

Inversions encounter relaxed genetic constraints and balance birth and death of *TPS* genes in *Curcuma*

Received: 9 April 2024

Accepted: 21 October 2024

Published online: 29 October 2024

 Check for updatesXuezhu Liao^{1,4}, Dejin Xie^{1,4}, Tingting Bao^{1,4}, Mengmeng Hou¹, Cheng Li¹, Bao Nie¹, Shichao Sun¹, Dan Peng¹, Haixiao Hu², Hongru Wang¹, Yongfu Tao¹, Yu Zhang³, Wei Li¹ & Li Wang¹✉

Evolutionary dynamics of inversion and its impact on biochemical traits are a puzzling question. Here, we show abundance of inversions in three *Curcuma* species (turmeric, hidden ginger and Siam tulip). Genes within inversions display higher long terminal repeat content and lower expression level compared with genomic background, suggesting inversions in *Curcuma* experience relaxed genetic constraints. It is corroborated by depletion of selected SNPs and enrichment of deleterious mutations in inversions detected among 56 Siam tulip cultivars. Functional verification of tandem duplicated terpene synthase (*TPS*) genes reveals that genes within inversions become pseudo-genes, while genes outside retain catalytic function. Our findings suggest that inversions act as a counteracting force against tandem duplication in balancing birth and death of *TPS* genes and modulating terpenoid contents in *Curcuma*. This study provides an empirical example that inversions are likely not adaptive but affect biochemical traits.

Chromosomal inversions are a principal driver of genome structure evolution, which shield inverted haplotypes from exchanging genetic variation and suppress recombination¹. As other type of genetic mutations, inversions are subject to both neutral evolutionary processes (such as genetic drift) and natural selection. The fate of inversion polymorphism will depend on the role of the affected genes in specific circumstances². First, inversions may be spread and eventually fixed in a population as they inhibit recombination and thus maintain adaptive combinations of multiple beneficial genes for certain traits^{3–5}. Numerous empirical studies have demonstrated their phenotypic consequences and adaptive significance^{6–10}. For example, an inversion polymorphism in the white-throated sparrow (*Zonotrichia albicollis*), spanning over 100 Mbp and repressing recombination, was found to be associated with phenotypic variation in plumage color, social behavior, and mate choice¹¹. A well-characterized 14 Mbp inversion on chromosome 4 of maize (known as Inv4m), originally introgressed

from its wild teosinte relative, was reported to promote earlier flowering in the shorter growing season at high elevations and was thus essential for highland adaptation of maize populations^{12–14}. Second, some inversions will be lost due to genetic drift or purifying selection, as inversions may accumulate deleterious mutations^{9,15}, disrupt gene sequences alter gene expression¹⁶, or cause negative epistatic variation². For example, a 415 Kbp polymorphic inversion in human populations disrupted the transcription factor gene *ZNF257* and led to a significant reduction of its expression level in lymphoblastoid cell lines despite the absence of significant phenotypic effects. This inversion was more consistent with the assumption of deleterious polymorphism based on modeling the change in frequency of the inversion during evolution¹⁷. In the third scenario, both the inverted and non-inverted haplotypes were maintained over extended periods of evolutionary time, as balancing selection, disassortative mating, antagonistic pleiotropy, or selection along ecological gradients were

¹Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China. ²Department of Plant Sciences, University of California Davis, Davis, CA 95616, USA. ³School of Agriculture, Sun Yat-sen University, Shenzhen 518107, China. ⁴These authors contributed equally: Xuezhu Liao, Dejin Xie, Tingting Bao. ✉e-mail: wangli03@caas.cn

involved^{18–20}. Thus, inversions can have both direct and indirect effects on phenotypic traits and be adaptive, deleterious or neutral.

Whether or which evolutionary forces act on inversions can be assessed by their effect on genome features, such as nucleotide substitution rate, transposable element (TE) insertions, gene expression levels, etc. Inversions primarily alter the positional relationship of genes to their regulatory elements or disrupt genes, thereby modifying gene expression^{8,21–23}. For instance, an inversion in maize repositioned the promoter of the *p2* gene near the enhancer of the *p1* gene, which activated the expression of *p2* and resulted in red kernels; while cultivars with the non-inverted haplotype showed white kernels as *p2* was not expressed in kernel pericarp²³. Inversions are generally considered to have milder effects on gene function unless they disrupt gene sequences²⁴. However, recent studies showed that chromosomal inversions might affect the expression of thousands of genes across the genome in maize²⁵, yeast⁴, and *Drosophila*²², suggesting that genomic structural variation (SV) itself could have a profound effect on global gene expression. However, comprehensive views of the functional consequences of inversions on genes within the inversions are not yet well understood. In addition to effects on the order of genes and their regulatory elements, genes within inversions may be affected by other unknown pressures. Besides gene expression, little is known about the detrimental consequences of inversions, such as increased TE activity and accumulation of deleterious mutations³, which may disable the function of genes within inversion. Therefore, it is intriguing to explore how inversions affect the evolutionary fate of genes within the non-breakpoint inversion regions, and thorough investigations of inversions in multiple individuals and populations are necessary to determine their functional and evolutionary implications.

Curcuma species are an ideal system to examine the functional and evolutionary consequences of inversions. Firstly, *Curcuma*, the third largest genus in the family Zingiberaceae, demonstrates a wide range of variation in chromosome base numbers and ploidy levels among species, which hints at possibly extensive genomic SVs among species. Its chromosome base numbers vary from $2n = 20$ in *C. harmonidii*, $2n = 77$ in *C. oligantha*, to $2n = 105$ in *C. raktakanta*²⁶. Frequent polyploidization ($2x - 15x$) and hybridization lead to varying degrees of genetic and morphological variation among species and blur species boundaries^{26–32}, making taxonomic classification of the genus a challenging task. Secondly, *Curcuma* appears to be undergoing rapid radiation³³, which is often accompanied by chromosomal SVs. Such a case has been reported in *Rhodiola*, where extensive chromosome fusion/fission events and SVs within the *Rhodiola* genome might have facilitated the initial rapid radiation³⁴. Thirdly, *Curcuma* species, mainly distributed in tropical and subtropical Asia, are widely used as medicinal plants, ornamentals, and sources of starch. Turmeric, *C. longa*, is the main ingredient in curry and is widely known for its anti-cancer properties^{30–32}. *Curcuma* species contain abundant volatile compounds, including terpenoids, in their vegetative reproduction organ—rhizomes, tuberous roots, and flowers in some ornamental species^{29,35} (Fig. 1). The medicinal values of *Curcuma* species make it an ideal system to investigate the effects of inversion on biochemical traits, which differs from most existing reports that focus on the effects of inversion on morphological and agricultural traits³⁶. Taken together, to detect the extent of inversions in *Curcuma* and assess its evolutionary and functional effects, especially on biochemical traits, is worthy of in-depth exploration.

Here, we report genome assemblies of three *Curcuma* species, *C. longa* ($2n = 3x = 63$), *C. petiolata* ($2n = 2x = 42$), and *C. alismatifolia* ($2n = 2x = 32$), representing different phylogenetic lineages, ploidy levels and chromosome counts in the genus. We examine the prevalence of inversions within and among species and assess their effects on genomic features and biochemical traits. It reveals that exceptional

inversions in *Curcuma* species correlate with elevated TE content, increased substitution rate and accumulated deleterious mutations, as well as reduced gene expression levels within inversion regions, suggesting relaxed selection on inversions. Further examination demonstrates that tandem duplicated *TPS* genes within inversions become pseudogenized, affecting terpene biosynthesis.

Results

Haplotype-resolved genome assemblies of three *Curcuma* species

We selected three species of the *Curcuma* genus with different ploidy levels and chromosome numbers to assemble their genomes, representing genomic diversity of *Curcuma* species (Fig. 1a). Genome survey revealed a genome size of 737.68 Mbp for *C. longa* (Supplementary Fig. 1), 1.00 Gbp for *C. petiolata* (Supplementary Fig. 2) and 979.41 Mbp for *C. alismatifolia* (Supplementary Fig. 3). Furthermore, *C. longa* demonstrated a high heterozygosity ratio (4.67%, not autopolyploid AAA status) compared to the other two *Curcuma* species (1.61% for *C. alismatifolia* and 0.64% for *C. petiolata*) (Supplementary Figs. 1–3; Supplementary Note 1).

Based on 30–42 Gbp HiFi data and 100–150 Gbp Hi-C sequencing data, the retained contig N50 of newly assembled chromosome-level genomes of *C. alismatifolia*, *C. petiolata*, and *C. longa* were 23.37 Mbp, 25.81 Mbp, and 25.27 Mbp, respectively, which was significantly higher than previous reports (Table 1, Supplementary Table 1, Liao et al.²⁹, Yin et al.³⁷). Subsequently, 94.70–97.28% of the contigs were successfully anchored to the corresponding chromosomes (Table 1).

The calculated switching error (percentage of single bases that were supposed to be present in one haplotype were incorrectly anchored onto another) was 0.52% between the two haplotypes of *C. alismatifolia*, profoundly lower than 5.9% in *Camellia sinensis*³⁸, which confirmed accurate separation of these highly heterozygous haplotypes. As a result, the assembled genome size of *C. alismatifolia* was ~970 Mbp (951.46 and 973.47 Mbp for the two haplotypes); the size of both haplotypes of *C. petiolata* was 1.04 Gbp; while the size of *C. longa* was ~700 Mbp (696.23 Mbp, 715.95 Mbp, and 736.66 Mbp), which was smaller than the chimeric monoploidy genome size (1.11 Gbp) reported in Yin et al.³⁷ (Table 1 and Supplementary Figs. 4 and 5).

Next, we utilized two methods to estimate the integrity of the assembled genomes. First, BUSCO (Benchmarking Universal Single-Copy Orthologs) statistics revealed 98.55%, 98.50%, and 98.80% completeness for the conserved reference gene set ($n = 1614$) in embryophyta_odb10 database for *C. alismatifolia*, *C. petiolata*, and *C. longa*, respectively (Table 1). Second, long terminal repeat (LTR) assembly index (LAI)^{39,40} was utilized to estimate the integrity of LTR-RTs to assess the completeness of non-coding sequences. The results showed that the LAI of *C. alismatifolia*, *C. petiolata*, and *C. longa* were 25.46, 23.58, and 21.63 (Table 1), respectively, indicating a golden level of assembly. Noteworthy, LAI of *C. longa* was greatly improved compared to 13.56 reported in the previous chimeric genome³⁷.

Subsequently, two layers of genome annotation were conducted for the three species. First, we annotated TE using EDTA software⁴¹ and identified 76.23%, 76.78%, and 69.40% of TE in *C. alismatifolia*, *C. petiolata*, and *C. longa*, respectively, which echoed previous results^{29,37} (Table 1 and Supplementary Data 1). Second, protein-coding genes were annotated with a combination of ab initio prediction, homology annotation, and transcript assembly. The number of genes in the three genomes was 50,239/48,465, 45,764/44,683, and 37,115/37,161/37,995 for *C. alismatifolia*, *C. petiolata* and *C. longa*, respectively (Table 1). The BUSCO statistics of the proteins showed an average completeness of over 95% (Table 1), much higher than previous reports^{29,37}. In summary, our assembled genomes and annotated genes were complete and convincing, which laid a good foundation for the following analyses.

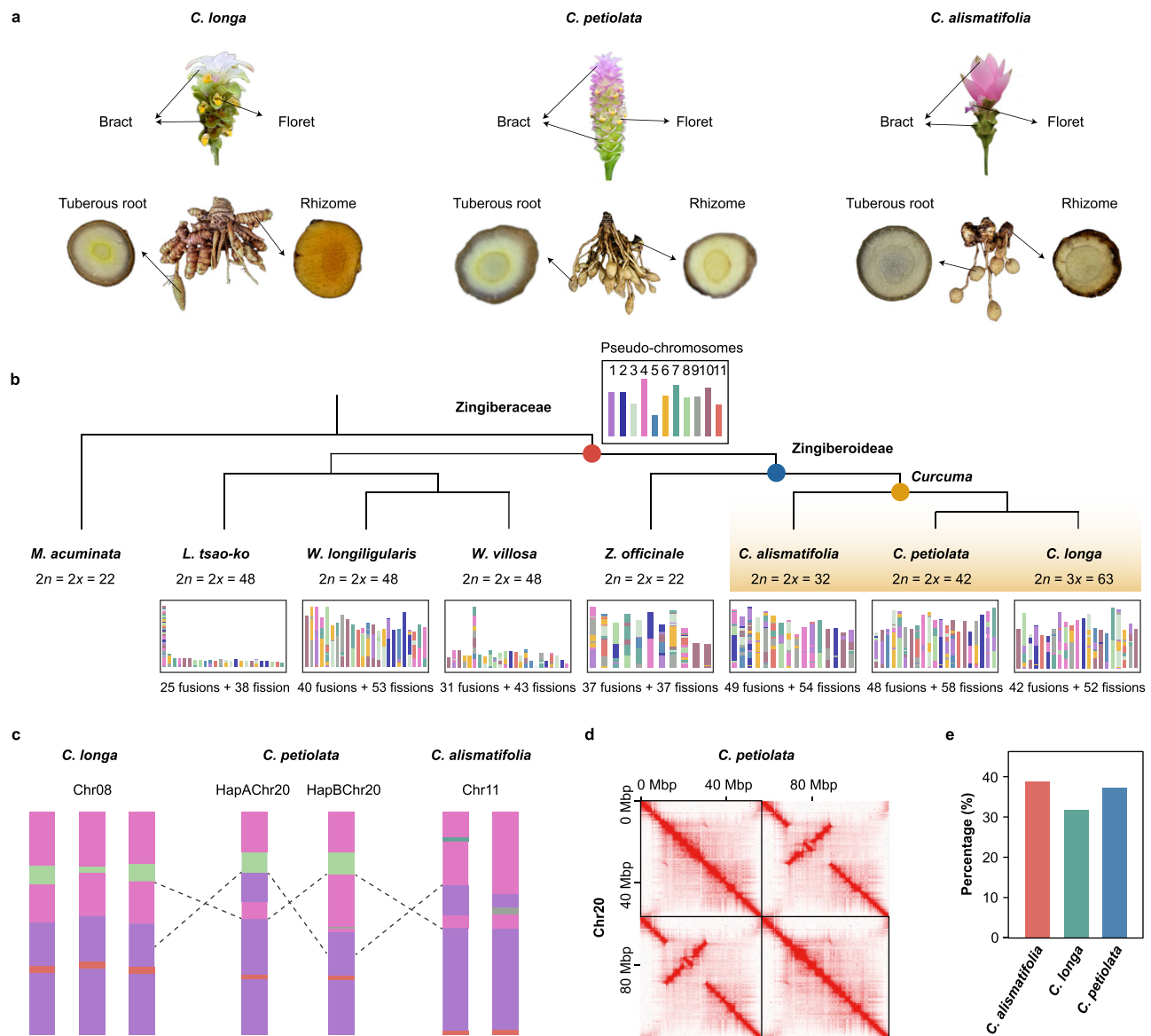


Fig. 1 | Phenotypic divergence and karyotype evolution among *Curcuma* species. **a** Images of bract, floret, rhizome, and tuberous root of *C. alismatifolia*, *C. petiolata*, and *C. longa*. **b** Ancestral chromosomal reconstruction of Zingiberaceous species. Phylogeny was constructed based on 1791 single-copy orthologous genes. 11 pseudo-chromosomes of the common ancestor of Zingiberaceae were marked with different colors. It showed that modern species experienced at least dozens of chromosome fusion and fission events. The red dot represents the species of

Zingiberaceae, the blue dot represents the species of Zingiberoideae, and the yellow dot represents the species of *Curcuma*. **c** A diagram showed the presence of inversions between haplotypes of *C. petiolata* and among the three *Curcuma* species. HapA: haplotype A; HapB: haplotype B. **d** Hi-C interaction signal validated an inversion between two haplotypes of Chromosome 20 in *C. petiolata*. **e** Percentage of the total length of inter-specific inversions between haplotypes from any two of the three *Curcuma* species.

Karyotype evolution and collinearity analyses in Zingiberaceae revealed pervasive inversions among *Curcuma* species

Various chromosome numbers and ploidy levels in *Curcuma* species motivated us to investigate the karyotype evolution in Zingiberaceae. First, we constructed the phylogenetic tree of seven Zingiberaceous species based on 1791 orthologous single-copy genes to infer karyotype evolution. It turned out *Curcuma* was sister to *Zingiber*, and *C. alismatifolia* was the early diverged species in the genus (Fig. 1b).

Next, 11 ancestral pseudo-chromosomes of Zingiberaceae were constructed with WGD software⁴² based on the inferred phylogeny and collinearity among *C. longa*, *C. petiolata*, *C. alismatifolia*, *Z. officinale*, and *Wurfbainia longiligularis* (*Lanxangia tsao-ko* and *W. villosa* were excluded owing to the poor genome quality) (Fig. 1b and Supplementary Fig. 6). The 1:2 collinearity between the ancestral chromosomes of Zingiberaceae and species in the family further validated a

common whole-genome duplication (WGD) event in Zingiberaceae (Fig. 1b and Supplementary Figs. 7 and 8). Meanwhile, compared with the ancestral chromosomes of Zingiberaceae, *Curcuma* species underwent at least 40 chromosome fissions and fusions to form the current karyotype (Fig. 1b). In addition, different parts of ancestral pseudo-chromosomes were retained among haplotypes of *Curcuma* genomes, suggesting that the chromosomes of *Curcuma* species were still in a state of dynamic change post species divergence (Supplementary Fig. 7).

To illustrate the karyotypic evolution in *Curcuma*, we analyzed all extant chromosomes in *Curcuma* containing ancestral pseudo-chromosome 11 as an example to illustrate possible fission, fusion and chromosome rearrangement events (Supplementary Fig. 8). The ancestral pseudo-chromosome 11 were present in multiple chromosomes of Zingiberaceous species (Supplementary Fig. 8 and 9),

Table 1 | Comparison of basic statistics for genome sequencing, assembly, and annotation of three *Curcuma* species

	<i>C. alismatifolia</i>			<i>C. petiolata</i>		<i>C. longa</i>			
	This study		Liao et al. ²⁹	This study		This study		Yin et al. ³⁷	
Haplotype	HapA	HapB	NA	HapA	HapB	HapA	HapB	HapC	NA
Number of chromosomes	16	16	16	21	21	21	21	21	21
Total length (Gbp)	0.95	0.97	0.99	1.04	1.04	0.70	0.72	0.74	1.11
GC%	39.31	39.32	39.98	39.32	39.31	38.81	38.83	38.90	40.18
Number of contigs	54	65	582	69	69	53	56	54	680
Scaffold N50 (Mbp)	59.67	61.96	57.51	52.23	51.60	33.54	34.06	36.15	50.12
Contig N50 (Mbp)	23.37		NA	25.81		25.27			2.34
Anchor rate (%)	94.70		NA	97.28		95.66			96.25
LAI	25.46		26.38	23.58		21.63			13.56
BUSCOs of genome (%)	98.55		96.53	98.50		98.80			95.20
Number of protein-coding genes	50,239	48,465	57,534	45,764	44,683	37,115	37,161	37,995	49,612
BUSCOs of protein set (%)	96.10	95.40	90.70	97.30	97.20	97.20	97.20	96.90	NA
Repeat sequences (%)	76.23		75.84	76.78		69.40			69.99

suggesting that the evolution of pseudo-chromosome 11 had included numerous chromosome fission and fusion events. In addition, an inversion was detected among chromosome 6 of *C. alismatifolia* and chromosomes 13, 15, and 8 of *C. petiolata* (Supplementary Fig. 9b), among chromosome 20 of *C. petiolata*, chromosome 8 of *C. longa* and chromosome 11 of *C. alismatifolia*, as well as between the two haplotypes of *C. petiolata* (Fig. 1c), which was verified by the conflict signals of Hi-C interaction (Fig. 1d).

The signals of frequent inversions in *Curcuma* motivated us to systematically identify inversions among the three *Curcuma* species via chromosome collinearity analysis based on 10 chromosomes with one-to-one correspondence. Inversions were identified between haplotypes from any two of the three *Curcuma* species as inter-specific inversions (named INTER hereafter). It revealed that the percentage of the total length of inter-specific inversion accounted for ~32% to 39% of the 10 chromosomes (Fig. 1e). The highest percentages appeared between *C. petiolata* and *C. alismatifolia*, accounting for ~17%–23% of the 10 chromosomes (Supplementary Fig. 10 and 11). Altogether, these findings suggested that chromosomal inversions frequently accompanied karyotype evolution in *Curcuma* species.

Pervasive inversions between haplotypes within *Curcuma* species and among populations of *C. alismatifolia*

For highly heterozygous horticultural species, SV between haplotypes within species has often been neglected or poorly studied at a basic level such as describing numbers of structure variations. To further explore whether inversions, especially between haplotypes, had an impact on the evolution of *Curcuma* species, we first characterized the inversions between haplotypes within *Curcuma* species (list INTRA; Fig. 2a). We found 63 inversions in *C. alismatifolia*, 47 inversions in *C. petiolata*, and 54 (HapA/HapB), 49 (HapA/HapC), and 59 (HapB/HapC) inversions in *C. longa*, respectively (Supplementary Data 2 and 3). The existence of these inversions was further confirmed by conflict signals of Hi-C interaction (Supplementary Fig. 12) and the intact/split HiFi reads mapping at the inversion boundary (Supplementary Fig. 13). The percentage of inversions between the two haplotypes of *C. alismatifolia* (HapA: 14.52%, 138,137,483 bp; HapB: 15.65%, 152,306,047 bp) was particularly high (Supplementary Data 2 and 3), and 16 inversions exceeded 1 Mbp in length, with 5 of them more than 10 Mbp (Supplementary Data 2 and 3).

Furthermore, to evaluate the prevalence of these inversions in *Curcuma* species, we compared the percentage of nucleotides within inversions in *Curcuma* with that of the published haplotype-resolved genome of representative diploid and allotetraploid species

(Supplementary Table 2). We found that the length percentage of INTRA inversions in *Curcuma* species was approximately 30 times higher than in *Camellia sinensis*, a diploid species (Theaceae) with a high proportion of translocations, and 8 times higher than that in *Vitis vinifera*⁴³, a diploid perennial species (Vitaceae) that mainly relied on asexual reproduction. Moreover, this percentage even exceeded that observed in *Oryza alta*, an allotetraploid species (Fig. 2b). Additionally, ~3000–5000 genes were located within inversions of Zingiberoideae species, representing a much higher percentage compared to other diploid species (Fig. 2c).

To further investigate the prevalence of inversions at the population level, we generated a genome-wide variation map containing >1.51 million SNPs from 56 *C. alismatifolia* cultivars based on whole genome resequencing data. First, the local PCA analysis was performed to identify genomic regions with abnormally high differentiation, which were typically considered potential inversion polymorphisms at the population level. It turned out that almost all INTRA inversions in *C. alismatifolia* overlapped with these highly differentiated regions (Supplementary Fig. 14), suggesting these inversions affected local population structures. Second, three software were utilized to detect SVs among populations of *C. alismatifolia*, and inversions predicted by at least two software were compared with those inversions between two haplotypes of *C. alismatifolia*. When the length of overlapped regions was over 50% of the length of corresponding inversions in the two lists, they were treated as the same inversion. It revealed that 33 (list POP) out of 63 inversions between haplotypes were present at the population level. The average frequency of these inversions was 18.67%, and shorter inversions exhibited higher frequency (Supplementary Fig. 15 and Supplementary Data 4).

Genes within inversions exhibited lower expression level, higher nucleotide substitution rates, and more frequent LTR insertions

We next evaluated how inversions affected gene evolution in the three inversion lists (INTER, INTRA, and POP) and found that gene pairs within inversions in *C. longa* exhibited significantly lower gene expression levels (two-sided Wilcoxon rank-sum test, $p < 0.001$) but higher nucleotide substitution rates (Ka/Ks , nonsynonymous substitution rates/synonymous substitution rates; two-sided Wilcoxon rank-sum test, $p < 0.001$) and LTR content in the 3 Kbp flanking regions of genes (two-sided Wilcoxon rank-sum test, $p = 0.189$) compared to the genomic background (Fig. 3a). The pattern held with the different inversion lists in the three *Curcuma* species (Fig. 3a), while inversions in *Zingiber officinale* showed an opposite trend (Supplementary Fig. 16).

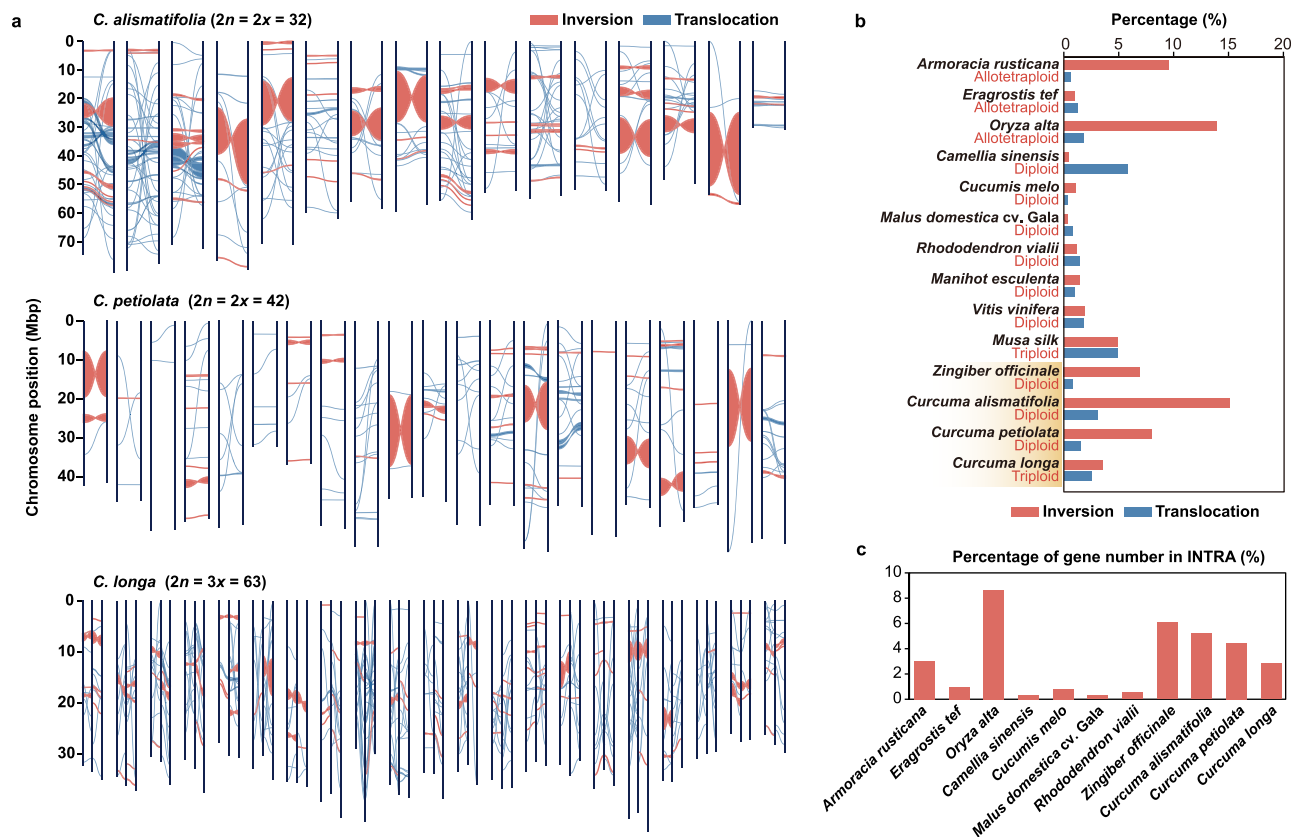


Fig. 2 | Pervasive inversions detected between haplotypes within species. **a** Inversions and translocations between haplotypes within each of the three *Curcuma* species. **b** Percentages of total length of inversions and translocations among

species with different ploidy levels. **c** Percentages of genes in inversions. INTRA: inversions between haplotypes within species.

To identify the potential causes of the low gene expression and high nucleotide substitution rate within inversion regions, such as chromosomal SV or TE insertion, we compared the statistics between genes within the inversion and in the high LTR region (where the LTR content was higher than the average LTR content of inversion). We found that the observed lower gene expression and higher nucleotide substitution rates in inversions persisted even when compared to genomic regions with higher LTR content (Fig. 3b). Interestingly, this pattern was restricted to LTRs, but not other categories of TEs, whose distribution patterns varied across the three *Curcuma* species (Supplementary Fig. 17). As *C. alismatifolia* demonstrated the highest percentage of intra-specific inversions, we explored how gene expression level, nucleotide substitution rate and LTR content varied when moving away from the inversion breakpoints and discovered that gene expression level was gradually increased ($R^2 = 0.91$, $p = 0.001$, Supplementary Data 5 and 6), while nucleotide substitution rate ($R^2 = 0.84$, $p = 0.004$) and LTR content ($R^2 = 0.45$, $p = 0.102$) were gradually decreased to the genomic background level (Fig. 3c–e).

Inversions within populations of *C. alismatifolia* were enriched with deleterious mutations and depleted with selected SNPs

Population structure analyses revealed that *C. alismatifolia* cultivars could be divided into two main groups, g1 and g2. Individuals in the g1 and g2 groups differed in several traits, such as inflorescence morphology, plant height, and leaf traits (Fig. 4a, Supplementary Figs. 18 and 19, Supplementary Data 7), which were likely polygenic traits as shown in model plant species^{44–46}. Thus, *C. alismatifolia* cultivars provided us with suitable contrasting populations to detect signals of selection and evaluate the prevalence of selected SNPs across the genome.

Subsequently, we detected SNPs under positive selection between the same number (seven) of non-admixed g1-L and g2-L samples (individuals retained after excluding heterozygous individuals based on population structure and morphological characteristics) via three approaches, including the cross-population composite likelihood ratio (XP-CLR), number of segregating sites by length (nSL), and the method combining both nucleotide diversity (π) and population divergence (F_{ST}) (Fig. 4b and Supplementary Figs. 20–25). We found that inversions were significantly depleted with selected SNPs in all three approaches (two-sided χ^2 test, $p < 0.001$, Supplementary Data 8) (Fig. 4b and Supplementary Table 3). However, we could not rule out that a few inversions might be subject to selection, such as the 16-19167948-19763037 inversion with lower π and the 6-5019267-5216546 inversion with higher F_{ST} compared with genomic background (Fig. 4 and Supplementary Fig. 26).

Next, we found the polymorphism information content (PIC) of multi-allelic SNPs (allele number ≥ 2) in genes within the inversion region was significantly higher than that in genes outside of inversions (two-sided Wilcoxon rank-sum test, $p < 0.001$). Fewer selected SNPs and higher PIC of SNPs in genes within the inversion regions implied the inversion regions contained enough polymorphic sites but were less likely to be selected (Fig. 4c).

Finally, we examined the recombination rate across the genome of *C. alismatifolia* populations and found it was significantly lower in the inversion regions than that of genomic background and even than that of genomic regions with high LTRs (two-sided Wilcoxon rank-sum test, $p < 0.001$) (Fig. 4d), implying the reduced crossover events and strengthened linkages within inversion regions. The count of deleterious mutations in *C. alismatifolia* was assessed in both additive and recessive modes. It indicated that the inversion

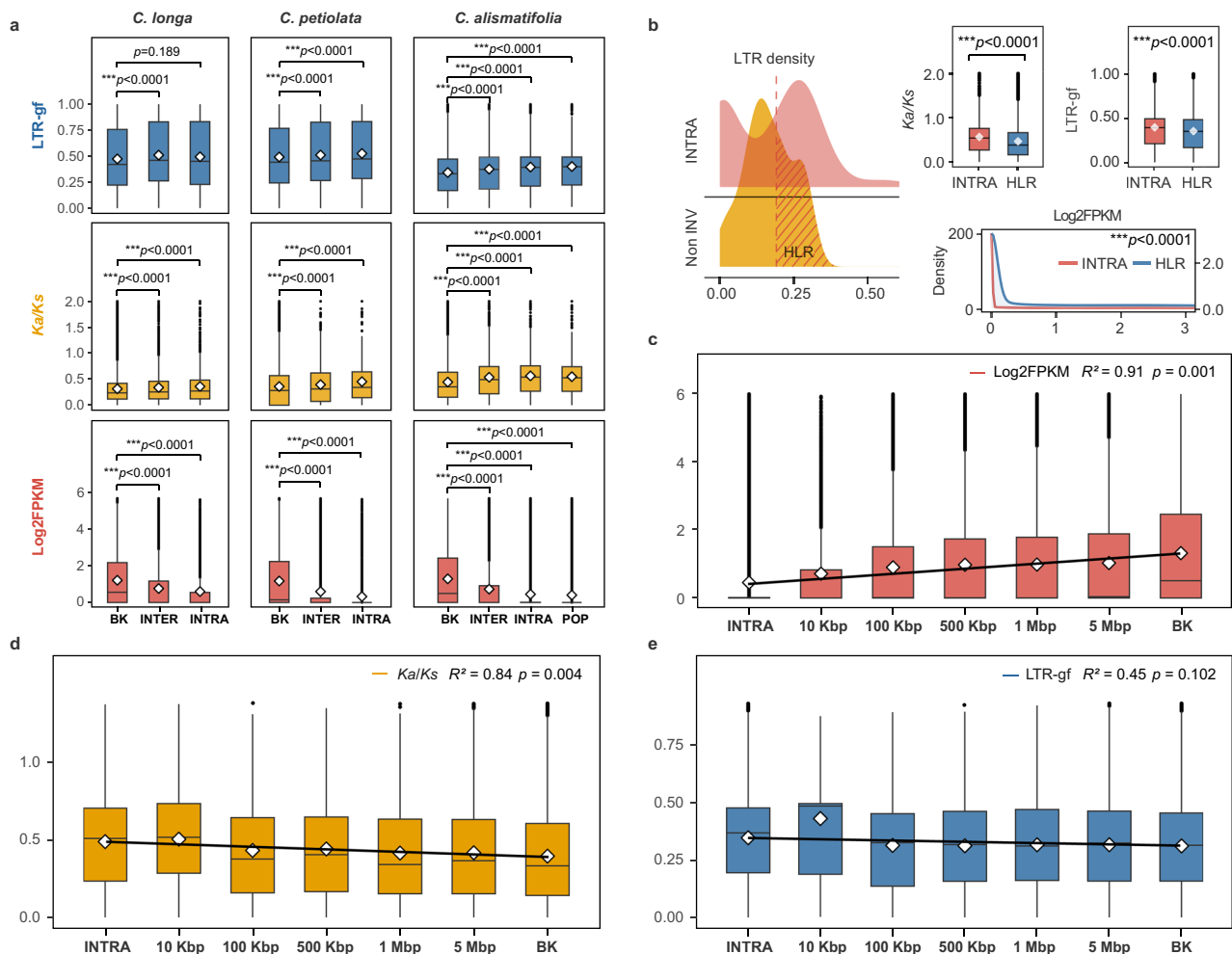


Fig. 3 | Genomic impacts of inversions in *Curcuma* species. **a** Comparison of gene expression, nucleotide substitution rates, and LTR content between inversions and genome-wide background in the three *Curcuma* species. **b** Comparison of gene expression, nucleotide substitution rates, and LTR content between inversions and high LTR regions in *C. alismatifolia*. The red dashed line indicated the mean LTR content in inversions (window size = 500 Kbp). The red shaded area indicated the non-inverted region where the LTR content exceeded the mean LTR content in the INTRA. The rhombuses represented the mean. The left y axis of the density map corresponded to INTRA, and the right y axis corresponded to HLR. **c** Linear regression of gene expression levels ($R^2 = 0.91$, $p = 0.001$) among genomic regions when departing away from inversions in *C. alismatifolia*. **d** Linear regression of nucleotide substitution rates ($R^2 = 0.84$, $p = 0.004$) among genomic regions when departing away from inversions in *C. alismatifolia*. **e** Linear regression of LTR

content ($R^2 = 0.45$, $p = 0.102$) among genomic regions when departing away from inversions in *C. alismatifolia*. **a–e** The box-plot elements were defined as: center line, median; box limits, upper and lower quartiles; whiskers, 1.5 \times interquartile range; points, outliers; rhombuses, mean. **a, b** Two-sided Wilcoxon rank-sum test was conducted for significance evaluation, and multiple comparisons were adjusted with the Bonferroni correction. Asterisks represented significant differences ($***p < 0.001$, adjusted, Supplementary Data 5). **c–e** The R^2 and p values were shown. FPKM fragments per kilobase of transcript per million fragments mapped, K_a nonsynonymous substitution rate, K_s synonymous substitution rate, LTR-gf The content of LTR in the 3 Kbp flanking regions of genes, INTRA inversions between haplotypes within species, INTER inter-specific inversions among the three species, POP inversions among populations of *C. alismatifolia*. BK genome-wide background. HLR high LTR regions. Source data are provided as a Source Data file.

regions accumulated more deleterious mutations in both modes (two-sided Wilcoxon rank-sum test, $p < 0.001$) (Fig. 4e), and the suppressed recombination failed at efficient removal of deleterious mutations in inversion regions. In addition, we detected significantly more deleterious mutations in the inversion “CaChr8-TPS-TD-INV” (8-10517073-28214380, the longest inversion on chromosome 8 of *C. alismatifolia*) compared to genomic background (Fig. 4e), which was heterokaryotype in the surveyed individuals, including nine *C. alismatifolia* individuals and two other *Curcuma* species (Supplementary Fig. 27, Supplementary Data 4). Furthermore, when the 33 inversions (list POP) were divided into two groups based on whether the two arrangements of inversions were heterokaryotypes (HETE) or homokaryotypes (HOMO), it showed HETE inversions (including CaChr8-TPS-TD-INV) in *C. alismatifolia* accumulated more deleterious mutations than HOMO inversions (Supplementary Fig. 28,

Supplementary Data 4). Taken together, all the above evidence suggested that inversions within *C. alismatifolia* suppressed recombination in heterokaryotype inversions, leading to the accumulation of deleterious mutations, and contained enough polymorphic sites, which were less likely to be selected (Fig. 4).

Inversions affected terpenoid biosynthesis via pseudogenization of tandemly duplicated TPS genes

To further determine how inversions affected functions or traits, we performed KEGG (Kyoto Encyclopedia of Genes and Genomes) enrichment analysis of genes within inversions, and found that the genes within inversions of three *Curcuma* species were significantly enriched in the terpene biosynthesis pathway (Fig. 5a and Supplementary Data 9), which was not detected in *Z. officinale* (Fig. 5a). To exclude the possibility that TPS genes are also enriched in other

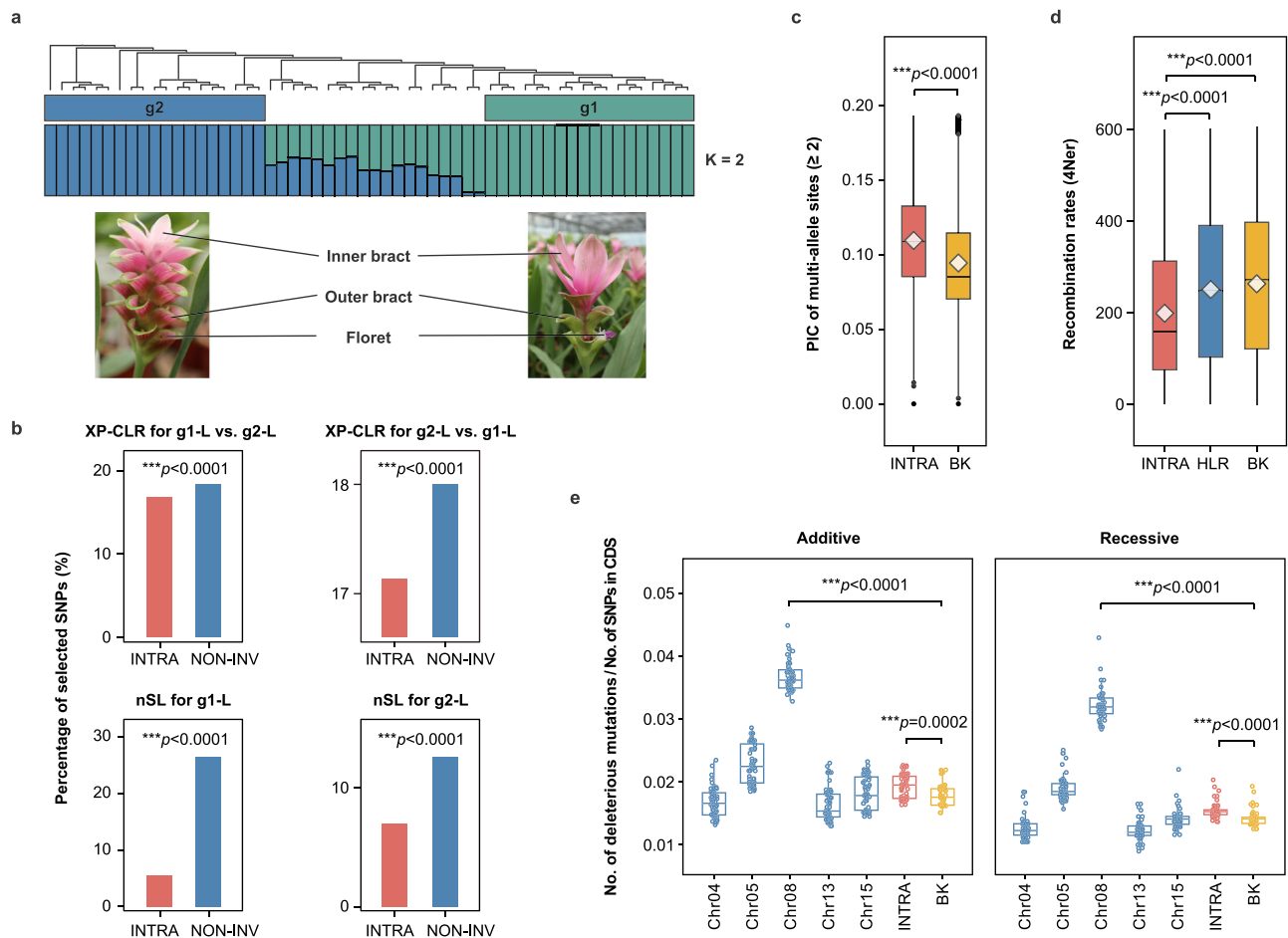


Fig. 4 | Inversions were depleted with selected SNPs in *C. alismatifolia*.

a Population structure of 56 *C. alismatifolia* cultivars at $K = 2$. g1 and g2 represented two groups with large differences in multiple traits. **b** Comparison of the proportion of selected SNPs (with 95% percentile as threshold for outliers; two-sided χ^2 test, $***p < 0.001$, Supplementary Data 8) between inversions and non-inversions based on XP-CLR (top), nSL analysis (down). **c** The polymorphism information content (PIC) of multi-allelic SNPs (≥ 2 alleles per site) in genes. **d** Comparison of recombination rates among inversions, high LTR regions, and genome-wide background in *C. alismatifolia*. **e** The box-plot elements are defined as: center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range; points, outliers; rhombuses, mean. **e** Comparison of percentage of deleterious mutations between inversions and genome-wide backgrounds in *C. alismatifolia*. Deleterious mutations of the top five longest inversions were also shown,

where Chr08 stood for “CaChr8-TPS-TD-INV” (8-10517073-28214380), the longest inversion on *C. alismatifolia* chromosome 8. Additive: the additive mode with deleterious alleles in both homozygous and heterozygous genotypes; Recessive: the recessive mode with deleterious alleles only in homozygous genotypes. CDS: coding sequence. Box-plot elements are defined as: center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range; points, all values, $n = 56$. INTRA inversions between haplotypes within species, NON-INV non-inversion region, BK genome-wide background, HLR high LTR regions. **c–e** two-sided Wilcoxon rank-sum test was employed for significance evaluation, and multiple comparisons were adjusted with the Bonferroni correction. Asterisks represented significant differences ($***p < 0.001$, adjusted, Supplementary Data 5). Source data are provided as a Source Data file.

genomic regions, we randomly selected 5133 genes (the number of genes contained in inversions) for KEGG enrichment in *C. alismatifolia* and repeated it five times. It turned out that the enriched items did not contain anything related to terpenoid biosynthesis pathways (Supplementary Fig. 29), suggesting that inversions in *Curcuma* might affect the content of terpenes. Next, we examined the genes that were enriched in terpene biosynthesis pathway, and found three tandem duplications of *TPS-a* genes (Supplementary Table 4, Supplementary Data 10, Supplementary Figs. 30 and 31). The *TPS* gene family of angiosperms could be divided into seven major groups: *TPS-a*, *TPS-b*, *TPS-c*, *TPS-d*, *TPS-e/f*, *TPS-g*, and *TPS-h*, of which the *TPS-a* mainly controlled the synthesis of sesquiterpenes in angiosperms⁴⁷. Based on gene annotation and phylogenetic analyses, the genes on chromosome 8 of *C. alismatifolia* and on chromosome 14 of *C. petiolata* were identified as the best hits for germacrene synthase and spanned both interior and exterior of inversion (Supplementary Figs. 32 and 33, Supplementary Table 4, Supplementary Data 9–12).

The *TPS* genes within the inversion on chromosome 8 of *C. alismatifolia* showed incomplete gene domains and the majority were not expressed (FPKM < 1) in 37 samples representing a complete set of tissues and developmental stages (except *Chr08HA1044* and *Chr08HB723*), whereas genes outside of the inversion retained complete domains and were normally expressed (Fig. 5b, Supplementary Fig. 33a, Supplementary Data 9). The incomplete genes within the inversion regions on the two haplotypes of chromosome 8 of *C. alismatifolia* lost different domains (Supplementary Fig. 33a), implying the independent evolution of the two haplotypes. The phenomenon was also observed in *C. petiolata*, where orthologous incomplete *TPS* genes within inversion regions on chromosome 14 were not expressed. The expression profile of all *TPS*s revealed that compared to *TPS*s outside of the inversion, *TPS*s within inversions had higher *Ka/Ks* and lower expression in *C. alismatifolia* (Supplementary Fig. 33b, Supplementary Data 9). 52.63% inversion-internal *TPS* genes (20/38) were not expressed (FPKM < 1) in 37 samples, and only 20.00% *TPS* genes

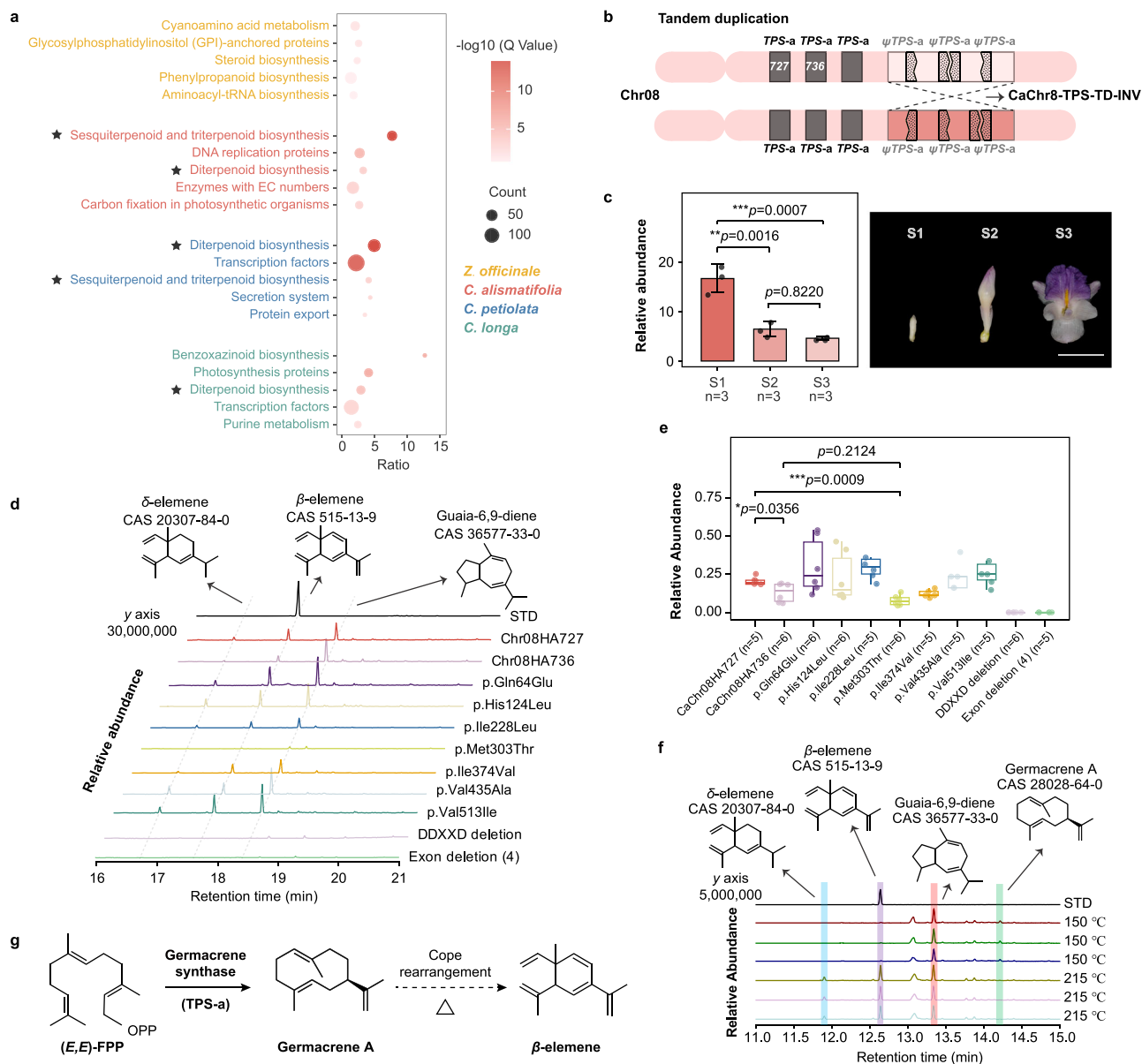


Fig. 5 | An inversion disabled function of tandemly duplicated *TPS* genes and impacted terpenoid biosynthesis. **a** KEGG enrichment of genes within inversions in three *Curcuma* species and *Z. officinale* performed with TBtools. Enrichment of KEGG pathways was calculated with the Hypergeometric test. Statistical tests were one-sided and multiple comparisons were adjusted with the Benjamini-Hochberg correction (Q value). **b** Schematic diagram showing tandemly duplicated *TPS*s on chromosome 8 (Chr08) of *C. alismatifolia*. Genes outside of the inversion maintained the intact structure (dark gray rectangle), and genes within the inversion were pseudogenes (hatching chevron). Genes within inversions on different haplotypes lost different domains. ψ : pseudogene (gray font); CaChr8-TPS-TD-INV: Inversion on *C. alismatifolia* Chr08 with tandem duplicated *TPS* genes spanning its inversion breakpoint (8-10517073-28214380), shown as the intersecting dashed lines. Light and dark pink boxes indicated inverted regions on two haplotypes. **c** Compound **1** content detected by GC-MS and floret phenotype at three developmental stages of florets. The bar indicated 1 cm. Error bars represent the standard deviations (SD), and data are presented as mean values \pm SD. Three biological replicates were performed. **d** GC-MS analysis of the main products formed by prokaryotic expression of germacrene synthases, compound **1** ($m/z = 93$), compound **3** ($m/z = 93$), and compound **4** ($m/z = 93$). Enzymes were incubated with compound **5**. Reaction products were identified by standard chemicals and comparison of their mass spectra and retention indices with authentic standards and

NIST libraries. Empty vector, pCold-TF. p.Gln64Glu, p.His124Leu, p.Ile228Leu, p.Met303Thr, p.Ile374Val, p.Val435Ala, and p.Val513Ile indicated seven different amino acid sites between Chr08HA727 and Chr08HA736 were mutated one by one accordingly. The direction of mutation is from Chr08HA727 to Chr08HA736. DDXXD deletion: DDXXD domain was deleted. Exon deletion (4): The sequence of exon 4 (containing the DDXXD domain) was deleted. All indicated assays were conducted for four or more repetitions. **e** Relative catalytic activity of germacrene synthase genes in producing compound **1**. Box-plot elements are defined as: center line, median; box limits, upper and lower quartiles; whiskers, 1.5 \times interquartile range; points, all values. 5–6 biological replicates were performed for each treatment. **f** GC-MS analysis of the products from prokaryotic expression of Chr08HA727 proteins at different injection temperatures. Reaction products were identified by comparison of their mass spectra and retention indices with authentic standards and NIST libraries. Empty vector, pCold-TF. When the injection temperature was lowered to 150 $^{\circ}$ C, the content of compound **1** was decreased, and compound **2** was increased significantly ($p < 0.01$). **g** Catalytic scheme of germacrene synthase. The triangle represented the cope rearrangement of compound **2** to form compound **1** that easily occurred at high injection temperatures. **c**, **e** Statistical tests were two-sided Student's t -test, and multiple comparisons were adjusted with the Bonferroni correction. Asterisks represented significant differences ($*p < 0.05$, $**p < 0.01$, $***p < 0.001$, adjusted).

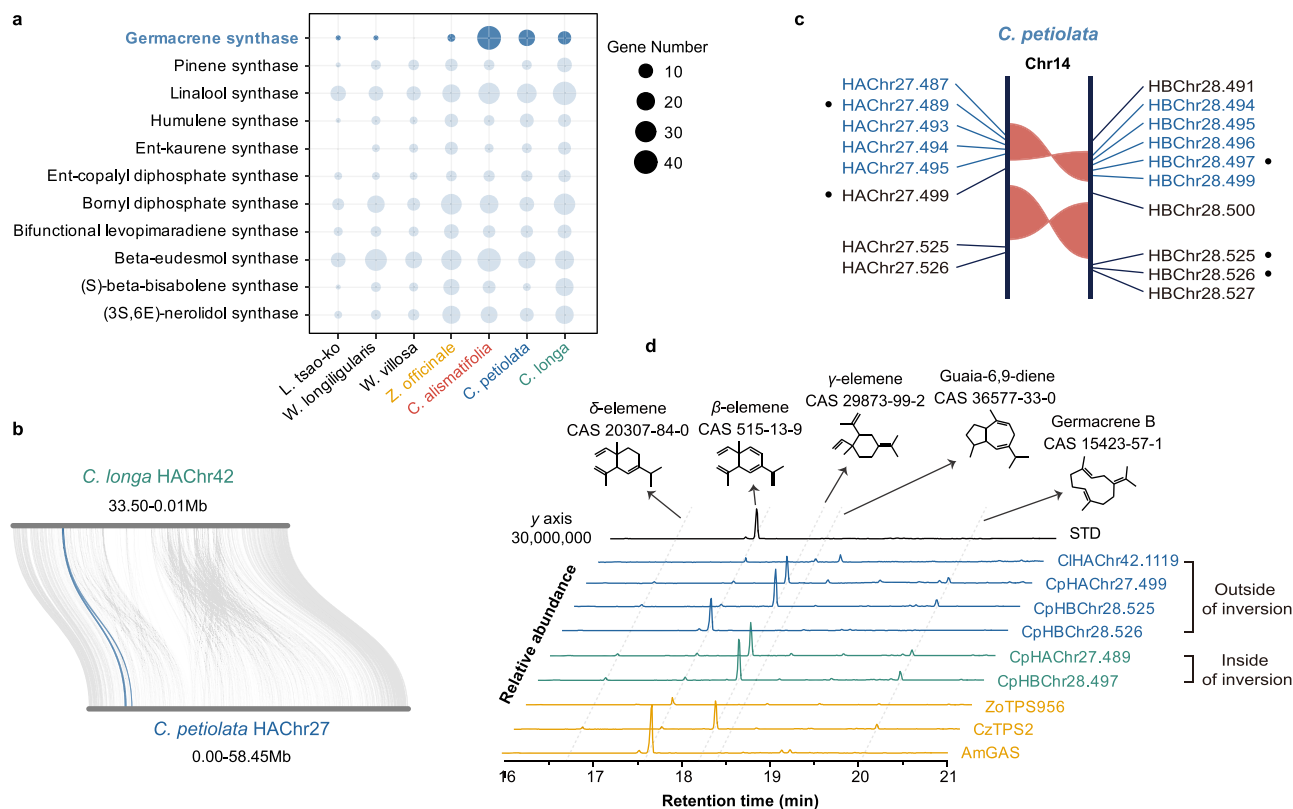


Fig. 6 | The evolution of germacrene synthase genes in Zingiberaceous species.

a Annotated functions of *TPS* genes in Zingiberaceous species in the NR (Non-Redundant Protein Sequence) database revealed that the annotated germacrene synthase genes were expanded in *Curcuma*. **b** Collinearity of tandem duplicated genes of germacrene synthase (blue) in *C. petiolata* and *C. longa*. **c** Germacrene

synthase tandem duplication genes spanning inversions on *C. petiolata* chromosome 14 (Chr14). Inversions between haplotypes were marked in red. Germacrene synthase genes in the inversion were labeled blue. Black dots denoted the genes with the complete gene structure. **d** GC-MS analysis of the products from prokaryotic expression of intact germacrene synthases in Zingiberaceae species.

(35/175) outside of the inversion were not expressed (two-sided χ^2 test, $p < 0.001$). Further investigation of gene expression data revealed that 18 *TPS* genes with complete gene structure were expressed highly in bracts and florets and lowly in rhizomes of *C. alismatifolia* (Supplementary Data 9), and these genes were down-regulated during floret development, consistent with the content levels of 1- β -elemene (C15H24), an important anticancer substance that was easily formed by cope rearrangement of 2-Germacrene A at high temperatures (Fig. 5c, Supplementary Figs. 33a and 34, Supplementary Tables 5 and 6, Supplementary Data 9, Liu et al.⁴⁸).

We functionally validated two representatives of these genes, *Chr08HA727* and *Chr08HA736*, which had a 7 amino acid difference (Supplementary Fig. 35). The main catalytic products of *Chr08HA727* and *Chr08HA736* included compound **1**, 3- δ -elemene (C15H24), and 4-Guaia-6,9-diene ($m/z = 93$), of which the compound **1** catalytic efficiency of *Chr08HA727* was 1.79-fold (two-sided Student's *t*-test, $p < 0.05$) higher than that of *Chr08HA736* (Fig. 5d, e, Supplementary Fig. 36, Supplementary Table 7, Supplementary Data 13). In addition, when the injection temperature of GC-MS was lowered from 215 °C to 150 °C, products of *Chr08HA727* changed from non-detection of compound **2** (CAS, 28028-64-0, the fragment ion was consistent with that of the published compound **2**⁴⁹, Supplementary Fig. 37) to traces of compound **2**, whereas compound **1** content was reduced 14.31-fold (two-sided Student's *t* test, $p < 0.001$) (Supplementary Fig. 37b), implying that *Chr08HA727* and *Chr08HA736* were germacrene synthases and the observed compound **1** was formed by the conversion of compound **2** at high temperatures (Fig. 5f, g, Supplementary Fig. 37, Supplementary Table 8). Moreover, we mutated the seven amino acids from that in *Chr08HA727* to that in *Chr08HA736* and evaluated the

catalytic activities of mutants. It showed that the catalytic activity of *Chr08HA727* was reduced by 3.63-fold (two-sided Student's *t* test, $p < 0.001$) when p.Met303Thr located close to the DDXXD structural domain (essential for terpenoid biosynthesis⁵⁰) was mutated, suggesting the significance of this locus in the catalytic activity (Fig. 5d, e, Supplementary Fig. 36a, Supplementary Table 7, Supplementary Data 13). The kinetic parameters of *Chr08HA727*, *Chr08HA736*, and p.Met303Thr indicated that *Chr08HA727* had the highest catalytic efficiency in producing compound **1** ($K_{cat}/K_m = 0.0883 \mu\text{M}^{-1} \text{min}^{-1}$) (Supplementary Fig. 36b).

In addition, most of the germacrene synthase genes within the inversion of chromosome 8 lost exon 4, which contains the DDXXD domain. Subsequently, we synthesized sequences lacking the DDXXD domain and exon 4, respectively (Fig. 5d, e, Supplementary Fig. 36a, Supplementary Table 7, Supplementary Data 13) and evaluated the function of the synthesized pseudogenes. Not surprisingly, they were no longer catalytically active, indicating that the function of the germacrene synthase within the inversion was impaired and catalytically inactivated.

Evolutionary history and dynamics of *TPS* tandem duplication in Zingiberaceae

To further explore whether and when the tandem duplicated germacrene synthase genes in *C. alismatifolia* occurred in other Zingiberaceous species, we first characterized all *TPS*s and classified them based on their functions in Zingiberaceae. It showed that germacrene synthase genes were significantly expanded in *Curcuma* species (Fig. 6a, Supplementary Table 9, Supplementary Data 14). Another tandem duplication of germacrene synthase genes was also detected in

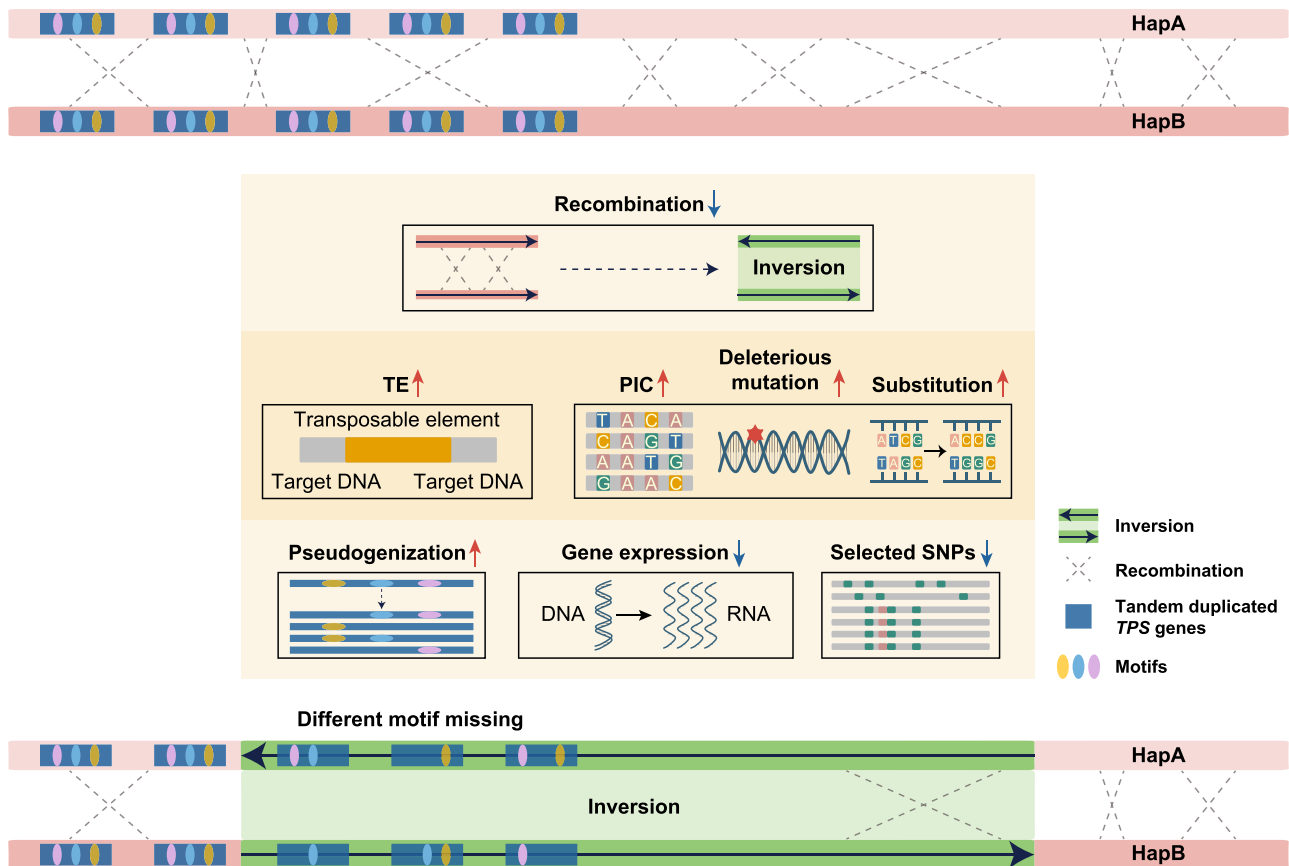


Fig. 7 | A schematic model depicting how inversions underwent relaxed genetic constraints and balanced birth and death of *TPS* genes in *Curcuma* species.

Inversions acted as a counteracting evolutionary force of gene tandem duplication to keep a balance between gene birth and death of *TPS* genes, and genomic characteristics associated with inversions indicated relaxed genetic constraints.

Recombination within inversions between haplotypes was suppressed, resulting in inefficient removal of TEs and deleterious mutations. Higher levels of TE content,

deleterious mutations, PICs, and base substitutions were observed within inversions. Genes within inversions were barely expressed, and inversion-internal genes on different haplotypes lost different domains, implying that these genes were pseudogenes and the two haplotypes evolved independently. And fewer selected SNPs were detected within inversions. Red and blue arrows indicated an increase and decrease within inversions relative to the genome-wide background, respectively.

collinear regions between *C. longa* and *C. petiolata* that spanned the inside and outside of the inversion (Fig. 6b, c; Supplementary Fig. 38), suggesting that the tandem duplication of germacrene synthase genes might occur at the common ancestor of *Curcuma* (Fig. 6a). Functional verification of the intact germacrene synthase genes of Zingiberoideae species, showed that genes in both *Z. officinale* and *Curcuma* possessed catalytic activity of converting 5-(*E,E*)-farnesyl diphosphate ((*E,E*)-FPP) into compound **2** and converted into compound **1** at high temperatures (Fig. 6d and Supplementary Figs. 39–41).

Finally, we have summarized our findings in a brief model illustrating how inversions affected genomic features (Fig. 7). Inversions broke the tandem duplicated *TPS*s, and caused genes within inversions to experience reduced genetic constraints. Specifically, recombination within inversions was inhibited, leading to inefficient removal of TEs and deleterious mutations, which further caused pseudogenization and the absence or decrease of gene expression within inversions. Despite sufficient polymorphic sites within inversions, SNPs within inversions were less likely to be selected. Taken together, the inversion caused the pseudogenization of the tandem duplicated germacrene synthase genes within the inversion, which maintained a balance between the birth and death of terpenoid synthase genes.

Discussion

Our study represented an extensive investigation of inversions in a plant genus and its evolutionary and functional effects. We detected

enormous inversions in *Curcuma*, and found they suffered from relaxed genetic constraints, allowing genes within inversions to exhibit higher nucleotide substitution rates, more deleterious mutations and TE contents, and lower expression levels. Meanwhile, SNPs within inversions were less likely to be under positive selection between two populations of *C. alismatifolia* differentiating at multiple traits. In addition, an explicit example of tandem duplicated *TPS* genes spanning inversion breakpoints showed that *TPS* genes within inversion were pseudogenes, and their function was disabled, while *TPS* genes outside of inversion were functional in catalyzing compound **5** into compound **2**, which was converted into compound **1** at high temperatures.

Distinct from numerous empirical studies that demonstrated the advantageous role of inversions as they preserved combinations of multiple adaptive loci via suppressed recombination^{6,8–10}, our study provided an empirical example that inversions experienced relaxed genetic constraints and were less likely under positive selection in *Curcuma*. On the other side of the coin, the suppressed recombination of inversions led to inefficient removal of TEs and deleterious mutations (Fig. 4), which could also be exaggerated by the asexual reproduction of *Curcuma* species. Meanwhile, the accumulation of TEs and deleterious mutations within inversions might lead to reduced expression levels and pseudogenization of genes within inversions in *Curcuma* (Figs. 3, 5, and 6). Our study uncovered the extensive impacts of inversions on all genes within inversion

regions, rather than restricting to its impacts on genes at breakpoints^{7,21–23}.

How SVs impacted biochemical traits, especially the diversity and content of secondary metabolites, was a rarely addressed field⁵¹. Our study originally shed light on the effect of inversions on terpenoid biosynthesis. As previously reported, *TPS* genes acted as metabolic gatekeepers of terpenoid diversity in plants⁴⁷, and tandem duplication events were common in the *TPS* gene family^{52,53}. However, *TPS* gene family expansion through tandem duplication was often accompanied by the lack of expression of most *TPS* genes, and the underlying genetic mechanism was yet unknown⁵². In *C. petiolata* and *C. alismatifolia*, we found *TPS* tandem duplicated genes spanning inversion breakpoints, and they were almost pseudogenized within inversion but retained complete structure and germacrene synthase activity outside of the inversion. Our study provided an empirical example of how inversion coupled with tandem duplications affected terpenoid accumulations, which filled in a gap in how SVs influence biochemical traits.

The evolutionary history of inversions has been mysterious. As for the origin of CaChr8-TPS-TD-INV, given that the tandem duplication of germacrene synthases was also detected in *C. longa* and *C. petiolata* but absent in *Z. officinale*, it was possible that the tandem duplicated germacrene synthases genes were originated in the common ancestor of *Curcuma* species. Meanwhile, haplotypes of CaChr8-TPS-TD-INV in *C. alismatifolia* lost different domains of *TPS* genes within the inversions, suggesting independent evolution of the two haplotypes. The above evidence stimulated us to propose that this inversion occurred after the tandem duplication event.

In fact, whether this *TPS*-related inversion is adaptive, deleterious, or neutral still depends on its interaction with the environment. In a mild environment, growth is often prioritized over secondary metabolite production regarding efficient resource allocation. In such a case, the dramatic expansion of *TPS* tandem duplicated genes possibly caused the exhausted supply of substrates, and the inversion acted as a “stop/pause” mechanism of terpenoid production by pseudogenization of gene copies within inversions. However, given that terpenes were essential for the resistance and adaptation of plants⁴⁷, in an adverse environment where enough terpenoids were beneficial for plants’ survival and fitness, inversions reducing terpenoid production played a negative role in plants’ adaptivity. Given the variability of environments following the establishment of the inversion, the real situation might be a mixture of multiple processes, and the relative role of selection, drift, mutation, and recombination might be dynamically variable over the lifetime of the inversion and within its distributional area².

Frequent occurrence of inversions in *Curcuma* species implied that some regions on chromosomes in the genus might be relatively susceptible to being broken. To predict inversion hotspots based on sequence and structural features with machine learning models might deserve further investigation. Such exploration, together with the identification of inversions affecting important phenotypic and/or biochemical traits, will provide insights into candidate sites for genome editing. Recent studies have demonstrated that not only genes but also entire chromosomes could be engineered with CRISPR-Cas9 technology to develop desirable traits in both *Arabidopsis* and maize^{54,55}. Therefore, future breeding efforts of economic plants are likely to involve the creation of cultivars with inversions underlying significant traits.

Methods

Genome assembly

The plants of *C. alismatifolia*, *C. petiolata*, and *C. longa* were cultivated in a greenhouse of the Institute of Agricultural Genomics at Shenzhen, Chinese Academy of Agricultural Sciences. DNA was extracted from the young leaves based on the Qiagen Genomic DNA kit (QIAGEN) for

PacBio HiFi, MGI-SEQ 2000, and Hi-C data sequencing. To well annotate the genome, total RNA was extracted from flowers, leaves, and young stems and sequenced on the MGI-SEQ 2000 platform. Quality control was conducted in Fastp v0.20.1⁵⁶. The genomes of three *Curcuma* species (*C. alismatifolia*, *C. petiolata*, and *C. longa*) were assembled using Hifiasm v0.16.1-r375⁵⁷. Chromosome-level assemblies were obtained after the confirmation of misjoins in contigs through Juicer v1.9.9⁵⁸ and 3D-DNA v180922⁵⁹. All details are available in Supplementary Note 1. The final distinction between HapA and HapB of *C. alismatifolia* and *C. petiolata* was based on the division of HapA and HapB in the initial assembly result of Hifiasm 0.16.1-r375⁵⁷. Meanwhile, the initial splitting of the haplotypes of *C. longa* was based on the separation in Subphaser⁶⁰. Furthermore, we calculated the switch error (a phasing error between the two haplotypes) taking the final HapA and HapB genome of *C. alismatifolia* as an example in calc_switchErr³⁸.

Genome annotation

Repeat annotation was implemented in EDTA v2.0.1⁴¹ and GETA was utilized to integrate annotations from homology-based, transcriptome-based, and ab initio predictions with a parameter setting of ‘max-intronlen 50000’ for gene structure annotations. Additional details are available in Supplementary Note 2. Next, the completeness of assembled genomes was evaluated with both BUSCO v5.3.2⁶¹ in the embryophyta_odb10 database and the LAI statistic in LTR_retriever v2.9.0^{39,40}. GO and KEGG functional enrichment analyses were conducted by TBtools v1.120⁶², based on the annotation results of eggno-mapper v2.0.1^{63,64}.

Phylogenetic relationships and ancestral chromosome reconstruction in Zingiberaceae

To reconstruct the phylogenetic tree of Zingiberaceae, a total of seven species, including *C. longa*, *C. petiolata*, *C. alismatifolia*, *W. villosa*⁶⁵, *L. tsao-ko*⁶⁶, *Z. officinale*⁶⁷, and *W. longiligularis*⁶⁸ were selected, with *Musa acuminata* (<https://banana-genome-hub.southgreen.fr/>) as an outgroup. To infer phylogenetic relationships, orthologous genes were obtained with OrthoFinder v2.5.2 (<https://github.com/davidemms/OrthoFinder>), and finally, a total of 1791 orthologous single-copy genes were used. Sequence alignments for each single-copy gene family were performed with MAFFT v7.464⁶⁹, and these alignments were converted into a nucleotide matrix by PAL2NAL v14⁷⁰. Under the GTRGAMMA model, a phylogenetic tree was constructed with RAxML v8.2.12⁷¹. To infer the ancestral karyotype of Zingiberaceae, *C. longa*, *C. petiolata*, *C. alismatifolia*, and *W. longiligularis*, along with their haplotypes, were compared with haplotype A of *Z. officinale*. Collinearity between *Curcuma* species was obtained with WGDI⁴², and further analyzed in JCVI v1.3.5⁷². Additional details are available in Supplementary Note 3.

Identification of SVs

To detect the SVs between haplotypes among *Curcuma* and several representative diploids and allotetraploid species in angiosperms. The whole-genome alignments were performed with Nucmer v4.0.0beta2⁷³, and the SVs were detected in SyRI v1.6⁷⁴ (Supplementary Fig. 42 and Supplementary Table 10). All details are available in Supplementary Note 4.

In order to identify the shared inversions among species, we first determined the chromosome correspondences among the three species using JCVI v1.3.5⁷², and selected 10 chromosomes with one-to-one correspondence. The alignments were also performed with Nucmer v4.0.0beta2⁷³ and filtered with parameter ‘-i 90 -l 100’, and then the SVs were detected in SyRI v1.6⁷⁴.

To explore the frequency of inversions identified between haplotypes within species, we re-analyzed the whole-genome resequencing data of 56 *C. alismatifolia* cultivars published in our previous study²⁹. The data was quality-controlled with Fastp v0.20.1⁵⁶.

Subsequently, the filtered data were aligned to haplotype A of *C. alismatifolia* as a reference in GTX v2.2.1 (<http://www.genetalks.com/gtxlab.html#clin-gtx-heal4>) and the BAM files were obtained. To obtain reliable inversions, we utilized three SV detection software, including Delly v1.2.6⁷⁵, Pindel v0.2.5b⁷⁶, and Smoove v0.2.5 (<https://github.com/brentp/smoove>) to call SVs, respectively, and the resultant SVs were merged with the parameter of '1000 2111 50' in Survivor v1.0.7⁷⁷. Inversions detected in at least two software were intersected with the inversions between haplotypes of *C. alismatifolia* in Bedtools v2.30.0 (<http://bedtools.readthedocs.org>). When the length of overlapped regions was over 50% of the length of corresponding inversions in the two lists, they were treated as the same inversion.

To validate the inversions with the conflict signals of Hi-C interaction, we merged the two assembled haplotype genomes as one reference genome. Subsequently, we reused Juicer v1.9.9⁵⁸ for Hi-C reads mapping against the reference, and 3D-DNA v180922⁵⁹ was used with parameter “-q 0” to retain all possible contact signals.

HiFi data were further mapped to the haplotypes with minimap2 v2.21-r1071⁷⁸ and the mapping was visualized in IGV v2.17.4⁷⁹ to verify the accuracy of inversion identification. The extraction of genes within the inversion was based on the gene annotation files resulting from SyRI v1.6⁷⁴ by Bedtools v2.30.0 (<http://bedtools.readthedocs.org>). The BK (genome-wide background) for comparison with INTRA (inversions between haplotypes within species) was the whole genomic background including the inversion region.

Population genetics analyses

Based on resequencing data, SNP calling was performed with GTX v2.2.1, followed by GATK v4.2.2.0-1⁸⁰ for VariantFiltration with ‘QD < 2.0, QUAL < 30.0, SOR > 3.0, FS > 60.0, and MQ < 40.0’. PLINK v2.00a3.3LM⁸¹ was used to further filter the low-quality SNPs with parameters of ‘-maf 0.05 --geno 0.2’ to obtain the SNP dataset D1 (13,517,360 SNPs), and LD pruning was further performed based D1 with ‘--indep-pairwise 50 10 0.2’ to obtain a more pivotal SNP dataset D2 (1,509,564 SNPs). With the D2 dataset, we first inferred population structure with Admixture v1.3.0⁸², and then constructed a phylogenetic tree using VCF2Dis v1.46 (<https://github.com/hewm2008/VCF2Dis>) with *Z. officinale* as a outgroup.

To detect genomic regions with abnormal population structure, we performed local PCA analysis based on the D2 dataset. By referring to methods in Todesco et al.⁸³, we used Lostruct program (https://github.com/petrelharp/local_pca) to run PCA on 500 SNP-wide non-overlapping sliding windows for each chromosome independently and calculate their similarity scores, and finally visualized the similarity scores using a multidimensional scaling transformation in R v4.3.3.

In order to detect SNPs under positive selection, we calculated F_{ST} and π using VCFtools v0.1.16⁸⁴ and pixy (<https://pixy.readthedocs.io/en/latest/>), and verified by xpcIcR v1.1.2 (<https://github.com/hardingnj/xpcIcR>) based on the D2 dataset and selscan v2.0.0 (<https://github.com/szpiech/selscan>) with ‘-nsl’ based on the D1 dataset to identify the selected regions. For selected SNPs, we filtered for windows with the top 5% outliers, and all SNPs contained in the window were considered as selected SNPs. All details are available in Supplementary Note 5.

The PIC value of an *l*-allele locus was calculated as following:

$$PIC = 1 - \sum_{i=1}^l P_i^2 - \sum_{i=1}^{l-1} \sum_{j=i+1}^l 2P_i^2 P_j^2 \quad (1)$$

where P_i and P_j were the allele frequencies of the *i*th and *j*th allele in the population, and *l* represented allele number⁸⁵. Finally, the mean value of PIC was calculated with the window size of 100 Kbp.

Recombination rate was estimated across the genome with sliding windows (window size = 10 Kbp) using FastEPRR v2.0⁸⁶. In order to evaluate whether deleterious SNPs were enriched in inversions, we first

annotated SNPs located in protein-coding regions with Sorting Intolerant From Tolerant 4 G (SIFT4G)⁸⁷, with Uniref90 as the reference protein database. The *Z. officinale* genome was used as an outgroup to identify the deleterious alleles in the *C. alismatifolia* genome. At a given position, if a *C. alismatifolia* reference allele matched the *Z. officinale* reference allele, the *C. alismatifolia* allele was categorized as an ancient allele. If a *C. alismatifolia* allele was different from the *Z. officinale* allele, it was defined as a derived allele. Only homozygous deleterious sites were polarized. Derived alleles with SIFT scores less than 0.05 were considered “deleterious”. Information belonging to each chromosome was extracted in Bedtools v2.30.0 (<http://bedtools.readthedocs.org>).

Collinear gene pairs

To detect collinear gene pairs, gene pairs from two haplotypes were retrieved in MCSanX⁸⁸. Gene pairs from the same pair of homologous chromosomes were retained. A final manual check was performed to ensure that the retained collinear gene pairs exhibited both sequence similarity and sequence continuity. Nucleotide substitution rate of the collinear gene pairs was calculated with the Simple *Ka/Ks* Calculator using TBtools v1.120⁶².

Gene expression

RNA-seq data were generated from 37 samples representing a complete set of tissues and developmental stages. 2–3 replicates for each sample. Additional details are available in Supplementary Note 6. Then, RNA from all samples was extracted and sequenced on the MGI-SEQ 2000 sequencing platform, yielding ~6 Gbp of raw data. Next, data was quality-controlled and filtered with Fastp v0.20.1⁵⁶. To ensure data volume uniformity, we randomly selected an identical amount (equal to the minimum amount among all samples) of sequencing reads of samples in *C. alismatifolia* for gene expression quantification with Seqtk v1.3-r106 (<https://github.com/lh3/seqtk>). Subsequently, to obtain read counts for genes, the gene annotation files from different haplotypes were merged for read mapping and counted with STAR v2.7.10a⁸⁹ with parameters ‘--alignIntronMax 50000 --alignMatesGapMax 50000 --outFilterMismatchNmax 1 --outSAMattrIHstart 0’. The stringent parameters were set to ensure that different copies of the gene could be uniquely matched by controlling the mismatch rate. Finally, FPKM values were calculated with StringTie v2.1.6⁹⁰.

Identification of terpene synthase genes

Based on the TPS protein sequences of *Arabidopsis thaliana* and other species listed in Supplementary Data 15, we scanned for candidate TPS genes in *C. alismatifolia* with BLASTP v2.12.0+ (<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>). The hit sequences were then filtered for those containing PF01397 and PF03936 domains. The remaining TPS sequences were structurally corrected with IGV-GSAMAN v0.6.76, based on transcriptome data. Subsequently, we aligned TPS sequences with MAFFT v7.464⁶⁹ software, and trimmed poorly aligned positions with TrimAL v1.4⁹¹ software. To reveal phylogenetic clustering of the TPS genes, we reconstructed the phylogenetic tree with 1000 ultrafast bootstrap replicates to assess branch support in IQ-TREE v2.2.0-beta⁹². Finally, tree visualization of phylogenetic tree was implemented through FigTree v1.4.3 (<http://tree.bio.ed.ac.uk/software/figtree>).

To further examine the structural integrity of candidate genes, the ten conserved motifs (Supplementary Table 11) were identified with MEME v5.5.2⁹³. The TPS gene identification in *C. petiolata* and *C. longa* was conducted by searching for *C. alismatifolia* TPS gene sequences within their respective genome sequences. Finally, all TPS genes were functionally annotated against the NR database (<https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz>), and the annotation accuracy was verified based on reported catalytic activity (Supplementary Table 12; Yang et al.^{65,68}). The positions of the TPS genes on the chromosome

were displayed via MG2C v2.1 (http://mg2c.iask.in/mg2c_v2.1/index.html).

Terpene synthase activity assay

To perform TPS activity assays, the coding sequences of selected *TPS* genes were synthesized by Tsingke company (China, www.tsingke.com.cn) and inserted into the protein expression vector pCold-TF. The *in vitro* products were detected with GC-MS, and the HPLC $\geq 98\%$ of β -elemene (CAS: 515-13-9, WiKeQi Company, China) was used as the standard. To further clarify whether the TPSs catalyze the formation of compound **2** or compound **1**, the products from prokaryotic expression of Chr08HA727 and reference gene AmGAS⁴⁹ (accession number, KC145534.1) were detected with GC-MS (ThermoTSQ9000-Trace1310, Thermo, USA). An empty vector (pCold-TF) was served as a negative control. All experiments were performed in three or more biological replicates. The injection volume was set to 1 μ l and the injection temperature at 150 and 215 °C. Data acquisition and subsequent analysis were executed employing Xcalibur v4.1 software (Thermo Fisher Scientific, USA). Compounds were identified by comparing mass spectra with the NIST2017 mass spectra library as well as standard samples. All details are available in Supplementary Note 7.

Quantification of secondary metabolites

Volatile compounds from flowers of *C. alismatifolia* cv. “Chiang Mai Pink” at three different stages (Flower1, Flower2, and Flower3) were gathered with a solid-phase microextraction method. Subsequently, 0.2 g of each sample was added to a 15 mL bottle along with 3 mL of MTBE for overnight extraction. The final volume was adjusted to 1 mL for thermal analysis. Three biological replicates were performed.

An internal standard of 1 μ l of 0.2 ng μ L⁻¹ Dodecane per 1 mL was added, and the samples were analyzed with GC-MS. The parameters for GC analyses were as follows: 50 °C for 3 min, ramp at a rate of 5 °C min⁻¹ to 250 °C and held for 0 min, ramp at a rate of 10 °C min⁻¹ to 290 °C and held for 5 min. The total run time was 52.0 min. The ion source temperature of MS was set to 300 °C and spectra were recorded from *m/z* = 50 to 500. Data acquisition and subsequent analysis were executed employing Xcalibur v4.1 (Thermo Fisher Scientific, USA). The compounds were then qualitatively analyzed using the NIST2017 library and compared with published plant volatiles in the pherobase database (www.pherobase.com).

Reporting summary

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

Data availability

The raw sequence data for the PacBio HiFi reads, ONT reads, MGI short reads, and Hi-C reads as well as the genomic assembly of *C. longa*, *C. petiolata*, and *C. alismatifolia* reported in this paper have been deposited in the Genome Sequence Archive at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation under BioProject number PRJCA017980. Genome assembly and annotation are available in Figshare [<https://doi.org/10.6084/m9.figshare.25303411.v2>]. Source data are provided with this paper.

Code availability

The python code used to perform PIC analysis, and the R code used for visualization are available in GitHub [<https://github.com/monian1113/Curcuma-scripts>].

References

- Knief, U. et al. Evolution of chromosomal inversions across an avian radiation. *Mol. Biol. Evol.* **41**, msae092 (2024).
- Faria, R., Johannesson, K., Butlin, R. K. & Westram, A. M. Evolving inversions. *Trends Ecol. Evol.* **34**, 239–248 (2019).
- Huang, K. & Rieseberg, L. H. Frequency, origins, and evolutionary role of chromosomal inversions in plants. *Front. Plant Sci.* **11**, 296 (2020).
- Said, I. et al. Linked genetic variation and not genome structure causes widespread differential expression associated with chromosomal inversions. *Proc. Natl Acad. Sci. USA* **115**, 5492–5497 (2018).
- Villoutreix, R. et al. Inversion breakpoints and the evolution of supergenes. *Mol. Ecol.* **30**, 2738–2755 (2021).
- Battlay, P. et al. Large haploblocks underlie rapid adaptation in the invasive weed *Ambrosia artemisiifolia*. *Nat. Commun.* **14**, 1717 (2023).
- Giner-Delgado, C. et al. Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nat. Commun.* **10**, 4222 (2019).
- Harringmeyer, O. S. & Hoekstra, H. E. Chromosomal inversion polymorphisms shape the genomic landscape of deer mice. *Nat. Ecol. Evol.* **6**, 1965–1979 (2022).
- Jay, P. et al. Mutation load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. *Nat. Genet.* **53**, 288–293 (2021).
- Zhou, Y. et al. Pan-genome inversion index reveals evolutionary insights into the subpopulation structure of Asian rice. *Nat. Commun.* **14**, 1567 (2023).
- Huynh, L. Y., Maney, D. L. & Thomas, J. W. Chromosome-wide linkage disequilibrium caused by an inversion polymorphism in the white-throated sparrow (*Zonotrichia albicollis*). *Heredity* **106**, 537–546 (2011).
- Calfee, E. et al. Selective sorting of ancestral introgression in maize and teosinte along an elevational cline. *PLoS Genet.* **17**, e1009810 (2021).
- Wang, L. et al. The interplay of demography and selection during maize domestication and expansion. *Genome Biol.* **18**, 215 (2017).
- Wang, L. et al. Molecular parallelism underlies convergent highland adaptation of maize landraces. *Mol. Biol. Evol.* **38**, 3567–3580 (2021).
- Huang, K. et al. Mutation load in sunflower inversions is negatively correlated with inversion heterozygosity. *Mol. Biol. Evol.* **39**, msac101 (2022).
- Hoffmann, A. A. & Rieseberg, L. H. Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annu. Rev. Ecol. Syst.* **39**, 21–42 (2008).
- Puig, M. et al. Functional impact and evolution of a novel human polymorphic inversion that disrupts a gene and creates a fusion transcript. *PLoS Genet.* **11**, e1005495 (2015).
- Kirkpatrick, M. & Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434 (2006).
- Knief, U. et al. A sex-chromosome inversion causes strong overdominance for sperm traits that affect siring success. *Nat. Ecol. Evol.* **1**, 1177–1184 (2017).
- White, M. J. D. *Animal cytology and evolution*, 3rd edn. (University Press, 1973).
- Damián, A. et al. Fine breakpoint mapping by genome sequencing reveals the first large X inversion disrupting the *NHS* gene in a patient with syndromic cataracts. *Int. J. Mol. Sci.* **22**, 12713 (2021).
- Lavington, E. & Kern, A. D. The effect of common inversion polymorphisms In(2L)t and In(3R)Mo on patterns of transcriptional variation in *Drosophila melanogaster*. *G3* **7**, 3659–3668 (2017).
- Sharma, S. P., Zuo, T. & Peterson, T. Transposon-induced inversions activate gene expression in the maize pericarp. *Genetics* **218**, iyab062 (2021).

24. Torgasheva, A. A. & Borodin, P. M. Synapsis and recombination in inversion heterozygotes. *Biochem. Soc. Trans.* **38**, 1676–1680 (2010).
25. Crow, T. et al. Gene regulatory effects of a large chromosomal inversion in highland maize. *PLoS Genet.* **16**, e1009213 (2020).
26. Leong-Skornickova, J. et al. Chromosome numbers and genome size variation in Indian species of *Curcuma* (Zingiberaceae). *Aob* **100**, 505–526 (2007).
27. Chen, J., Xia, N., Zhao, J., Chen, J. & Henny, R. J. Chromosome numbers and ploidy levels of Chinese *Curcuma* species. *Hortscience* **48**, 525–530 (2013).
28. Ramachandran, K. Chromosome numbers in Zingiberaceae. *Cytologia* **34**, 213–221 (1969).
29. Liao, X. et al. The genomic and bulked segregant analysis of *Curcuma alismatifolia* revealed its diverse bract pigmentation. *abIO-TECH* **3**, 178–196 (2022).
30. Liang, H. et al. Inferring the phylogeny and divergence of Chinese *Curcuma* (Zingiberaceae) in the Hengduan mountains of the Qinghai-Tibet plateau by reduced representation sequencing. *Forests* **12**, 520 (2021).
31. Závěská, E. et al. Phylogeny of *Curcuma* (Zingiberaceae) based on plastid and nuclear sequences: proposal of the new subgenus *Ecomata*. *Taxon* **61**, 747–763 (2012).
32. Heng, L., Jiabing, D., Shanshan, T., Zhonghao, Z. & Ruiwu, Y. Research progress on the systemic relationship of *Curcuma* plants in China. *Mol. Plant Breed.* **17**, 3695–3702 (2019).
33. Wu, Z. *Flora of China*. (HarperCollins, 2012).
34. Zhang, D. Q. et al. Two chromosome-level genome assemblies of *Rhodiola* shed new light on genome evolution in rapid radiation and evolution of the biosynthetic pathway of salidroside. *Plant J.* **117**, 464–482 (2024).
35. Zagórska, J., Kukula-Koch, W., Czop, M., Iłowiecka, K. & Koch, W. Impact of thermal processing on the composition of *Curcuma longa* rhizome. *Foods* **12**, 3086 (2023).
36. Hu, H. et al. Unravelling inversions: technological advances, challenges, and potential impact on crop breeding. *Plant Biotechnol. J.* **22**, 544–554 (2024).
37. Yin, Y. et al. A chromosome-scale genome assembly of turmeric provides insights into curcumin biosynthesis and tuber formation mechanism. *Front. Plant Sci.* **13**, 1003835 (2022).
38. Zhang, X. et al. Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. *Nat. Genet.* **53**, 1250–1259 (2021).
39. Ou, S. J., Chen, J. F. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
40. Ou, S. J. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
41. Su, W., Ou, S., Hufford, M. B. & Peterson, T. A tutorial of EDTA: extensive de novo TE annotator. *Methods Mol. Biol.* **2250**, 55–67 (2021).
42. Sun, P. et al. WGDl: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant* **15**, 1841–1851 (2022).
43. Calderon, L. et al. Diploid genome assembly of the Malbec grapevine cultivar enables haplotype-aware analysis of transcriptomic differences underlying clonal phenotypic variation. *Hortic. Res.* **11**, uhae080 (2024).
44. Lin, Y. R., Schertz, K. F. & Paterson, A. H. Comparative analysis of QTLs affecting plant height and maturity across the Poaceae, in reference to an interspecific *sorghum* population. *Genetics* **141**, 391–411 (1995).
45. Tian, F. et al. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* **43**, 159–162 (2011).
46. Zhao, K. et al. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* **2**, 467 (2011).
47. Karunanithi, P. S. & Zerbe, P. Terpene synthases as metabolic gatekeepers in the evolution of plant terpenoid chemical diversity. *Front. Plant Sci.* **10**, 1166 (2019).
48. Liu, Y., Chen, R. & Xie, T. Research on biosynthesis of anticancer drug elemene and key enzyme germacrene A synthase. *Chin. J. Biochem. Mol. Biol.* **38**, 1322–1334 (2022).
49. Pazouki, L., Memari, H. R., Kannaste, A., Bichele, R. & Niinemets, U. Germacrene A synthase in yarrow (*Achillea millefolium*) is an enzyme with mixed substrate specificity: gene cloning, functional characterization and expression analysis. *Front. Plant Sci.* **6**, 111 (2015).
50. Jia, Q. et al. Origin and early evolution of the plant terpene synthase family. *Proc. Natl Acad. Sci. USA* **119**, e2100361119 (2022).
51. Yang, H. et al. A haplotype-resolved gap-free genome assembly provides novel insight into monoterpenoid diversification in *Mentha suaveolens* ‘Variegata’. *Hortic. Res.* **11**, uhae022 (2024).
52. Shen, Y. et al. Chromosome-level and haplotype-resolved genome provides insight into the tetraploid hybrid origin of *Patchouli*. *Nat. Commun.* **13**, 3511 (2022).
53. Xia, E. et al. The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into its genome evolution and adaptation. *Mol. Plant* **13**, 1013–1026 (2020).
54. Schmidt, C., Pacher, M. & Puchta, H. Efficient induction of heritable inversions in plant genomes using the CRISPR/Cas system. *Plant J.* **98**, 577–589 (2019).
55. Schwartz, C. et al. CRISPR-Cas9-mediated 75.5-Mb inversion in maize. *Nat. Plants* **6**, 1427–1431 (2020).
56. Chen, S. F., Zhou, Y. Q., Chen, Y. R. & Gu, J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, 884–890 (2018).
57. Cheng, H. Y., Concepcion, G. T., Feng, X. W., Zhang, H. W. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
58. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
59. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
60. Jia, K. H. et al. SubPhaser: a robust allopolyploid subgenome phasing method based on subgenome-specific k-mers. *N. Phytol.* **235**, 801–809 (2022).
61. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
62. Chen, C. et al. TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* **13**, 1194–1202 (2020).
63. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, 309–314 (2019).
64. Cantalapiedra, C. P., Hernandez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
65. Yang, P. et al. Chromosome-level genome assembly and functional characterization of terpene synthases provide insights into the volatile terpenoid biosynthesis of *Wurfbainia villosa*. *Plant J.* **112**, 630–645 (2022).
66. Li, P. et al. Chromosome-level genome assembly of *Amomum tsaoko* provides insights into the biosynthesis of flavor compounds. *Hortic. Res.* **9**, uhac211 (2022).
67. Cheng, S. P. et al. Haplotype-resolved genome assembly and allele-specific gene expression in cultivated ginger. *Hortic. Res.* **8**, 188 (2021).

68. Yang, P. et al. Comparing genomes of *Fructus Amomi*-producing species reveals genetic basis of volatile terpenoid divergence. *Plant Physiol.* **193**, 1244–1262 (2023).
69. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
70. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, 609–612 (2006).
71. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
72. Tang, H. et al. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
73. Marcais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
74. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
75. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, 333–339 (2012).
76. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
77. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
78. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
79. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
80. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
81. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
82. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
83. Todesco, M. et al. Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* **584**, 602–607 (2020).
84. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
85. Nagy, S. et al. PICcal: an online program to calculate polymorphic information content for molecular genetic studies. *Biochem. Genet.* **50**, 670–672 (2012).
86. Gao, F., Ming, C., Hu, W. & Li, H. New software for the fast estimation of population recombination rates (FastEPRR) in the genomic era. *G3* **6**, 1563–1571 (2016).
87. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).
88. Wang, Y. et al. MCSanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
89. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
90. Perteza, M., Kim, D., Perteza, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
91. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
92. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
93. Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, 202–208 (2009).

Acknowledgements

We acknowledge the following funding sources: the National Science Foundation of China (grant 32470245), the Innovation Program of the Chinese Academy of Agricultural Sciences, the Chinese Academy of Agricultural Sciences Elite Youth Program (grants 110243160001007), and the Shenzhen Fundamental Research Program (grant JCYJ20220818103212025). We thank Professor Xiang Gao from Northeast Normal University for providing the experimental platform to quantify volatile compounds via GC-MS, and Doctor Yuanjun Ye of Environmental Horticulture Research Institute, Guangdong Academy of Agricultural Sciences for providing images of *C. alismatifolia*.

Author contributions

L.W. and X.Z.L. conceptualized the project. X.Z.L. acquired materials. X.Z.L., D.J.X., T.T.B., D.P., and C.L. conducted genome assembly and analyzed the results. D.J.X., T.T.B., and X.Z.L. performed the verification of TPS function. X.Z.L. and L.W. wrote the manuscript, W.L., H.X.H., M.M.H., B.N., S.C.S., H.R.W., Y.F.T. and Y.Z. revised the manuscript. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-53719-y>.

Correspondence and requests for materials should be addressed to Li Wang.

Peer review information *Nature Communications* thanks Jihai Gao and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024