

Haplotype-resolved T2T genome reveals an ancient hybrid ancestry and functional specialization in red-skinned longan (*Dimocarpus longan*)

Jiaming Zhang

zhangjiaming@catasitbb.cn

Chinese Academy of Tropical Agricultural Sciences

Article

Keywords: Dimocarpus, haplotype-resolved genome, interlineage hybrid, germplasm, population genomics, telomere-to-telomere assembly

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8014891/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

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

Abstract

Cultivated longan (*Dimocarpus longan*) has undergone extensive selection, resulting in reduced genetic diversity and loss of ancestral alleles associated with important horticultural traits. The rare red-skinned germplasms exhibit many valuable breeding traits that cultivated longans do not have, such as purple-red peel pigmentation, high seed abortion rates, persistent flowering, and natural dwarfism. To dissect the genetic basis of these characteristics, we generated the first telomere-to-telomere (T2T) haplotype-resolved genome assemblies for a red-skinned accession (RP1901). The two haplotypes (RPhapA, 427.67 Mb; RPhapB, 427.98 Mb) are reference-grade, revealing an unexpectedly high divergence, with an average nucleotide identity (ANI) of only 94.06% between haplotypes, substantially lower than that among cultivated longans. Phylogenomic and synonymous substitution rate (Ks) analysis indicated that the two RP haplotypes originated from deeply diverged *Dimocarpus* lineages that separated approximately 8.9–10.6 million years ago (Mya), predating the divergence of major cultivated lineages. Population genomic analyses of 91 accessions identified a distinct ancestral component specific to red-skinned longan, which was nearly absent from Chinese cultivars but retained in Southeast Asian germplasms, supporting an ancient hybrid ancestry involving divergent lineages maintained in mainland Southeast Asia. Furthermore, haplotype-specific gene (HSG) and allele-specific expression (ASE) analysis demonstrated functional specialization between the two haplotypes, with RPhapA preferentially contributing to regulatory processes and RPhapB enriched for genes associated with secondary metabolism and pigmentation. Together, our findings demonstrate that red-skinned longan represents a hybrid diploid originating from two deeply divergent ancestral lineages, and highlight the roles of admixture and haplotype specialization in shaping novel horticultural traits.

Significance Statement

We present the first telomere-to-telomere (T2T), haplotype-resolved genome for longan, revealing a hybrid origin rooted in mainland Southeast Asia. The two haplotypes exhibit stark functional specialization: one governs transcriptional regulation, while the other drives secondary metabolism and pigmentation, revealing a complementary functional architecture in this tropical fruit crop.

Introduction

Longan (*Dimocarpus longan* Lour.) has been cultivated in southern China and mainland Southeast Asia for approximately two thousand years. Over this period, cultivated forms have gradually accumulated through the selection and vegetative propagation of superior individuals, resulting in several hundred recognised cultivars today. These cultivars exhibit substantial variation in multiple horticultural traits, including fruit size, sugar content, and maturation period, as well as differences in pericarp colour ranging from yellow-brown to brown and greyish brown. However, unlike its close relatives lychee and rambutan¹, no longan cultivars display the striking red or pink pigmentation characteristic of these Sapindaceae fruits.

Red-skinned longan germplasms provide a rare and valuable opportunity to explore this colour divergence. Their vivid purple-red peel results from anthocyanin accumulation dominated by cyanidin and petunidin derivatives^{2,3,4}. Intriguingly, red-skinned accessions also exhibit a combination of distinctive agronomic traits, including high seed abortion, persistent flowering, and natural dwarfism, several of which are important targets for longan improvement. The exploitation of such agronomic traits based on current longan cultivars, however, is constrained by the unusually narrow genetic base of cultivated longan^{5,6}. Most cultivars derive from a limited number of clonally propagated selections, and the available variations within this pool for genetic improvement are correspondingly limited. The conventional solution—introgression from more divergent germplasm—requires prior knowledge of which germplasm is divergent and to what degree. For red-skinned longan, this relationship has never been resolved. Horticultural literature has generally treated it as a colour variant of *D. longan*⁷. However, whether it represents a simple variant, a landrace, or material of more distant origin has never been tested with genomic data. This distinction is not merely academic. Traits segregating as alleles within the cultivated gene pool can be recovered through conventional crossing, whereas traits originating from a lineage outside that pool require a different transfer strategy and entail different expectations of linkage drag. Morphological assessment and pedigree records alone are insufficient to resolve this question, especially when historical cultivation and introgression have blurred the boundaries between these categories. Resolving which case applies necessitates a genome-level investigation.

Our preliminary sequencing of a red-skinned accession revealed an exceptionally high heterozygosity rate (~ 4%), substantially exceeding the 0.36–0.88% reported for cultivated longan genomes⁵. Such extensive heterozygosity presents a major challenge for genome reconstruction because conventional consensus assemblies collapse divergent alleles and obscure haplotype-specific structural variation⁸. These unresolved regions may contain the very genomic signals required to determine whether red-skinned longan represents a deeply differentiated lineage. The available longan references, including the recent Shixia telomere-to-telomere assembly⁹, are cultivar-derived and were not constructed to phase haplotypes at this level of divergence. Therefore, a fully phased assembly is essential for resolving its genomic architecture.

Generating phased genomes for perennial crops with unknown pedigrees remains challenging. In conventional trio-based approaches, parental genomes provide direct haplotype assignment; however, such information is unavailable for long-lived clonally propagated trees with undocumented breeding histories. The red-skinned accession investigated here was introduced from Vietnam, and its parental origin is unknown. Although Hi-C sequencing enables chromosome-scale haplotype separation, it does not directly determine parental origin. Consequently, additional genomic strategies are required to resolve haplotype relationships and investigate their evolutionary history.

High heterozygosity itself may provide insights into the evolutionary processes shaping perennial fruit crops. In lychee (*Litchi chinensis*), another important Sapindaceae fruit tree, a haplotype-resolved genome of the cultivar Feizixiao revealed two highly divergent haplotypes with 2.27% heterozygosity,

reflecting contributions from independently domesticated early- and late-maturing germplasm¹⁰. Similarly, haplotype-specific structural variation has been shown to influence gene expression and phenotypic divergence in other perennial species⁸. These findings suggest that divergent haplotypes within a single diploid genome can preserve distinct evolutionary histories and contribute to trait variation. However, the extent of haplotype divergence within a single cultivated perennial, and the functional consequences of maintaining such divergent genomes in one nucleus, remains largely unexplored.

To address these questions, we generated the first reference-quality, gap-free, haplotype-resolved T2T genome assemblies of the red-skinned longan accession RP1901. Without parental information, we phased the two haplotypes using an ANI-based anchoring strategy that exploits sequence divergence between haplotypes to establish chromosome-level separation. By integrating comparative genomics, population genomics, haplotype-specific gene analysis, and allele-specific expression analysis across 91 accessions, we uncovered two major features of red-skinned longan: a hybrid ancestry involving deeply divergent *Dimocarpus* lineages and extensive functional specialization between haplotypes. Our results redefine the genomic origin of this valuable germplasm and provide a foundation for understanding the evolutionary mechanisms underlying pigmentation, reproductive development, and plant architecture in longan and related Sapindaceae crops.

Results

Gap-free telomere-to-telomere haplotype-resolved assemblies of red-skinned longan

The red-skinned longan germplasm RP1901 (RP) exhibits several distinctive phenotypic characteristics compared with cultivated longan varieties, including purple-red pigmentation in leaves and pericarp, high seed abortion, persistent flowering, and natural dwarfism (Fig. 1a). To investigate the genomic basis underlying these traits, we generated a high-quality reference genome assembly for this germplasm by integrating PacBio HiFi, Oxford Nanopore, Illumina, and Hi-C sequencing. A total of 31.3 Gb HiFi reads (~73× coverage), 40.8 Gb Illumina short reads (~95×), 14.5 Gb ultralong ONT reads (~34×), and 44.6 Gb Hi-C data (~103×) were used in this study (Table S1). Survey analysis based on 21-mer profiling estimated the genome size at ~431 Mb, with a heterozygosity rate of 4.04% and a repetitive content of 42.3% (Figure S1). This unusually high heterozygosity suggested extensive divergence between allelic sequences and highlighted the necessity of a haplotype-resolved strategy.

The two haplotype assemblies showed high contiguity, with contig N50 values of 24.93 Mb for RPhapA and 27.19 Mb for RPhapB. Hi-C scaffolding anchored the contigs into 15 pseudochromosomes for each haplotype, and all remaining gaps (8 in RPhapA; 5 in RPhapB) were subsequently closed using corrected Oxford Nanopore reads, resulting in final assemblies of 427.67 Mb (RPhapA) and 427.98 Mb (RPhapB). Individual chromosome sizes ranged from 22.7 to 47.0 Mb (Table S2). Centromeric regions were identified by integrating multiple genomic features, including low frequency unique 21-mer, tandem-repeat (TR) enrichment, LTR accumulation, and gene density minima (Fig. 1b). Telomeric repeat

motifs (AAACCCT/AGGGTTT) were detected at both ends of all chromosomes, confirming complete chromosome reconstruction (Fig. 1c, Figure S2).

Table 1
Genome assembly, annotation and assessment.

Items	Genomic feature	RP		JDB	SX
		hapA	hapB		
Assembly	Total sequence length of contigs (Mb)	441.54	437.57	454.50	441.51
	Total number of contigs	227	113	250	98
	Contig N50 (Mb)	24.93	27.19	12.10	28.34
	Total sequence length of scaffolds (Mb)	427.67	427.98	455.5	441.52
	Total number of scaffolds	15	15	90	95
	Scaffold N50 (Mb)	27.92	27.19	29.6	28.34
	Number of gaps	0	0	222	3
	Number of telomeres	30	30	/	26
Annotation	Number of genes	32080	32116	32062	29325
	Repeat content (%)	43.42	42.89	41.70	54.96
	GC content (%)	33.84	33.65	43.90	33.90
Quality assessment	QV	55.80	54.71	/	68.1
	K-mer Completeness (%)	99.36	99.36	/	99.97
	BUSCO Completeness (%)	98.5	98.6	98.1	98.0
	LAI score	21.42	18.98	/	21.2

Comprehensive evaluation validates the accuracy and completeness of the two haplotype assemblies

The quality of the two haplotype assemblies was evaluated using multiple complementary approaches. At the read-alignment level, HiFi, ONT, Hi-C, and NGS reads were mapped onto the assemblies, showing consistent coverage patterns across both haplotypes (Fig. 2a). The mapping rates reached 99.4%, 98.6%, 100%, and 91.1% for HiFi, ONT, Illumina, and Hi-C reads, respectively, with 100% of the assembly covered by at least 1× depth for HiFi, ONT, and NGS reads, and 91% for Hi-C reads. The Hi-C interaction matrices further displayed clear chromosomal interaction patterns for both haplotypes (Fig. 2b).

At the gene-space completeness level, BUSCO analysis recovered 98.5% (RPhapA) and 98.6% (RPhapB) of complete orthologs, exceeding the completeness values reported for previously available longan

genomes, including JDB and SX (Fig. 2c; Table 1).

Base-level accuracy was further assessed using Merqury k-mer analysis¹¹. The assembly achieved 99.36% k-mer completeness, with only 0.17% and 0.08% of k-mers absent from heterozygous and homozygous peaks, respectively (Fig. 2d), indicating that nearly all genomic information present in the Illumina short reads was faithfully captured. The consensus quality values (QV) were 55.80 for RPhapA and 54.71 for RPhapB (55.22 for the combined assembly), corresponding to a nucleotide accuracy greater than 99.999%.

The k-mer multiplicity spectrum exhibited the canonical diploid pattern, with clearly resolved heterozygous and homozygous peaks at approximately 34× and 71× coverage, respectively (Fig. 2d). Two-copy k-mers accounted for only 0.58% of the heterozygous peak, whereas one-copy k-mers constituted merely 0.71% of the homozygous peak, suggesting minimal evidence of artificial duplication or sequence collapse.

The accuracy of haplotype separation was further evaluated using haplotype-specific k-mer analysis (Fig. 2e). RPhapA and RPhapB contained comparable numbers of haplotype-specific k-mers (1.748×10^8 vs. 1.779×10^8 ; ratio = 0.983), indicating balanced representation of the two haplotypes. Notably, haplotype-specific k-mers were substantially more abundant than shared k-mers (3.53×10^8 vs. 1.40×10^8), consistent with the high genomic heterozygosity estimated by GenomeScope2.

The long-range continuity of the assemblies was further supported by LTR Assembly Index (LAI) scores of 21.4 for RPhapA and 19.0 for RPhapB. Together, these assessments demonstrate that the two RP haplotype assemblies accurately capture chromosome structure, gene content, nucleotide sequence, and haplotype-specific variation.

Functional annotation identified more than 32,000 protein-coding genes and over 1,300 non-coding RNA genes in each haplotype assembly (Table S3 and S4). More than 91% of predicted genes were functionally annotated. Repetitive sequences constituted approximately 42% of each haplotype genome, among which long terminal repeats (LTRs) were predominant, accounting for about 29% (Figure S3).

Deeply diverged haplotypes reveal a hybrid ancestry

The RP genome exhibits an exceptionally high heterozygosity rate of 4.04%, far exceeding the 0.36–0.88% range reported for cultivated longan genomes⁵. This extensive allelic divergence was reflected by widespread sequence variation between the two haplotypes, including 12.35 million single-nucleotide variants (SNVs), 5.86 million insertions/deletions (InDels), and 264,875 structural variants (SVs) (Fig. 1c, Fig. 3a). The magnitude of divergence between RPhapA and RPhapB suggested that the two haplotypes may have originated from genetically distinct ancestral lineages.

To determine the evolutionary relationship between RP and cultivated longan, we first compared genome-wide average nucleotide identity (ANI) values among available longan genomes. Cultivated

accessions JDB, SX, and HHZ showed high genome-wide similarity, with ANI values exceeding 98.3%. In contrast, the ANI values between either RP haplotype and cultivated longan genomes averaged 94.37% (Fig. 3b). Remarkably, the inter-haplotype ANI within RP (94.06%) was nearly identical to that between RP and the cultivars (Table S5). These divergence levels exceed those typically observed among cultivated longan accessions and are comparable to lineage-level divergence reported in other perennial crops⁸. By comparison, intergeneric comparisons within Sapindaceae, including longan–lychee and lychee–rambutan, showed substantially lower ANI values (< 90%) (Fig. 3b). Together, these comparisons place the RP haplotypes at an intermediate evolutionary position between typical intraspecific variation and deeper intergeneric divergence.

Genome synteny analysis revealed extensive 1:1 collinearity between RPhapA and RPhapB, as well as between cultivated longan genomes JDB and SX, indicating that recent whole genome duplication has not contributed substantially to genome evolution within the genus *Dimocarpus* (Fig. 3c; Table S6). In contrast, comparisons with *Citrus* species showed prevalent 1:2 synteny relationships, consistent with lineage-specific genome duplication events in the Rutaceae after its divergence from Sapindaceae¹⁵.

The evolutionary divergence of RP haplotypes was further assessed using synonymous substitution rate (Ks) analysis of syntenic orthologs (Table S7). The Ks peak between RPhapA and RPhapB (0.0235) was higher than those observed between RPhapA and cultivated longan genomes JDB or SX, but substantially lower than those between RPhapA and the more distant Sapindaceae plants lychee and rambutan (Fig. 3d). This hierarchical pattern was consistent with the ANI-based comparisons and supports the placement of the two RP haplotypes as deeply diverged but related genomic lineages. In contrast, the similar paralogous Ks distributions between the two haplotypes (Fig. 3e) indicate that their divergence is primarily restricted to orthologous regions rather than resulting from genome-wide acceleration of sequence evolution.

Phylogenomic reconstruction based on 776 single copy orthologs further resolved the evolutionary position of the RP haplotypes within the *Dimocarpus* lineage. Both RPhapA and RPhapB clustered within the longan clade but represented early-diverging lineages relative to cultivated accessions. Molecular dating estimated the divergence times of RPhapA and RPhapB lineages at approximately 8.9 and 10.6 million years ago (Mya), respectively, compared with approximately 3.7 Mya for the divergence between JDB and SX (Fig. 3f).

While both RP haplotypes were positioned early in the phylogeny, they exhibited asymmetric relationships with cultivated longan. This asymmetry was evident at the chromosome level, where RPhapA consistently showed higher ANI with JDB and SX than RPhapB did (Figure S4), suggesting that RPhapA has retained greater genomic affinity to the cultivated gene pool. To test whether this pattern is a localized or genome-wide phenomenon, we calculated Δ ANI across 200-kb windows. The results supported a genome-wide trend: the median Δ ANI (RPhapA - RPhapB) was + 0.27 for JDB and + 0.28 for SX, with 70.0% and 68.6% of windows, respectively, favouring RPhapA (Fig. 3g). Notably, the approximately 30% of windows showing the opposite signal were non-randomly distributed, with a

pronounced concentration in the distal ~ 10 Mb of chromosome 9, where Δ ANI values were consistently negative in both cultivar comparisons, suggesting recombinations between RPhapA and RPhapB happened in the evolutionary history.

Together, these analyses indicate that RP contains two highly divergent haplotypes derived from distinct ancestral lineages. RPhapA shows a consistently closer relationship to cultivated longan across the genome, whereas RPhapB represents a more divergent ancestral contribution. The genome-wide and spatially coherent nature of this asymmetry argues against a pattern driven solely by localized introgression and instead supports broad retention of distinct haplotype ancestries. These features are consistent with a hybrid ancestry involving divergent *Dimocarpus* lineages and provide a genomic framework for understanding the unusual genetic composition of red-skinned longan.

Population genomics identifies a red-skinned-longan-associated ancestral lineage

To further investigate the evolutionary origin of red-skinned longan, we performed whole genome resequencing of 91 longan accessions, including five red-skinned accessions (RP, RPD, RPL, RPM, RPT) and 86 yellow-skinned cultivars from major longan-growing regions, with *L. chinensis* (FZX) as an outgroup. All reads were mapped to the RPhapA T2T genome assembly to characterize population relationships and ancestral components (Table S8; Figure S5).

ADMIXTURE analysis¹⁶ revealed similar support for $K = 2$ and $K = 3$ population models, with cross-validation errors of 0.45754 and 0.45753, respectively (Figure S6). At $K = 2$, all five RP accessions were almost exclusively assigned to a single genetic component ($Q_1 = 100\%$), which was largely absent from Chinese cultivars (2.6%) but retained at substantial levels in the Vietnamese accession YNSJ (62.7%) and the Thai accessions (23.7%) (Fig. 4a). At $K = 3$, an additional component further differentiated Chinese cultivars into Fujian- and Guangdong-associated genetic groups, whereas red-skinned accessions remained dominated by Q_1 with negligible contributions from these cultivated components. These results indicate that red-skinned longan retains a distinct ancestral genetic component that is rare or absent in modern Chinese cultivars.

Principal component analysis (PCA) further supported the distinct genetic position of red-skinned accessions. PC1 explained 6.84% of the total genetic variation and separated all RP accessions from cultivated longan populations (Fig. 4b). Vietnamese and Thai accessions occupied intermediate positions between RP and Chinese cultivars, suggesting partial genetic affinity between Southeast Asian germplasm and the red-skinned lineage. A maximum-likelihood phylogenetic analysis rooted with *L. chinensis* placed all five RP accessions as a strongly supported monophyletic clade at the base of the longan ingroup, with markedly shorter branch lengths (0.27) than those of Vietnam (0.85), Thailand (1.35), or Chinese cultivars (1.66) (Figure S7). Genome-wide F_{ST} calculations in 100-kb sliding windows showed that the genetic differentiation between RP and Chinese cultivars was higher than that observed for Vietnam–China and Thailand–China comparisons (Fig. 4c).

To further evaluate potential admixture relationships and identify populations sharing genetic affinity with RP, we performed f3-, f4-, and D-statistics using ADMIXTOOLS¹⁷. Both f4- and D-statistics revealed significant excess sharing of derived alleles between RP and Vietnamese accessions relative to Chinese cultivars, whereas Thai accessions did not show significant deviation from the expected tree topology (Fig. 4d; Tables S9). The f3-statistics were positive for all tested combinations ($Z = 57.3\text{--}141.0$; Table S10), but these values were not informative for detecting admixture because the putative parental lineages were not represented in the sampled populations. Together, these allele-sharing analyses suggest a closer genetic relationship between RP and Vietnamese germplasm than between RP and Chinese cultivated populations.

TreeMix analysis¹⁸ further supported historical gene-flow relationships among longan populations. The optimal model containing two migration edges (selected based on the likelihood elbow criterion) inferred gene-flow signals from Chinese cultivars toward the Vietnam + Thailand lineage and specifically toward Thai populations (Fig. 4e). When three migration edges were considered, an additional edge connecting RP-related ancestry and Vietnam was inferred, with a migration weight of 14.7%. This pattern was consistent with the ADMIXTURE result showing substantial representation of the red-skinned-associated component in YNSJ and with the allele-sharing patterns detected by f4 and D statistics.

Collectively, population structure, phylogenetic relationships, and formal allele-sharing analyses identify red-skinned longan as a genetically distinct lineage that retains ancestral variation largely absent from modern Chinese cultivars. The close relationship between RP and Vietnamese germplasm provides population-level support for the involvement of mainland Southeast Asian lineages in the evolutionary history of red-skinned longan.

Functional specialization between haplotypes contributes to regulatory and pigmentation-related pathways

Given the deep evolutionary divergence between the two RP haplotypes, we investigated whether the two haplotypes exhibit complementary functional signatures. Haplotype-specific gene (HSG) analysis identified 5,549 genes unique to RPhapA and 5,483 genes unique to RPhapB (Table S11). Functional enrichment analysis revealed distinct patterns between the two haplotypes. RPhapA HSGs were enriched in processes associated with transposable element regulation, DNA integration, RNA-directed DNA polymerase activity, heme A metabolism, nicotinamide metabolism, and cytokinin biosynthesis (Fig. 5a, Table S12). KEGG enrichment further highlighted regulatory processes, including pathways associated with basal transcriptional regulation (Fig. 5b).

In contrast, RPhapB HSGs showed enrichment for functions related to UV-responsive pigment accumulation, floral organ identity maintenance, and flavonoid, isoflavonoid, and other secondary-metabolism pathways at the KEGG level (Fig. 5b). These contrasting enrichment patterns indicate that the two RP haplotypes have accumulated complementary functional repertoires, with RPhapA showing

stronger representation of regulatory-associated genes and RPhapB contributing more genes associated with secondary metabolism and pigmentation-related processes.

To further investigate whether the two haplotypes exhibit transcriptional asymmetry, we performed allele-specific expression (ASE) analysis using developing RP pericarp tissues. The number of ASE genes was comparable between RPhapA-dominant and RPhapB-dominant categories at S1 (1,889 versus 1,936) and S2 (1,888 versus 1,966), but declined sharply at S3 (1,113 and 1,114 genes, respectively; Figure S8). Functional enrichment of ASE genes revealed developmental stage- and haplotype-specific expression patterns (Fig. 5c), with RPhapB-dominant genes at S3 enriched for RNA stabilization, double-strand break repair, and protein depalmitoylation, consistent with post-transcriptional remodelling during fruit maturation.

To determine whether ASE patterns were associated with differences in protein-coding evolution, we compared Ka/Ks ratios between ASE and non-ASE genes. Both RPhapA-dominant and RPhapB-dominant genes exhibited significantly lower Ka/Ks ratios than non-ASE genes (Table S13), together with a higher proportion of gene pairs under purifying selection (Fig. 5d; Table S13). The similar Ka/Ks distributions between the two ASE categories suggest that the observed expression asymmetry is unlikely to be driven by extensive divergence in protein-coding sequences, but instead may primarily reflect cis-regulatory differences between haplotypes.

Together, these analyses demonstrate that the two RP haplotypes have undergone divergent evolutionary trajectories, resulting in complementary functional differentiation at both gene-content and transcriptional levels. Such haplotype-specific functional biases may contribute to the unique combination of pigmentation, developmental, and architectural traits observed in red-skinned longan.

Discussion

Exceptional haplotype divergence within a diploid longan genome

The defining feature of the RP genome is not simply its elevated heterozygosity, but the evolutionary history encoded within this variation. At 4.04%, RP exhibits four to ten times higher heterozygosity than cultivated longan genomes (0.36–0.88%)^{5, 9, 12} and nearly twice that reported for the lychee cultivar Feizixiao (2.27%), previously considered one of the most heterozygous Sapindaceae genomes assembled¹⁰. This level of heterozygosity also exceeds that reported for the pear hybrids 'Yuluxiang' and 'Hongxiangsu' (1.6–2.0%)¹⁹ and citrus species (1.5–2.2%)²⁰. However, in these previously characterised cases, divergent haplotypes originated from cultivars or landraces within the same crop lineage. In contrast, the two RP haplotypes trace to lineages that diverged during the late Miocene, with estimated separation times of 8.9 Mya for RPhapA and 10.6 Mya for RPhapB, substantially predating the divergence between cultivated JDB and SX (~ 3.7 Mya).

Thus, the high heterozygosity of RP does not primarily represent variation accumulated during crop diversification, but rather reflects the coexistence of two deeply divergent evolutionary histories within a single diploid genome. The sequence divergence between these haplotypes supports this interpretation. The inter-haplotype ANI is 94.06%, compared with values exceeding 98.3% among cultivated genomes JDB, SX, and HHZ. Moreover, ANI between RP haplotypes and cultivated longan (94.37%) is comparable to the divergence between the two RP haplotypes themselves, indicating that neither haplotype corresponds simply to a typical cultivated genome.

Chromosome-scale ANI comparisons revealed that RPhapA is consistently closer to cultivated longan than RPhapB, allowing haplotype orientation relative to the cultivated lineage (Fig. 3g, Figure S4). However, this relationship represents only a minor asymmetry superimposed on a much deeper divergence. Both haplotypes retain evolutionary histories distinct from the modern cultivated gene pool.

These observations have important implications for interpreting diploid plant genomes. When haplotypes are separated by recent variation, they can generally be considered alternative allelic states of the same gene repertoire. In RP, however, the depth of divergence suggests that the two haplotypes have accumulated substantial differences in gene content and regulation, raising the possibility that a single diploid individual can preserve complementary genetic functions inherited from divergent ancestral lineages.

Although the divergence between RP haplotypes approaches levels reported among some plant lineages (Shi et al., 2024), we caution against interpreting this value as evidence for species-level separation. The 95–96% ANI threshold commonly used for prokaryotic species delimitation²¹ was developed for clonally reproducing microorganisms and is not applicable to comparisons between haplotypes within a sexually reproducing diploid plant. Determining the taxonomic significance of this divergence will require broader sampling of wild *Dimocarpus* populations.

Hybrid ancestry of the red-skinned lineage

The presence of two highly divergent haplotypes within a single diploid genome raises questions about their evolutionary origin. Three scenarios could explain this pattern: hybridization between long-separated lineages, ancient introgression involving divergent genomic backgrounds, or retention of ancestral polymorphism through incomplete lineage sorting.

Our data are most consistent with an ancient admixture history involving divergent lineages. Under localized introgression, ancestry tracts are broken up by recombination, producing interdigitated blocks of comparable extent favouring each haplotype; under lineage-level mixing, the asymmetry remains uniform along the chromosomes for as long as recombination has not dispersed the parental contributions. The observed pattern is unambiguously the latter: concordant segments favouring RPhapA span 100.2 Mb across 89 runs, against only 8.8 Mb across 12 runs for the reverse, with the longest run reaching 5.2 Mb (Fig. 3g). The two haplotypes are therefore retained as largely intact

parental chromosome sets rather than as recombined mosaics. Population-level data are consistent with this reading: a red-skinned-longan-associated ancestral component was predominant in all red-skinned accessions but nearly absent from Chinese cultivars, while being retained at intermediate frequencies in Vietnamese and Thai germplasm, and f_4 - and D-statistics detected stronger affinity between RP and Vietnamese accessions than between RP and Chinese cultivars. Retention of ancestral polymorphism cannot be completely excluded, and would produce the same genome-wide pattern if the ancestral population had been strongly structured, but the intact organization of the chromosome sets argues against ancient introgression as the source of the divergence.

These observations bear on two distinct timescales. The parental lineages themselves separated in the late Miocene and evolved independently for millions of years thereafter, a period during which the traits now carried by red-skinned longan may have been shaped. Their combination within a single genome, by contrast, has not been followed by sufficient recombination to disperse the parental chromosome sets. In a long-lived perennial with a juvenile phase of six to eight years and a high rate of seed abortion, opportunities for recombination are correspondingly infrequent, so the intact organization of the haplotypes constrains the number of intervening generations rather than the absolute age of the combination.

We therefore characterize this as hybrid ancestry rather than homoploid hybrid speciation. The formal criteria for the latter require genetic admixture from two parental lineages, reproductive isolation between the hybrid lineage and both parents, and evidence that the isolation arose directly from hybridization^{22, 23}. We satisfy the first criterion but have not tested the others. This distinction is not pedantic: homoploid hybrid speciation is considered rare precisely because it lacks the immediate reproductive isolation conferred by polyploidy²⁴, and genomic admixture alone, while necessary, is not sufficient to establish it.

A more tractable question is whether this represents a recurring hybrid or an established lineage. RPD and RPL are independent field collections that differ from RP in leaf morphology, fruit morphology and pericarp colour (Figure S9); yet both carry the same ancestral component—a configuration that a repeatedly re-formed F_1 would not be expected to produce across morphologically differentiated genotypes. The material reproduces sexually, and the morphological differences among independent collections indicate segregation rather than clonal identity.

Its high rate of seed abortion nonetheless complicates this interpretation. Reduced fertility is characteristic of divergent hybrids—across more than 2,000 documented crosses, F_1 hybrids average 0.393 pollen fertility at the genus level, compared with 0.82 for non-hybrid species²⁴—but established homoploid hybrid lineages typically show fertility restored to near-parental levels, as lineages that fail to recover it are unlikely to persist. Red-skinned longan has evidently not followed that path, and two features of the system may explain why. Longevity can substitute for fecundity in a perennial tree, where an individual may persist for decades on infrequent successful reproduction; vegetative propagation removes the constraint altogether once the material enters cultivation. Meiotic chromosome spreads

and pollen viability assessment in RP, benchmarked against a cultivar control, would distinguish these possibilities and test whether pairing is indeed impaired at this level of divergence.

The timing places the underlying lineage split in a documented biogeographic context. A genomic analysis of 60 citrus accessions concluded that the genus diversified during the late Miocene through a rapid Southeast Asian radiation, correlated with a marked weakening of the monsoons²⁰. Our estimates of 8.9 and 10.6 Mya fall within the same temporal window and the same region, raising the possibility that the divergence recorded in the RP genome was part of a broader late Miocene diversification of the Southeast Asian woody flora, rather than an isolated event. Testing this—and identifying the parental lineages themselves—will require sequencing wild *Dimocarpus* taxa, including *D. longan* subsp. *malesianus*, *D. longan* var. *obtusus* and *D. confinis*, together with broader sampling across mainland Southeast Asia.

A cross-haplotype regulatory architecture

The divergent evolutionary histories of RP haplotypes provide a potential genomic basis for a key horticultural target trait: pigmentation. Haplotype-specific gene analysis revealed complementary functional enrichment patterns between RPhapA and RPhapB. Flavonoid biosynthesis and pigmentation-related genes are preferentially enriched on RPhapB, whereas RPhapA-specific genes are biased toward transcriptional regulation. This complementary distribution of functional repertoires suggests that the two haplotypes contribute different biological capacities within the same diploid genome.

Similar haplotype-specific regulatory patterns have been increasingly recognized in heterozygous perennial crops. Haplotype-resolved analyses in hybrid poplar demonstrated that sequence divergence and transposable element occupancy contribute to allele-specific expression patterns⁸. In citrus, haplotype-resolved chromatin landscape analysis revealed that structural variants in promoter regions, particularly transposon insertions, are associated with allele-specific expression, with approximately 30% of variation in ASE attributable to haplotype-associated factors²⁵. In blood oranges, a Copia-like retrotransposon insertion upstream of the Ruby gene introduces cis-regulatory modules that confer cold-responsive anthocyanin accumulation²⁶, demonstrating how structural variation at a single locus can generate a novel trait phenotype. These findings collectively suggest that transposon-mediated cis-regulatory divergence may represent a general mechanism underlying trait variation in heterozygous perennial crops.

ASE analysis in RP further demonstrated extensive transcriptional asymmetry between haplotypes during pericarp development. However, the comparable Ka/Ks distributions between RPhapA-dominant and RPhapB-dominant genes suggest that these expression differences are unlikely to result from widespread protein-coding divergence, but instead are more consistent with regulatory variation between haplotypes. The coexistence of regulatory-associated genes on one haplotype and pigmentation-related genes on the other suggests a potential inter-haplotype functional complementation pattern that extends beyond simple allelic dominance.

More broadly, cross-haplotype regulation may represent an underappreciated mechanism by which heterozygous horticultural crops, particularly those of hybrid or introgression origin, maintain trait expression across generations of vegetative propagation. This is consistent with the observation that in clonally propagated polyploids such as potato, strawberry, and banana, the maintenance of heterozygosity is critical for agronomic performance, and cis-regulatory variation is increasingly recognized as a major contributor to phenotypic diversity²⁷. In the context of heterosis, allele-specific expression arising from cis-regulatory divergence has been proposed to enable functional complementation between haplotypes, where a beneficial allele from one parent compensates for a less active or deleterious allele from the other parent²⁸. The configuration we observe, regulatory capacity enriched on one haplotype and biosynthetic capacity, represents an inter-haplotype complementation mechanism that extends beyond classical dominance models. Whether such cross-haplotype regulation contributes more generally to hybrid vigour in fruit crops remains an open question, but the T2T haplotype-resolved reference now makes such experiments technically feasible.

Implications for longan improvement and future directions

The genomic architecture revealed here has important implications for longan improvement. Red-skinned longan represents a source of genetic variation that is largely absent from the currently characterised cultivated gene pool. The ancestral component associated with this lineage was nearly absent from the Chinese cultivars examined, suggesting that the distinctive traits maintained in red-skinned germplasm, including red pericarp pigmentation, reduced plant stature, and altered flowering behaviour, may not be readily recovered through selection within existing elite germplasm. Instead, their utilization will likely require targeted introgression strategies and careful management of linkage drag associated with transferring variation from deeply divergent lineages.

The haplotype-resolved genome provides a framework for guiding such efforts. The asymmetric distribution of regulatory and metabolic functions between haplotypes may influence how specific traits are inherited and expressed during breeding. Furthermore, Southeast Asian germplasm retaining related ancestral components may represent valuable intermediate resources for broadening the genetic base of cultivated longan. Given that red-skinned lineages remain poorly characterised, expanding genomic surveys across mainland Southeast Asia will be important for identifying related diversity and informing conservation strategies.

Several limitations should be considered. First, population analyses were performed using a red-skinned reference, and although we repeated the workflow against cultivar assemblies, potential mapping bias cannot be completely excluded in highly divergent genomic regions. Second, the current population panel includes limited representation from Southeast Asian germplasm, with only one Vietnamese and three Thai accessions available, preventing precise inference of the geographic origin of the red-skinned lineage. Third, because parental genomes have not been identified, the inference of ancient hybrid ancestry is based on genome-wide divergence patterns and population genetic evidence rather than

direct comparison with candidate parental lineages. Finally, although the present study reveals genomic features associated with pigmentation and other distinctive traits, functional validation of the underlying causal variants will be required to establish their molecular mechanisms.

Materials and Methods

Plant materials and sample collection

The longan accessions used in this study were maintained at the National Germplasm Repository for Tropical Fruit Trees in Zhanjiang, Guangdong Province, China (21°09'52"–21°11'01" N, 110°16'24"–110°16'39" E). Young leaves were collected for genomic DNA extraction using the cetyltrimethylammonium bromide (CTAB) method. For RNA-seq, pericarp tissues were collected from red-skinned longan RP1901 (RP) at three developmental stages (10, 45, and 90 days after flowering, DAF), with three biological replicates per stage. Total RNA was extracted using the HiPure HP Plant RNA Kit (Magen, R4165, China).

Sequencing and library construction

PacBio HiFi libraries with an average insert size of approximately 15 kb were prepared using the SMRTbell Prep Kit 3.0 (102-182-700, PacBio, USA) and sequenced on the PacBio Revio platform. Oxford Nanopore (ONT) libraries were constructed using the Ligation Sequencing Kit 1D (SQK-LSK109, ONT, UK) and sequenced on the PromethION 48 platform. Hi-C libraries, whole-genome sequencing libraries and RNA-seq libraries were prepared following standard protocols and sequenced on the Illumina NovaSeq 6000 platform with 150-bp paired-end reads.

Genome assembly and quality assessment

Genome characteristics were first estimated using a k-mer ($k = 21$) frequency analysis performed with Jellyfish (v2.3.0)²⁹ on 40 Gb of high-quality Illumina reads. The resulting k-mer distribution was analysed using GenomeScope2 (v2.0)³⁰ to estimate genome size, heterozygosity, and repeat content.

Haplotype-resolved genome assemblies were generated using hifiasm (v0.19.9)³¹ with PacBio HiFi and Hi-C data. Hi-C reads were processed using HiC-Pro (v3.1.0)³² to obtain valid read pairs, and contigs were anchored and ordered into chromosome-scale scaffolds using YaHS (v1.2)³³ based on Hi-C contact-frequency matrices. Remaining gaps were filled using corrected Oxford Nanopore reads with TGS-GapCloser (v1.2.1)³⁴. Hi-C interaction matrices were generated using HapHiC (v1.0.6)³⁵.

Assembly completeness was evaluated using BUSCO v5.7.1 with the embryophyta_odb10 database³⁶. Consensus quality value (QV) and k-mer completeness were estimated using Merquy¹¹ by comparing the assembled genome to the Illumina read k-mer set. The assembly quality of long terminal repeat (LTR) retrotransposons was assessed by calculating the LTR Assembly Index (LAI) using the LTR_retriever (v2.9.0) package³⁷.

Genome annotation and repeat sequence identification

Structural gene annotation was conducted using a comprehensive strategy integrating homology-based, *ab initio*, and transcriptome-based predictions. For homology-based prediction, protein sequences from JDB longan, SX longan, lychee, rambutan, yellowhorn, soapberry, and *Arabidopsis thaliana* were aligned to the red-skinned longan genome to identify conserved gene structures. *Ab initio* gene prediction was performed using Augustus (v3.4.0)³⁸, Genscan (v1.0)³⁹, and GlimmerHMM (v3.0.4)⁴⁰, which infer gene models from intrinsic sequence features. Transcriptome evidence from RNA-seq data was incorporated to support and refine gene models. Outputs from all three approaches were merged using the MAKER pipeline (v3.1.4) to generate a high-confidence⁴¹, non-redundant gene set. Functional annotation was performed by aligning predicted proteins to multiple public databases, including SwissProt, TrEMBL, KEGG, InterPro, Gene Ontology (GO), and NCBI's non-redundant (NR) protein database.

Non-coding RNA (ncRNA) was also annotated. Transfer RNA (tRNA) genes were identified using tRNAscan-SE (v2.0.12)⁴², and ribosomal RNA (rRNA) genes were annotated using Barrnap (v0.9). MicroRNA (miRNA) and small nuclear RNA (snRNA) genes were predicted using INFERNAL (v1.1.5)⁴³ with covariance models from the Rfam database. Transposable elements (TEs) were identified using the EDTA (v2.2.1) pipeline⁴⁴ with default parameters, which integrates multiple tools to classify and annotate both class I (retrotransposons) and class II (DNA transposons) elements in plant genomes.

Telomere detection and centromere localisation

Telomeric repeats were detected using TeloExplorer from the quarTeT toolkit⁴⁵ with parameters (-c plant -m 50). Missing telomeres were complemented using Teloclip v0.3.4. Centromeric regions were predicted using multiple complementary approaches. First, unique 21-mer distributions across each chromosome were analysed with GenomeTools v1.6.5⁴⁶; unique k-mer counts were calculated in non-overlapping 100-kb windows, and regions with markedly reduced unique k-mer density were considered potential centromeric regions. Tandem repeats were identified using Tandem Repeats Finder (TRF) v4.09.1⁴⁷ with the parameters 2 7 7 80 10 50 600 -f -d -m. Repeats 30 to 600 bp in length were extracted, clustered using Centromics based on pairwise similarity, and the cluster with the highest genomic abundance was designated as the putative centromeric repeat type.

The CentroMiner module from the quarTeT toolkit was used to integrate TR-enriched regions with the TE annotation from EDTA. Final centromere positions were determined by integrating multiple lines of evidence, including unique k-mer distribution, TR clustering, repeat monomer enrichment, LTR content, and gene density across both RPhapA and RPhapB assemblies. Circos v0.69-9⁴⁸ was used to visualise the distribution patterns of unique k-mers, TR clusters, repeat monomers, LTRs, and gene density.

Haplotype comparison and identification of variant sites

Whole-genome alignments between RPhapA and RPhapB were performed using Minimap2 (v2.28-r1209)⁴⁹. Structural rearrangements, including inversions, translocations, and duplications, were identified with SyRI (v1.7.0)⁵⁰. Plotsr (v1.1.1)⁵¹ was used to visualise genome-wide structural differences.

Genome-wide and sliding-window ANI analysis

Pairwise genome alignments were performed using MUMmer (v4.0.1)⁵², and ANI values between homologous chromosome pairs were calculated using dnadiff in MUMmer. For sliding-window comparison, each haplotype was aligned separately to JDB and SX with Minimap2 (v2.28-r1209; asm10 preset, --secondary = no)⁴⁹; block identity was taken from the gap-compressed de:f tag, blocks shorter than 200 bp were discarded, and identities were assigned to 200-kb windows weighted by overlap length. Homologous windows between haplotypes were defined from SyRI syntenic and inverted regions ($\geq 70\%$ mapped). Δ ANI was calculated as RPhapA minus RPhapB identity against the same reference. Windows required $\geq 50\%$ alignment coverage for both haplotypes and a coverage ratio over 0.70, ensuring that Δ ANI could not arise from asymmetric alignment completeness between the two assemblies.

Comparative genomics and phylogenomic analysis

For comparative genomic analyses, we selected 11 representative plant species spanning major angiosperm lineages, including *Arabidopsis thaliana* (*A. thaliana*), *Carica papaya* (*C. papaya*), *Theobroma cacao* (*T. cacao*), *Citrus sinensis* (*C. sinensis*), *Xanthoceras sorbifolium* (*X. sorbifolium*), *Nephelium lappaceum* (*N. lappaceum*), *Litchi chinensis* (*L. chinensis*), *Dimocarpus longan* (*D. longan*), *Malus domestica* (*M. domestica*), *Vitis vinifera* (*V. vinifera*), and *Solanum lycopersicum* (*S. lycopersicum*). In addition, multiple Sapindaceae genome assemblies were incorporated to provide a more comprehensive evolutionary framework, including two longan genomes^{5,9}, one lychee genome¹⁰, and one rambutan genome⁵³. Genome assemblies and GFF3 annotation files were downloaded from NCBI.

Redundant gene features were removed using JCVI (v1.5.6)⁵⁴. Protein-coding sequences and CDS were extracted from GFF3 files using gffread (v0.12.7)⁵⁵. Gene-family clustering and identification of single copy orthologs were performed with OrthoFinder (v2.5.5)⁵⁶. Protein-coding sequences of single copy orthologs were concatenated in a consistent order and aligned using MAFFT (v7.505)⁵⁷. Paralogous and orthologous gene pairs were identified using the collinearity module in WGDl (v0.74)⁵⁸, and Ka and Ks substitution rates were calculated with the Ks module using CDS and protein sequences. The Ks_YN00 values were used to visualise duplication events, and the results were plotted using ggplot2 in R.

For phylogenomic analysis, coding DNA sequences of single copy genes were combined in a consistent order and aligned using MAFFT. A total of 776 single copy genes and 961,373 aligned sites were used. Substitution model selection was performed with ModelFinder⁵⁹, and a GTR + F+I+R5 model was selected for further analysis. The phylogenetic tree was inferred under maximum likelihood in IQ-TREE 2⁶⁰, rooted with *S. lycopersicum*, and used for molecular clock analysis. Molecular clock analysis was performed using a Bayesian Markov chain Monte Carlo (MCMC) approach in BEAST 2.7.8⁶¹, with four

fossil calibration points. The GTR model with site heterogeneity (G + I) ⁶² was used for nucleotide substitution evaluation, with a Bayesian Skyline coalescent model ⁶³ and a relaxed molecular clock. The MCMC process was accelerated using the BEAGLE library ⁶⁴, run for 10 billion states, sampled every 5,000 states. The resulting phylogenetic tree was visualised using MEGA (v12) (Kumar et al., 2024).

Population resequencing and variant calling

A panel of 91 longan accessions, comprising five red-skinned longan lineage accessions (RP, RPD, RPL, RPM, RPT), one Vietnamese accession (YNSJ), three Thai accessions (YD, PT, SZP), and 82 Chinese cultivars, was used for population genomic analyses, with *L. chinensis* accession FZX as outgroup. Resequencing data for the five red-skinned lineage longan accessions were newly generated in this study; data for the remaining 86 longan accessions and the *L. chinensis* outgroup were obtained from public databases (PRJCA005305 and PRJNA747875).

Clean reads from 91 longan accessions and the FZX outgroup were aligned to the RPhapA T2T genome using BWA-MEM (v2.2.1) ⁶⁵, sorted and deduplicated with SAMtools ⁶⁶ and GATK MarkDuplicates ⁶⁷. SNPs were jointly called across all accessions using GATK HaplotypeCaller followed by GenotypeGVCFs, and filtered by bcftools (v1.21) (variants with QUAL < 30 or DP < 10 were removed), retaining bi-allelic SNPs with missing rate < 10% and minor allele frequency ≥ 0.05 for downstream analyses. SNPs were LD-pruned by PLINK (v2.0.0) ⁶⁸ with $-\text{indep-pairwise } 50 \ 10 \ 0.1$. Population structure was inferred with ADMIXTURE ¹⁶ for K values ranging from 2 to 10 with default ADMIXTURE parameter. Genome-wide F_{ST} was computed in 100-kb sliding windows. f3-, f4-, and D-statistics were computed using ADMIXTOOLS ¹⁷, and TreeMix ¹⁸ analyses were performed with migration edge numbers ranging from 0 to 3.

Principal component analysis (PCA) was performed using smartpca in EIGENSOFT (v7.2.1) ⁶⁹, and the first two principal components were visualised in R with ggplot2. For phylogenetic reconstruction of the 92 accessions (91 longan + FZX), a maximum-likelihood tree was constructed from concatenated genome-wide SNPs using IQ-TREE (v2.1.4) with the MFP + ASC model and 1,000 ultrafast bootstrap replicates; the tree was visualised using MEGA (v12) ⁷⁰. Genome-wide population differentiation was quantified by computing weighted F_{ST} in non-overlapping 100-kb sliding windows for all pairwise population comparisons using VCFtools (v0.1.16) ⁷¹.

Haplotype-specific gene identification and functional enrichment

The allelic gene pairs between haplotypes were identified referring to a previous method used in pear ⁷². Haplotype-specific genes (HSGs) were defined as genes annotated in one haplotype assembly (RPhapA or RPhapB) that lacked a corresponding allele on the other haplotype. Pairwise syntenic blocks between RPhapA and RPhapB were first constructed using JCVI MCscan ⁵⁴ and 1:1 orthologous gene pairs were extracted, after which non-syntenic genes were aligned to the opposite haplotype assembly using GMAP v2025-07-30 ⁷³ with a cDNA-mapping criterion (> 50% reciprocal coordinate overlap, > 80% sequence identity, > 80% query coverage). Genes lacking a supported counterpart by either criterion were classified

as HSGs. GO and KEGG enrichment analyses were performed using the clusterProfiler R package (v4.10.0) ⁷⁴ with a hypergeometric test and Benjamini–Hochberg multiple-testing correction ($P_{\text{adj}} < 0.05$); terms with adjusted $P < 0.05$ were considered significantly enriched.

Allele-specific expression analysis

Clean RNA-seq reads from RP pericarp at three developmental stages (S1, S2, S3) were mapped to a combined diploid reference (RPhapA + RPhapB) using STAR (v2.7) ⁷⁵, with `--outFilterMultimapNmax 1` to retain only uniquely mapped reads. The resulting alignments were split into haplotype-specific BAM files based on chromosome assignment using SAMtools. Read counts for each 1:1 RPhapA–RPhapB orthologous gene pair (defined above) were then quantified separately on each haplotype using featureCounts (v2.1.1) ⁷⁶ at the mRNA level. ASE was tested using DESeq2 ⁷⁷ with biological replicates treated as paired samples and haplotype identity as the comparison factor. Genes were classified as showing significant ASE when the Benjamini–Hochberg-adjusted P-value was < 0.05 and the absolute log2 fold change between RPhapA and RPhapB exceeded 0.5. Positive and negative log2 fold-change values were classified as RPhapA-dominant and RPhapB-dominant expression, respectively. Pairwise Ka and Ks substitution rates were calculated for all 1:1 RPhapA–RPhapB orthologous gene pairs using the `-ks` module of WGDl (v0.74) ⁵⁸ with the YN model. Ka/Ks distributions across ASE categories were compared using two-sided Wilcoxon rank-sum tests in R.

Declarations

Data availability statement

The authors certify that all data presented in this report are available either in the supporting information tables, figures, or in publicly available databases. The PacBio HiFi, Oxford Nanopore, Hi-C, Illumina, and transcriptome data, haplotype-resolved genome assemblies, and annotations generated in this study have been deposited at the Genome Sequence Archive (GSA) of the National Genomics Data Centre under BioProject accession PRJCA048647 with GSA accession CRA031706 (<https://ngdc.cncb.ac.cn/>). Resequencing data for 86 additional longan accessions were downloaded from GSA under accession CRA004281 with BioProject accession PRJCA005305, and *Litchi chinensis* FZX resequencing data at NCBI SRA BioProject PRJNA747875. All other datasets supporting the conclusions of this article are included within the article and its additional files.

Conflict of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions

XH, JZ and SS conceived the project, coordinated research activities, and designed the experiments. XH, LL, XL and TL collected plant materials and investigated phenotypes. XH assembled and annotated the genomes. XH and JZ carried out the genome comparisons and phylogenomic analyses. XH designed and analysed the RNA-seq experiments. XH, JZ, YY and DT analysed the gene expression network. XH, SS and JZ wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank BerryGenomics Co., Ltd. for sequencing support. This work was supported by the Project of State Key Laboratory of Tropical Crop Breeding (No. NKLTBCBXTD11), the Natural Science Foundation of Hainan Province (No. 325MS122), the Project of Science and Technology Innovation of National Tropical Agricultural Science Centre of Chinese Academy of Tropical Agricultural Sciences (No. CATAS202618). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Wang, J., et al.: Comparative metabolomic and transcriptomic analysis reveals that variations in flavonoids determine the colors of different rambutan cultivars. *Horticulturae*. **10**, 263 (2024). <https://doi.org/10.3390/horticulturae10030263>
2. Hu, X., et al.: DIMYB1 positively regulates anthocyanin biosynthesis and contributes to red exocarp coloration in red-skinned longan. *Gene*. **960**, 149522 (2025). <https://doi.org/10.1016/j.gene.2025.149522>
3. Yi, D., et al.: Integrative analysis of the coloring mechanism of red longan pericarp through metabolome and transcriptome analyses. *J. Agric. Food Chem.* **69**, 1806–1815 (2021). <https://doi.org/10.1021/acs.jafc.0c05023>
4. Castillo-Olvera, G., et al.: *Litchi chinensis*: nutritional, functional, and nutraceutical properties. *Food Prod. Process. Nutr.* **7**, 3 (2025). <https://doi.org/10.1186/s43014-024-00275-z>
5. Wang, J., et al.: Genomic insights into longan evolution from a chromosome-level genome assembly and population genomics of longan accessions. *Hortic. Res.* **9** (2022). <https://doi.org/10.1093/hr/uhac021>
6. Sarkar, T., et al.: Minor tropical fruits as a potential source of bioactive and functional foods. *Crit. Rev. Food Sci. Nutr.* **63**, 6491–6535 (2023). <https://doi.org/10.1080/10408398.2022.2033953>
7. Hung, N.Q., Hung, V.V., Hong, N.T.B., Bay, D.T.B.: Current status of longan production in Vietnam. *Acta Hortic.* **1293**, 15–20 (2020). <https://doi.org/10.17660/ActaHortic.2020.1293.3>
8. Shi, T.L., et al.: High-quality genome assembly enables prediction of allele-specific gene expression in hybrid poplar. *Plant. Physiol.* **195**, 652–670 (2024). <https://doi.org/10.1093/plphys/kiae078>
9. Zhao, G., et al.: Improved longan genome assembly reveals insights into flowering mechanisms. *Plant. Biotechnol. J.* **24**, 1678–1696 (2026). <https://doi.org/10.1111/pbi.70433>

10. Hu, G., et al.: Two divergent haplotypes from a highly heterozygous lychee genome suggest independent domestication events for early and late-maturing cultivars. *Nat. Genet.* **54**, 73–83 (2022). <https://doi.org/10.1038/s41588-021-00971-3>
11. Rhie, A., Walenz, B.P., Koren, S., Phillippy, A.M.: Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020). <https://doi.org/10.1186/s13059-020-02134-9>
12. Lin, Y., et al.: Genome-wide sequencing of longan (*Dimocarpus longan* Lour.) provides insights into molecular basis of its polyphenol-rich characteristics. *Gigascience.* **6**, 1–14 (2017). <https://doi.org/10.1093/gigascience/gix023>
13. Li, H.T., et al.: Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants.* **5**, 461–470 (2019). <https://doi.org/10.1038/s41477-019-0421-0>
14. Zhang, L., et al.: The water lily genome and the early evolution of flowering plants. *Nature.* **577**, 79–84 (2020). <https://doi.org/10.1038/s41586-019-1852-5>
15. Zhu, C., et al.: The gap-free genome and multi-omics analysis of *Citrus reticulata* 'Chachi' reveal the dynamics of fruit flavonoid biosynthesis. *Hortic. Res.* **11**, uhae177 (2024). <https://doi.org/10.1093/hr/uhae177>
16. Alexander, D.H., Novembre, J., Lange, K.: Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009). <https://doi.org/10.1101/gr.094052.109>
17. Patterson, N., et al.: Ancient admixture in human history. *Genetics.* **192**, 1065–1093 (2012). <https://doi.org/10.1534/genetics.112.145037>
18. Pickrell, J., Pritchard, J.: Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012). <https://doi.org/10.1371/journal.pgen.1002967>
19. Li, Q., et al.: Haplotype-resolved T2T genome assemblies and pangenome graph of pear reveal diverse patterns of allele-specific expression and the genomic basis of fruit quality traits. *Plant. Commun.* **5** (2024). <https://doi.org/10.1016/j.xplc.2024.101000>
20. Wu, G.A., et al.: Genomics of the origin and evolution of Citrus. *Nature.* **554**, 311–316 (2018). <https://doi.org/10.1038/nature25447>
21. Richter, M., Rossello-Mora, R.: Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U S A.* **106**, 19126–19131 (2009). <https://doi.org/10.1073/pnas.0906412106>
22. Schumer, M., Rosenthal, G.G., Andolfatto, P.: How common is homoploid hybrid speciation? *Evolution* **68**, 1553–1560 (2014). <https://doi.org/10.1111/evo.12399>
23. Wang, Z., et al.: Genomic evidence for homoploid hybrid speciation between ancestors of two different genera. *Nat Commun* **13**, (1987). (2022) <https://doi.org/10.1038/s41467-022-29643-4>
24. Sanders, H.: Hybrid fertility and the rarity of homoploid hybrid speciation. *AoB PLANTS.* **17**, plaf035 (2025). <https://doi.org/10.1093/aobpla/plaf035>

25. Diaz, I.A., et al.: A haplotype-resolved chromatin landscape connects cis-regulatory variants to trait variation in Citrus. *BMC Genom.* **26**, 978 (2025). <https://doi.org/10.1186/s12864-025-12137-0>
26. Butelli, E., et al.: Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant. Cell.* **24**, 1242–1255 (2012). <https://doi.org/10.1105/tpc.111.095232>
27. Sakthivel, S.K., Vennapusa, A.R., Melmaiee, K.: Enhancing quality and climate resilient traits in vegetatively propagated polyploids: transgenic and genome editing advancements, challenges and future directions. *Front. Genet.* **16**, 1599242 (2025). <https://doi.org/10.3389/fgene.2025.1599242>
28. Botet, R., Keurentjes, J.: The role of transcriptional regulation in hybrid vigor. *Front. Plant. Sci.* **11**, 410 (2020). <https://doi.org/10.3389/fpls.2020.00410>
29. Marcais, G., Kingsford, C.: A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* **27**, 764–770 (2011). <https://doi.org/10.1093/bioinformatics/btr011>
30. Ranallo-Benavidez, T.R., Jaron, K.S., Schatz, M.C.: GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020). <https://doi.org/10.1038/s41467-020-14998-3>
31. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li, H.: Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods.* **18**, 170–175 (2021). <https://doi.org/10.1038/s41592-020-01056-5>
32. Servant, N., et al.: HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015). <https://doi.org/10.1186/s13059-015-0831-x>
33. Zhou, C., McCarthy, S.A., Durbin, R.: YaHS: yet another Hi-C scaffolding tool. *Bioinformatics.* **39**, btac808 (2023). <https://doi.org/10.1093/bioinformatics/btac808>
34. Xu, M., et al.: TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience.* **9** (2020). <https://doi.org/10.1093/gigascience/giaa094>
35. Zeng, X., et al.: Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nat. Plants.* **10**, 1184–1200 (2024). <https://doi.org/10.1038/s41477-024-01755-3>
36. Manni, M., Berkeley, M.R., Seppey, M., Zdobnov, E.M.: BUSCO: assessing genomic data quality and beyond. *Curr. Protoc.* **1**, e323 (2021). <https://doi.org/10.1002/cpz1.323>
37. Ou, S., Jiang, N.: LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant. Physiol.* **176**, 1410–1422 (2018). <https://doi.org/10.1104/pp.17.01310>
38. Nachtweide, S., Stanke, M.: Multi-genome annotation with AUGUSTUS. *Methods Mol Biol* 139–160 (2019). (1962). https://doi.org/10.1007/978-1-4939-9173-0_8
39. Korf, I., Flicek, P., Duan, D., Brent, M.R.: Integrating genomic homology into gene structure prediction. *Bioinformatics.* **17**, S140–S148 (2001). https://doi.org/10.1093/bioinformatics/17.suppl_1.S140

40. Majoros, W.H., Pertea, M., Salzberg, S.L.: TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*. **20**, 2878–2879 (2004).
<https://doi.org/10.1093/bioinformatics/bth315>
41. Campbell, M.S., Holt, C., Moore, B., Yandell, M.: Genome annotation and curation using MAKER and MAKER-P. *Curr. Protoc. Bioinf.* **48**, 4–11 (2014). <https://doi.org/10.1002/0471250953.bi0411s48>
42. Lowe, T.M., Chan, P.P.: tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **44**, W54–57 (2016). <https://doi.org/10.1093/nar/gkw413>
43. Nawrocki, E.P., Eddy, S.R.: Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. **29**, 2933–2935 (2013). <https://doi.org/10.1093/bioinformatics/btt509>
44. 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). https://doi.org/10.1007/978-1-0716-1134-0_4
45. Lin, Y., et al.: quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic. Res.* **10**, uhad127 (2023). <https://doi.org/10.1093/hr/uhad127>
46. Gremme, G., Steinbiss, S., Kurtz, S.: GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **10**, 645–656 (2013). <https://doi.org/10.1109/TCBB.2013.68>
47. Gary, B.: Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 573–580 (1999). <https://doi.org/10.1093/nar/27.2.573>
48. Krzywinski, M., et al.: Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009). <https://doi.org/10.1101/gr.092759.109>
49. Li, H.: Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. **34**, 3094–3100 (2018). <https://doi.org/10.1093/bioinformatics/bty191>
50. 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).
<https://doi.org/10.1186/s13059-019-1911-0>
51. Goel, M., Schneeberger, K.: plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics*. **38**, 2922–2926 (2022).
<https://doi.org/10.1093/bioinformatics/btac196>
52. Marcais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., Zimin, A.: MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
<https://doi.org/10.1371/journal.pcbi.1005944>
53. Zheng, J., Meinhardt, L.W., Goenaga, R., Matsumoto, T., Zhang, D., Yin, Y.: The chromosome-level rambutan genome reveals a significant role of segmental duplication in the expansion of resistance genes. *Hortic. Res.* **9**, uhac014 (2022). <https://doi.org/10.1093/hr/uhac014>
54. Tang, H., et al.: JCVI: A versatile toolkit for comparative genomics analysis. *IMeta* **3**, e211 (2024).
<https://doi.org/10.1002/imt2.211>

55. Pertea, G., Pertea, M.G.F.F., Utilities: GffRead and GffCompare. *F1000Res* 9, ISCB-Comm (2020). <https://doi.org/10.12688/f1000research.23297.2>
56. Emms, D.M., Kelly, S.: OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019). <https://doi.org/10.1186/s13059-019-1832-y>
57. Katoh, K., Kuma, K., Toh, H., Miyata, T.: MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005). <https://doi.org/10.1093/nar/gki198>
58. Sun, P., et al.: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant.* **15**, 1841–1851 (2022). <https://doi.org/10.1016/j.molp.2022.10.018>
59. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermini, L.S.: ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods.* **14**, 587–589 (2017). <https://doi.org/10.1038/nmeth.4285>
60. 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). <https://doi.org/10.1093/molbev/msaa015>
61. Bouckaert, R., et al.: BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019). <https://doi.org/10.1371/journal.pcbi.1006650>
62. Yang, Z.: Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994). <https://doi.org/10.1007/BF00160154>
63. Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G.: Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**, 1185–1192 (2005). <https://doi.org/10.1093/molbev/msi103>
64. Ayres, D.L., et al.: BEAGLE 3: Improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Syst Biol* 68, 1052–1061 (2019). <https://doi.org/10.1093/sysbio/syz020>
65. Vasimuddin, M., Misra, S., Li, H., Aluru, S.: Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In: *2019 IEEE international parallel and distributed processing symposium (IPDPS)*. IEEE (2019). <https://doi.org/10.1109/IPDPS.2019.00041>
66. Li, H., et al.: The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* **25**, 2078–2079 (2009). <https://doi.org/10.1093/bioinformatics/btp352>
67. McKenna, A., et al.: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010). <https://doi.org/10.1101/gr.107524.110>
68. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J.: Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4, s13742-13015 (2015). <https://doi.org/10.1186/s13742-015-0047-8>
69. Patterson, N., Price, A.L., Reich, D.: Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006). <https://doi.org/10.1371/journal.pgen.0020190>

70. Kumar, S., Stecher, G., Suleski, M., Sanderford, M., Sharma, S., Tamura, K.: MEGA12: molecular evolutionary genetic analysis version 12 for adaptive and green computing. *Mol. Biol. Evol.* **41** (2024). <https://doi.org/10.1093/molbev/msae263>
71. Danecek, P., et al.: Twelve years of SAMtools and BCFtools. *Gigascience* 10, (2021). <https://doi.org/10.1093/gigascience/giab008>
72. Sun, M., et al.: Haplotype-resolved, gap-free genome assemblies provide insights into the divergence between Asian and European pears. *Nat. Genet.* **57**, 2040–2051 (2025). <https://doi.org/10.1038/s41588-025-02273-4>
73. Wu, T.D., Watanabe, C.K.: GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics.* **21**, 1859–1875 (2005). <https://doi.org/10.1093/bioinformatics/bti310>
74. Wu, T., et al.: clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation.* **2**, 100141 (2021). <https://doi.org/10.1016/j.xinn.2021.100141>
75. Dobin, A., et al.: STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* **29**, 15–21 (2013). <https://doi.org/10.1093/bioinformatics/bts635>
76. Liao, Y., Smyth, G.K., Shi, W.: featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* **30**, 923–930 (2014). <https://doi.org/10.1093/bioinformatics/btt656>
77. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014). <https://doi.org/10.1186/s13059-014-0550-8>

Figures

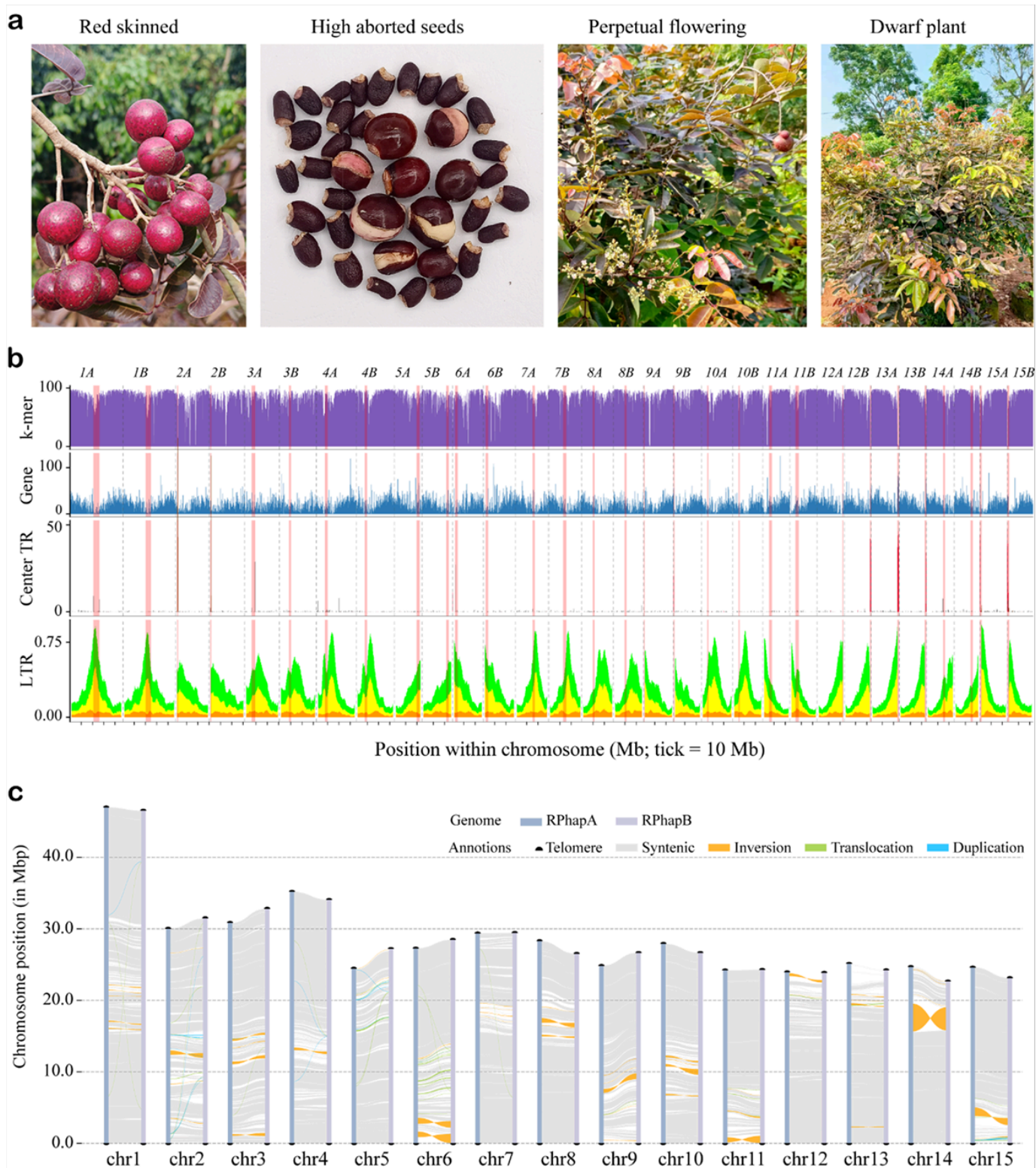


Figure 1

Complete red-skinned longan reference-grade genome T2T-RP. **(a)** Agronomic traits of high breeding value. **(b)** Chromosomal characterization by k-mer and gene densities, repeat distribution, and centromere and telomere prediction. **(c)** Pairwise comparison of homologous chromosomes.

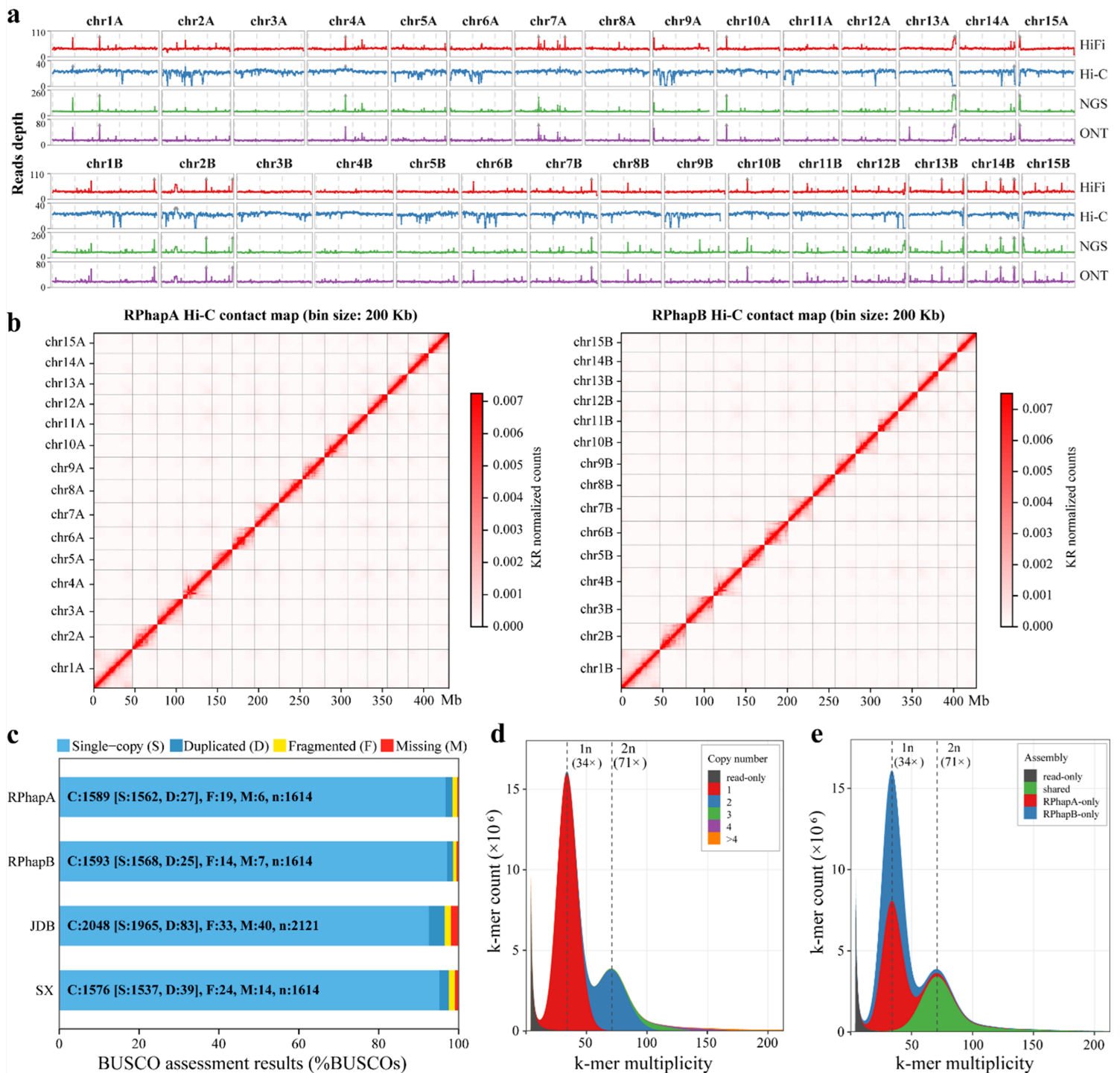


Figure 2

Quality assessments of the T2T-RP genome using multi-layered approaches. **(a)** Sequencing depth along the 15 chromosomes of haplotype RPhapA and haplotype RPhapB, calculated in 200-kb non-overlapping windows for PacBio HiFi (red), Hi-C (blue), Illumina short reads (NGS, green) and ONT (purple) data. Arrows indicate windows whose depth exceeds the plotted axis range. **(b)** Genome-wide Hi-C contact maps of haplotype RPhapA and haplotype RPhapB at 200-kb resolution, coloured by KR-normalised contact frequency. **(c)** BUSCO completeness of the two haplotypes and of two previously published longan genomes (JDB and SX) used for comparison. **(d)** Merqury copy-number spectrum. Dashed lines

mark the 1n and 2n coverage peaks. **(e)** Merqury assembly spectrum. Dashed lines mark the 1n and 2n coverage peaks.

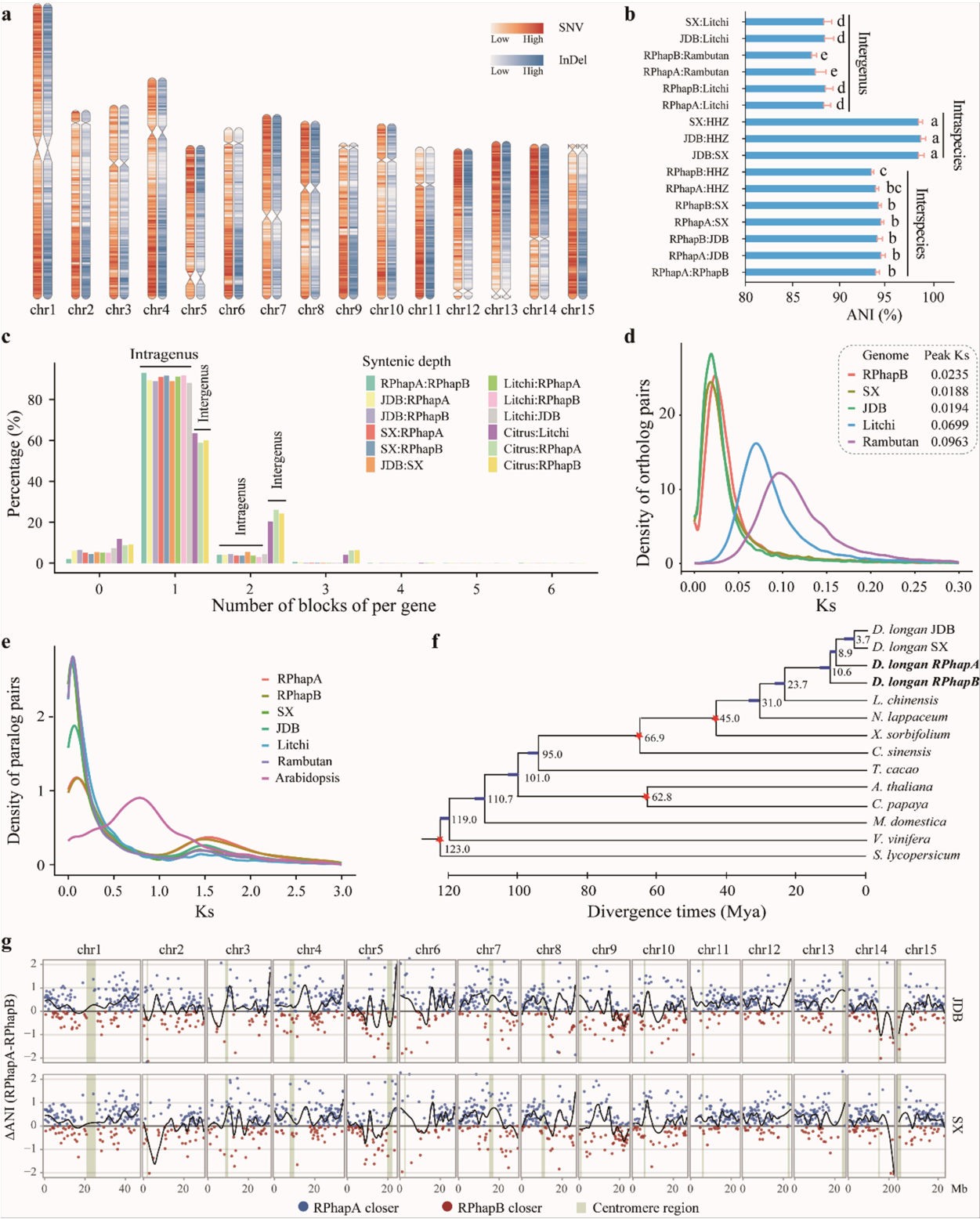


Figure 3

The two RP haplotypes derive from deeply diverged *Dimocarpus* lineages. **(a)** Divergence between homologous chromosomes, as revealed by single-nucleotide polymorphisms (SNPs) and

insertions/deletions (InDels). Chromosome numbering and orientation are based on previously published genome assemblies ^{5, 9, 12}. **(b)** Average nucleotide identity (ANI) among RP haplotypes and other Sapindaceae genomes. **(c)** Syntenic depth analysis between longan and related species. **(d)** Distribution of synonymous substitution rates (K_s) among orthologous gene pairs. **(e)** K_s distribution among paralogous gene pairs, indicating whole genome duplication events. **(f)** Phylogenetic reconstruction and molecular dating of longan and 10 other species based on single copy orthologs. The red diamond marks the fossil calibration point ^{13, 14}, and blue bars denote 95% highest posterior density intervals for divergence time estimates, *S. lycopersicum* was used as the outgroup. **(g)** Δ ANI along the 15 chromosomes of RPhapA against two cultivar references. Δ ANI = ANI(RPhapA–reference) – ANI(RPhapB–reference).

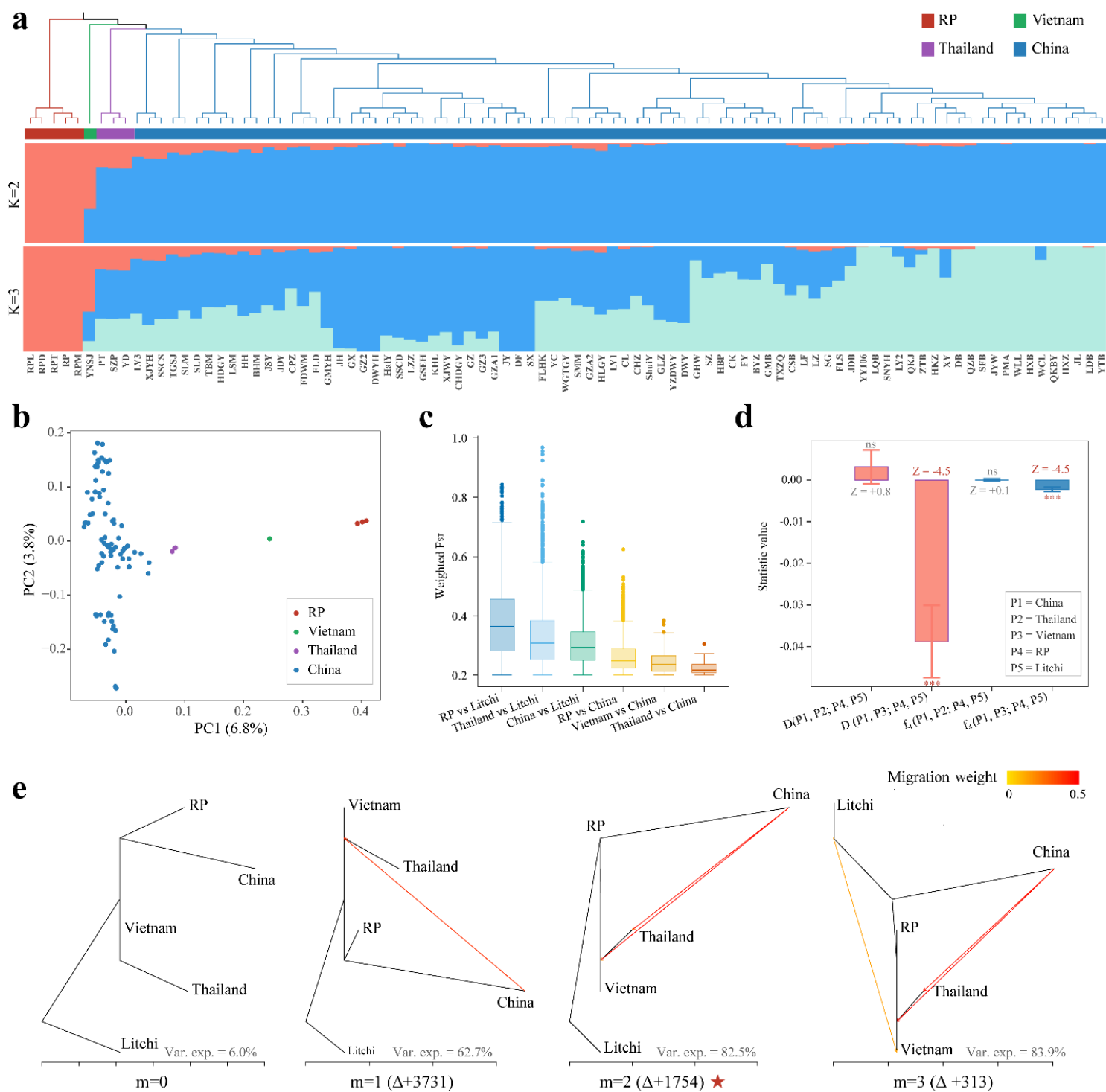


Figure 4

Population genomics reveals a red-skinned-longan-specific ancestral component. **(a)** Maximum-likelihood phylogenetic tree and ADMIXTURE bar plots for $K = 2$ and $K = 3$. The tree is rooted with *Litchi chinensis* as the outgroup. Branch colours indicate population affiliation. **(b)** Principal component analysis (PCA) of 91 longan accessions. PC1 explains 6.8% and PC2 explains 3.8% of total variance. **(c)** Genome-wide distributions of weighted F_{ST} for six pairwise population comparisons, calculated using 100-kb sliding windows. **(d)** f_4 -statistics and D-statistics testing for excess allele sharing between

population pairs relative to RP and the Litchi outgroup. Error bars represent $\pm$ SE. Z-scores are annotated; significant results ($|Z| \geq 3$) are indicated by asterisks (***) $P < 0.001$; ns, not significant. **(e)** TreeMix graphs for $m = 0$ to 3 migration events, rooted with *L. chinensis*. Migration edges are coloured by weight. The optimal model ($m = 2$, $\boxtimes$) is highlighted.

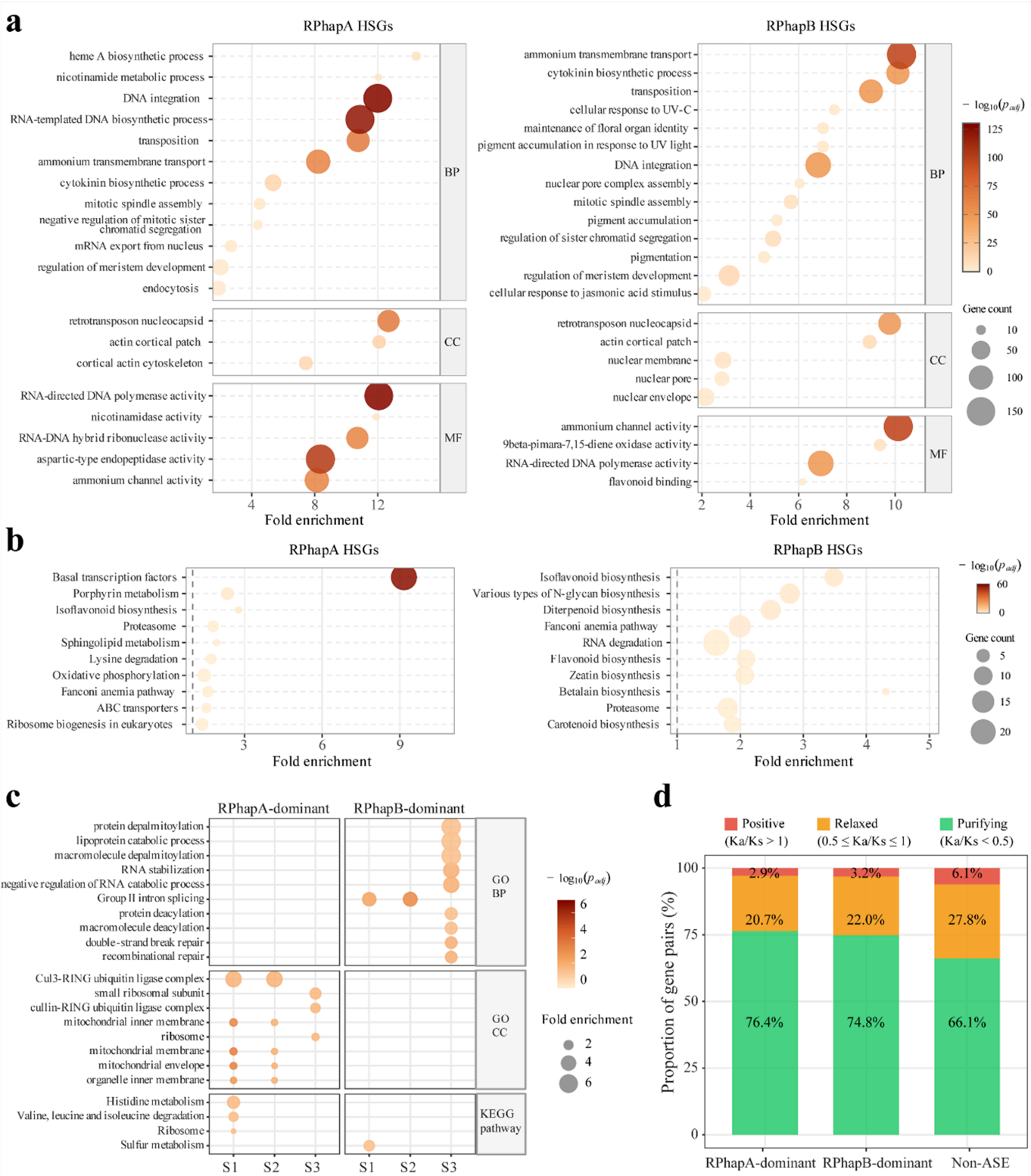


Figure 5

Functional characterisation of haplotype-specific genes and allele-specific expression in RP longan. **(a)** GO enrichment bubble plots for RPhapA and RPhapB specific genes (HSGs). **(b)** KEGG pathway enrichment for RPhapA and RPhapB HSGs. **(c)** Bubble heatmap of significantly enriched GO and KEGG terms for RPhapA-dominant and RPhapB-dominant ASE gene sets at each developmental stage. **(d)** Proportion of gene pairs under positive ($K_a/K_s > 1$), relaxed (0.5 to 1), and purifying selection ($K_a/K_s < 0.5$) for each ASE category.

Supplementary Files

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

- [SupplementaryFigures.docx](#)
- [supplementaryTables.xlsx](#)