y. Li, T.X. Dou, R. Zhou, Y.X. Guo, G. M. Deng, W.H. Zhang, et al., Two haplotype-resolved genome assemblies for AAB allotriploid bananas provide insights into banana subgenome asymmetric evolution and *Fusarium wilt* control, Plant Commun. 5 (2024) 100766.
[21] J.S. Heslop-Harrison, T. Schwarzacher, Q. Liu, Polyploidy: its consequences and enabling role in plant diversification and evolution, Ann. Bot. 131 (2023) 1–10.
[22] Q. Liu, L.H. Ye, M.Z. Li, Z.W. Wang, G. Xiong, Y.S. Ye, T.Y. Tu, T. Schwarzacher, J.S. Heslop-Harrison, Genome-wide expansion and reorganization during grass evolution: from 30 Mb chromosomes in rice and *Brachypodium* to 550 Mb in *Avena*, BMC Plant Biol. 23 (2023) 627.
[23] Y. Bao, Z. Zeng, W. Yao, X. Chen, M.W. Jiang, A. Sehrish, B. Wu, C.A. Powell, B. S. Chen, J.L. Xu, et al., A gap-free and haplotype-resolved lemon genome provides insights into flavor synthesis and huan-glongbing (HLB) tolerance, Hortic. Res. 10 (2023) uhad020.
[24] G. Li, L. Tang, Y. He, Y.Y. Xu, A. Bendahmane, J. Garcia-Mas, T. Lin, G.W. Zhao, The haplotype-resolved T2T reference genome highlights structural variation underlying agronomic traits of melon, Hortic. Res. 10 (2023) uhad182.
[25] K. Li, R.H. Chen, A. Abudoukayoumu, Q. Wei, Z.B. Ma, Z.Y. Wang, Q. Hao, J. Huang, Haplotype-resolved T2T reference genomes for wild and domesticated accessions shed new insights into the domestication of jujube, Hortic. Res. 11 (2024) uhae071.
[26] R.U. Joshi, A.K. Singh, V.P. Singh, R. Rai, P. Joshi, A review on adaptation of banana (*Musa* spp.) to cold in subtropics, Plant Breed. 142 (2023) 269–283.
[27] Q.S. Yang, J. Gao, W.D. He, T.X. Dou, L.J. Ding, J.H. Wu, C.Y. Li, X.X. Peng, S. Zhang, G.J. Yi, Comparative transcriptomics analysis reveals difference of key gene expression between banana and plantain in response to cold stress, BMC Genomics 16 (2015) 1–8.
[28] Y. Yang, W. Du, Y. Li, J. Lei, W. Pan, Recent advances and challenges in *de novo* genome assembly, Genom. Commun. 2 (2025) e014.
[29] G.W. Vurture, F.J. Sedlazeck, M. Nattestad, C.J. Underwood, H. Fang, J. Gurtowski, M.C. Schatz, GenomeScope: fast reference-free genome profiling from short reads, Bioinformatics 33 (2017) 2202–2204.
[30] M. Pfenninger, P. Schönenbeck, T. Schell, ModEst: accurate estimation of genome size from next-generation sequencing data, Mol. Ecol. Resour. 22 (2022) 1454–1464.
[31] M. Manni, M.R. Berkeley, M. Seppey, F.A. Simão, E.M. Zdobnov, BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes, BMC Plant Biol. 38 (2021) 4647–4654.
[32] K.D. Pruitt, T. Tatusova, D.R. Maglott, NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts, and proteins, Nucleic Acids Res. 33 (2005) D501–D504.
[33] T. Paysan-Lafosse, T. Blum, S. Chuguransky, T. Grego, B.L. Pinto, G.A. Salazar, M. L. Bileschi, P. Bork, A. Bridge, L. Colwell, et al., InterPro in 2022, Nucleic Acids Res. 51 (2023) D418–D427.
[34] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A. P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al., Gene ontology: tool for the unification of biology, Nat. Genet. 25 (2001) 25–29.
[35] R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, et al., The COG database: an updated version includes eukaryotes, BMC Bioinf. 4 (2003) 41.
[36] M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima, M. Ishiguro-Watanabe, KEGG for taxonomy-based analysis of pathways and genomes, Nucleic Acids Res. 51 (2023) D587–D592.
[37] A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, Nucleic Acids Res. 28 (2000) 45–48.
[38] J.R. Conway, A. Lex, N. Gehlenborg, UpSetR: an R package for the visualization of intersecting sets and their properties, Bioinformatics 33 (2017) 2938–2940.
[39] H.L. Li, Z.M. Dong, Y.S. Jiang, Y.S. Jiang, S.J. Jiang, H.T. Xing, Q. Li, G.C. Liu, S. M. Tian, Z.Y. Wu, et al., Haplotype-resolved genome of diploid ginger (*Zingiber officinale*) and its unique gingerol biosynthetic pathway, Hortic. Res. 8 (2021) 189.
[40] Y.P. Li, Y.T. Shi, M.Z. Li, D.Y. Fu, S.F. Wu, J.G. Li, Z.Z. Gong, H.T. Liu, S.H. Yang, The CRY2–COP1–HY5–BBX7/8 module regulates blue light-dependent cold acclimation in *Arabidopsis*, Plant Cell 33 (2021) 3555–3573.
[41] A.R. Pashapu, G. Statkevičiūtė, F. Sustek-Sánchez, M.R. Kovi, O.A. Rognli, C. Sarmiento, N. Rostoks, K. Jaškūnė, Transcriptome profiling reveals insight into the cold response of perennial ryegrass genotypes with contrasting freezing tolerance, Plant Stress 14 (2024) 100598.
[42] M. Goel, H. Sun, W.B. Jiao, K. Schneeberger, SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies, Genome Biol. 20 (2019) 1–13.
[43] Q. Liu, H.Y. Yuan, J.X. Xu, D.L. Cui, G. Xiong, T. Schwarzacher, J.S. Heslop-Harrison, The mitochondrial genome of the diploid oat *Avena longiglumis*, BMC Plant Biol. 23 (2023) 218.
[44] P. Novák, P. Neumann, J. Macas, Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2, Nat. Protoc. 15 (2020) 3745–3776.
[45] A. D'Hont, F. Denoeud, J.M. Aury, F.C. Baurens, F. Carreel, O. Garsmeur, B. Noel, S. Bocs, G. Droc, M. Rouard, et al., The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants, Nature 488 (2012) 213.
[46] J. Čížková, E. Hřibová, L. Humplikova, P. Christelová, P. Suchánková, J. Doležel, Molecular analysis and genomic organization of major DNA satellites in banana (*Musa* spp.), PloS One 8 (2013) e54808.
[47] E. Hřibová, P. Neumann, T. Matsumoto, N. Roux, J. Macas, J. Dolezel, Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing, BMC Plant Biol. 10 (2010) 204.
[48] J.O. Osuji, J. Crouch, G. Harrison, J.S. Heslop-Harrison, Molecular cytogenetics of *Musa* species, cultivars and hybrids, localisation of 18S-5.8S-25S and 5S rDNA and telomere-like sequences, Ann. Bot. 82 (1998) 243–248.
[49] Y. Wang, H. Tang, J.D. Debarry, X. Tan, J.P. Li, X.Y. Wang, T.H. Lee, H.Z. Jin, B. Marler, H. Guo, et al., MCScanX, a toolkit for detection and evolutionary analysis of gene synteny and collinearity, Nucleic Acids Res. 40 (2012) e49.
[50] V. Bandi, C. Gutwin, Interactive exploration of genomic conservation, in: Proceedings of the 46th Graphics Interface Conference on Proceedings of Graphics Interface 2020 (GI'20), Canadian Human-Computer Communications Society, Waterloo, Canada, 2020.
[51] J. Bartoš, O. Alkhimova, M. Doleželová, E. De Langhe, J. Dolezel, Nuclear genome size and genomic distribution of ribosomal DNA in *Musa* and *Ensete* (Musaceae): taxonomic implications, Cytogenet. Genome Res. 109 (2005) 50–57.
[52] W. Li, C. Chu, H. Li, H.T. Zhang, H.C. Sun, S.Y. Wang, Z.J. Wang, Y.Q. Li, T. M. Foster, E. López-Girona, et al., Near-gapless and haplotype-resolved apple genomes provide insights into the genetic basis of rootstock-induced dwarfing, Nat. Genet. 56 (2024) 505–516.
[53] D.L. Cui, G. Xiong, L.H. Ye, R. Gornall, T. Schwarzacher, J.S. Heslop-Harrison, Q. Liu, Genome-wide analysis of flavonoid biosynthetic genes in Musaceae (*Ensete*, *Musella*, and *Musa* species) reveals amplification of flavonoid 3'5'-hydroxylase, AoB Plants 16 (2024) plae049.
[54] G. Xiong, D.L. Cui, Y.Q. Tian, T. Schwarzacher, J.S. Heslop-Harrison, Q. Liu, Genome-wide identification of lectin receptor-like kinases gene family in *Avena sativa* and their roles in salt stress tolerance, Int. J. Mol. Sci. 25 (2024) 12754.
[55] Z.Y. Luo, Z.C. Zhou, Y.Y. Li, S.T. Tao, Z.Y. Hu, J.S. Yang, X.J. Cheng, R.S. Hu, W. L. Zhang, Transcriptome-based gene regulatory network analyses of differential cold tolerance of two tobacco cultivars, BMC Plant Biol. 22 (2022) 369.
[56] Y.B. Lu, X. Chen, H. Yu, C. Zhang, Y.J. Xue, Q. Zhang, H.F. Wang, Haplotype-resolved genome assembly of *Phanera championii* reveals molecular mechanisms of flavonoid synthesis and adaptive evolution, Plant J. 118 (2024) 488–505.
[57] J. Sardos, C. Breton, X. Perrier, I.V. den Houwe, S. Carpentier, J. Paofa, M. Rouard, N. Roux, Hybridization, missing wild ancestors and the domestication of cultivated diploid bananas, Front. Plant Sci. 13 (2022) 969220.
[58] A.T. Haile, M.R. Kovi, S.S. Johnsen, T. Hvoslef-Eide, B. Tesfaye, O.A. Rognli, Limited genetic diversity found among genotypes of the Entada landrace (*Ensete ventricosum*, (Welw.) Chessman) from Ethiopia, Front. Plant Sci. 15 (2024) 1336461.
[59] G. Martin, F. Carreel, O. Coriton, C. Hervouet, C. Cardi, P. Derouault, D. Roques, F. Salmon, M. Rouard, J. Sardos, et al., Evolution of the banana genome (*Musa acuminata*) is impacted by large chromosomal translocations, Mol. Biol. Evol. 34 (2017) 2140–2152.

[60] A.S. Nair, C.H. Teo, T. Schwarzacher, J.S. Heslop-Harrison, Genome classification of banana cultivars from South India using IRAP markers, Euphytica 144 (2005) 285–290.
[61] D. Beránková, J. Čížková, G. Majzlíková, A. Doležalová, H. Mduma, A. Brown, R. Swennen, E. Hřibová, Striking variation in chromosome structure within *Musa acuminata* subspecies, diploid cultivars, and F1 diploid hybrids, Front. Plant Sci. 15 (2024) 1387055.
[62] H. Li, E. Berent, S. Hadjipanteli, M. Galey, N. Muhammad-Lahbabi, D.E. Miller, K. N. Crown, Heterozygous inversion breakpoints suppress meiotic crossovers by altering recombination repair outcomes, PLoS Genet. 3 (2023) 1010702.
[63] A.Z. Liu, W.J. Kress, D.Z. Li, Insect pollination of *Musella lasiocarpa* (Musaceae): a monotypic genus endemic to China, Plant Syst. Evol. 235 (2002) 135–146.
[64] C.Y. Xue, H. Wang, D.Z. Li, Female gametophyte and seed development in *Musella lasiocarpa* (Musaceae), a monotypic genus endemic to Southwestern China, Can. J. Bot. 85 (2007) 964–975.
[65] Q. Liu, X.Y. Li, X.Y. Zhou, M.Z. Li, F.J. Zhang, T. Schwarzacher, P. Heslop-Harrison, The DNA landscape in *Avena*: chromosome and genome evolution defined by major repetitive DNA classes in whole-genome sequence reads, BMC Plant Biol. 19 (2019) 226.
[66] R. Zhou, J.W. Jenkins, Y. Zeng, S.Q. Shu, H.S. Jang, S.A. Harding, M. Williams, C. Plott, K.W. Barry, M. Koriabine, et al., Haplotype-resolved genome assembly of *Populus tremula* × *P. alba* reveals aspen-specific megabase satellite DNA, Plant J. 116 (2023) 1003–1017.
[67] A. Sepsi, J.D. Higgins, J.S. Heslop-Harrison, T. Schwarzacher, CENH3 morphogenesis reveals dynamic centromere associations during synaptonemal complex formation and the progression through male meiosis in hexaploid wheat, Plant J. 89 (2017) 235–249.
[68] M.Z. Franchet, Un nouveau type de *Musa lasiocarpa*, J. Bot. 3 (1889) 329–331.
[69] E.E. Cheesman, Classification of the bananas, Kew Bull. 2 (1947) 97–117.
[70] R. Zhou, S. Wang, N. Zhan, W.D. He, G.M. Deng, T.X. Dou, X.T. Zhu, W.Z. Xie, Y. Y. Zheng, C.H. Hu, et al., High-quality genome assemblies for two *Australimusa* bananas (*Musa* spp.) and insights into regulatory mechanisms of superior fiber properties, Plant Commun. 5 (2024) 100681.
[71] C. Jenny, V. Guignon, I. Manyer, F. Ballester, M. Ruas, M. Rouard, Collecting and managing *in situ* banana genetic resources information (*Musa* spp.) using online resources and citizen science, Database 2024 (2024) baae036.
[72] J.M. Belton, R.P. McCord, J.H. Gibcus, N. Naumova, Y. Zhan, J. Dekker, Hi-C: a comprehensive technique to capture the conformation of genomes, Methods 58 (2012) 268–276.
[73] S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (2018) i884–i890.
[74] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (2014) 550.
[75] J.P. Jin, F. Tian, D.C. Yang, Y.Q. Meng, L. Kong, J.C. Luo, et al., PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants, Nucleic Acids Res. 45 (2017) D1040–D1045.
[76] C.J. Chen, Y. Wu, J.W. Li, X. Wang, Z.H. Zeng, J. Xu, Y.L. Liu, J.T. Feng, H. Che, Y. H. Ye, et al., Tbtools-II: a "one for all, all for one" bioinformatics platform for biological big-data mining, Mol. Plant 16 (2023) 1733–1742.
[77] G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, Bioinformatics 27 (2011) 764–770.
[78] R.R. Wick, L.M. Judd, C.L. Gorrie, K.E. Holt, Completing bacterial genome assemblies with multiplex MinION sequencing, Microb. Genom. 3 (2012) (2017) e000.132.
[79] H. Cheng, G.T. Concepcion, X. Feng, H.W. Zhang, H. Li, Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm, Nat. Methods 18 (2021) 170–175.
[80] N.C. Durand, M.S. Shamim, I. Machol, S.S.P. Rao, M.H. Huntley, E.S. Lander, E. L. Aiden, Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments, Cell Syst. 3 (2016) 95–98.
[81] N.C. Durand, J.T. Robinson, M.S. Shamim, I. Machol, J.P. Mesirov, E.S. Lander, E. L. Aiden, Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom, Cell Syst. 3 (2016) 99–101.
[82] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (2015) 3210–3212.
[83] J.M. Flynn, R. Hubley, C. Goubert, J. Rosen, A.G. Clark, C. Feschotte, A.F. Smit, RepeatModeler2 for automated genomic discovery of transposable element families, Proc. Natl. Acad. Sci. U. S. A. 117 (2020) 9451–9457.
[84] S.J. Ou, W.J. Su, Y. Liao, K. Chougule, J.R.A. Agda, A.J. Hellinga, C.S.B. Lugo, T. A. Elliott, D. Ware, T. Peterson, et al., Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline, Genome Biol. 20 (2019) 275.
[85] R.G. Zhang, G.Y. Li, X.L. Wang, J. Dainat, Z.X. Wang, S.J. Ou, Y.P. Ma, TEsorter, an accurate and fast method to classify LTR-retrotransposons in plant genomes, Hortic. Res. 9 (2022) uhac017.
[86] J. Keilwagen, F. Hartung, M. Paulini, S.O. Twardziok, J. Grau, Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi, BMC Bioinf. 19 (2018) 189.
[87] A.D. Johnson, R.E. Handsaker, S.L. Pulit, M.M. Nizzari, C.J.O. Donnell, P.I.W. de Bakker, SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap, Bioinformatics 24 (2008) 2938–2939.
[88] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. Madden, BLAST+: architecture and applications, BMC Bioinf. 10 (2009) 421.
[89] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T.R. Gingeras, STAR: ultrafast universal RNA-seq aligner, Bioinformatics 29 (2013) 15–21.
[90] S. Kovaka, A.V. Zimin, G.M. Pertea, R. Razaghi, S.L. Salzberg, M. Pertea, Transcriptome assembly from long-read RNA-seq alignments with StringTie2, Genome Biol. 20 (2019) 278.
[91] B.J. Haas, A.L. Delcher, S.M. Mount, J.R. Wortman, R.K. Smith Jr., L.I. Hannick, R. Maiti, C.M. Ronning, D.B. Rusch, C.D. Town, et al., Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies, Nucleic Acids Res. 31 (2003) 5654–5666.
[92] B.J. Haas, S.L. Salzberg, W. Zhu, M. Pertea, J.E. Allen, J. Orvis, O. White, C. R. Buell, J.R. Wortman, Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments, Genome Biol. 9 (2008) R7d.
[93] B. Buchfink, K. Reuter, H.G. Drost, Sensitive protein alignments at tree-of-life scale using DIAMOND, Nat. Methods 18 (2011) 366–368.
[94] P. Jones, D. Binns, H.Y. Chang, M. Fraser, W.Z. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, et al., InterProScan 5: genome-scale protein function classification, Bioinformatics 30 (2014) 1236–1240.
[95] R.D. Finn, J. Clements, S.R. Eddy, HMMER web server: interactive sequence similarity searching, Nucleic Acids Res. 39 (2011) W29–W37.
[96] R.D. Finn, A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, et al., Pfam: the protein families database, Nucleic Acids Res. 42 (2014) D222–D230.
[97] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, M.A. Marra, Circos: an information aesthetic for comparative genomics, Genome Res. 19 (2009) 1639–1645.
[98] D.M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics, Genome Biol. 20 (2019) 238.
[99] Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood, BMC Plant Biol. 24 (2007) 1586–1591.
[100] P.C. Sun, B.B. Jiao, Y.Z. Yang, L. Shan, T. Li, X.N. Li, Z.X. Xi, X.Y. Wang, J.Q. Liu, WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes, Mol. Plant 15 (2022) 1841–1851.
[101] G. Marçais, A.L. Delcher, A.M. Phillippy, R. Coston, S.L. Salzberg, A. Zimin, MUMmer4: a fast and versatile genome alignment system, PLoS Comput. Biol. 14 (2018) e1005944.
[102] T. Schwarzacher, P. Heslop-Harrison, Practical *In Situ* Hybridization, BIOS Scientific Publishers Ltd., 2000.