Haplotype-resolved ancestry profiling reveals mosaic architecture of the lemon genome

Xiaobo Long
Hunan University
Yushi Wei
Huazhong Agricultural University
Haomin Lyu
Haomin.Lyu@hotmail.com

Higentec Breeding Innovation (ZheJiang) Co.,Ltd
Fanzai Zeng
Hunan University

Research Article

Keywords: haplotype, Citrus, ancestral profiling, functional divergence, Ka/Ks, expression bias

Posted Date: July 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10109139/v1

Additional Declarations: No competing interests reported.

Abstract

Lemon (*Citrus × limon*) is a highly heterozygous allopolyploid derived from three ancestral species: mandarin, pummelo, and citron. Despite its agronomic importance, the functional consequences of this multi-parental origin at the haplotype level remain unexplored. Here we re-analysed publicly available telomere-to-telomere (T2T) genome assemblies and transcriptomes to perform haplotype-resolved ancestry assignment, sequence divergence (Ka/Ks) analysis, and allele-specific expression profiling. Using a phylogenomic best-reciprocal-hit approach, we demonstrate that the two lemon haplotypes exhibit a striking mosaic architecture: the sour-orange-like haplotype (Cli_hapA) derives predominantly from mandarin (38.3%) and pummelo (23.5%), whereas the citron-like haplotype (Cli_hapB) is primarily citron-derived (47.4%). Sliding-window visualisation reveals large, pure ancestry blocks with abrupt transitions. Ks distributions confirm these affinities, and approximately 120–130 positively selected genes (Ka/Ks > 1) were identified per comparison. Allele-specific expression analysis identified 7,515 significantly biased gene pairs. Notably, mandarin- and pummelo-derived genes in Cli_hapA are expressed at significantly higher levels than their citron-derived orthologs in Cli_hapB, indicating that ancestral components differ not only in presence but also in transcriptional activity. Our study provides the first haplotype-resolved demonstration that citrus reticulate evolution produces not only structural mosaicism but also functional mosaicism at the subgenome level. The systematically higher expression of mandarin- and pummelo-derived alleles in the sour-orange-like haplotype identifies a set of transcriptionally active ancestral haplotypes that can be prioritised in breeding. The integrated framework—applicable to any multi-parental hybrid with publicly available T2T genomes—offers new opportunities for understanding heterosis and provides a marker-ready inventory of ancestry-biased alleles for breeding programmes targeting advantageous alleles from multiple progenitors.

1. Introduction

The phenomenon of heterosis—hybrid vigour—has long fascinated biologists and breeders, yet its mechanistic basis remains incompletely understood, particularly in allopolyploids that merge two or more divergent genomes into a single nucleus (Castle 1903; Chen 2013). An allopolyploid must orchestrate the expression of two (or more) sets of alleles, often derived from different parental species, to achieve superior phenotypes without detrimental conflicts (Van de Peer et al. 2009; Cheng et al. 2018). How this coordination is achieved at the molecular level is a central question in evolutionary and functional genomics.

Two major explanatory frameworks have emerged. The subgenome dominance hypothesis posits that one parental genome globally suppresses the other in terms of expression level and functional contribution, often driven by epigenetic asymmetries and transposable element dynamics (Cheng et al. 2018; Bird et al. 2018; Alger and Edger 2020). The subgenome functional complementation hypothesis, by contrast, suggests that the two subgenomes divide labour across different biological pathways, each taking primary responsibility for a distinct set of functions, and together they exceed the adaptive capacity of either progenitor (Conant and Wolfe 2008; Jiao et al. 2011). These models are not mutually

exclusive; indeed, both mechanisms could operate simultaneously. However, a critical unknown remains: how does such a division of labour arise from the hybrid's multi-parental evolutionary history? Answering this question requires moving beyond bulk genome assemblies to a haplotype-resolved, ancestry-aware framework.

In plants, complex, multi-parental hybridisation patterns are widespread across many lineages. For instance, the genus *Rosa* comprises approximately 150 species, and hybridisation and polyploidy are common throughout its evolution (Debray et al. 2022; Zhang et al. 2024b, 2026). The Orchidaceae family frequently undergoes interspecific hybridisation and reticulate evolution, posing significant challenges for phylogenetic studies (Givnish et al. 2015; Pérez-Escobar et al. 2024; Freudenstein 2025). Recent research on *Dendrobium* has also revealed haplotype-level variation and allelic imbalance linked to diverse ancestors in hybrid genomes (Wang et al. 2025; Chen et al. 2025; Li et al. 2026). In the octoploid species of Phyllanthaceae, haplotype-resolved genome assemblies have similarly highlighted the critical roles of polyploidisation and hybridisation in speciation (Li et al. 2024). Among these, the genus *Citrus* offers an exceptionally well-documented case of reticulate evolution driven by a small number of ancestral species (Wu et al. 2018; Droc et al. 2026).

Cultivated citrus species arose through reticulate evolution among a handful of basic species—primarily mandarin (*C. reticulata*), pummelo (*C. maxima*), and citron (*C. medica*)—through repeated hybridisation and backcrossing (Wu et al. 2018; Rao et al. 2021). As a result, most citrus genomes are not simple admixtures but structural mosaics of ancestral chromosome segments. This structural mosaicism has been elegantly documented using comparative genomics and k-mer-based ancestry deconvolution (Wu et al. 2014, 2018; Wang et al. 2018). For example, the sweet orange genome is known to be a mosaic of pummelo and mandarin segments, and lemon inherits blocks from all three basic species (Wu et al. 2023; Liu et al. 2025b). As recently highlighted in the hexaploid oat genome, such mosaic architectures can result from extensive inter-subgenome translocations and can create breeding barriers, yet they also offer opportunities to study subgenome expression balance and functional partitioning (Kamal et al. 2022; Avni et al. 2026). However, a similar approach has not been systematically applied to citrus.

Lemon (*C. × limon*) offers an ideal system to bridge this gap. Lemon is a highly heterozygous allopolyploid derived from citron, pummelo, and mandarin. Its breeding history is well understood: lemon originated from a cross between sour orange (itself a pummelo-mandarin hybrid) and citron (Wu et al. 2014, 2018). Crucially, a fully assembled, gap-free, haplotype-resolved telomere-to-telomere (T2T) genome of lemon is publicly available (Bao et al. 2023), together with transcriptome data from multiple tissues. In addition, high-quality gap-free or chromosome-scale genomes of the three progenitor species—mandarin (Zhu et al. 2024), pummelo (Liu et al. 2025a), and citron (Feng et al. 2025)—have recently been published. This suite of resources allows us to perform a high-resolution, haplotype-level dissection of ancestry and its functional consequences using only re-analysed public data.

We therefore tested the hypothesis that citrus reticulate evolution not only produced structural mosaicism of ancestral segments but also drove functional mosaicism between haplotypes. Using a

phylogenomic best-reciprocal-hit pipeline, we assigned parental ancestry at the gene level, computed Ks and Ka/Ks for orthologous pairs, called allele-specific expression bias, and performed functional enrichment. This integrated approach allowed us to directly test whether the structural mosaic inherited from reticulate evolution is converted into a haplotype-level functional mosaic, providing a generalisable computational framework for studying multi-parental hybrids using only public data.

2. Materials and Methods

2.1. Data preparation

All analyses are based on the systematic re-use of publicly available genomic and transcriptomic resources; no new sequencing data were generated. The haplotype-resolved telomere-to-telomere (T2T) genome assembly of lemon (*Citrus × limon*) was retrieved from the Genome Warehouse (GWH) database of the National Genomics Data Center (https://ngdc.cncb.ac.cn/gwh/) under accession number GWHCBFQ00000000.1 (Bao et al. 2023). For ancestral origin assignment and comparative evolutionary analyses, we obtained three additional genomes representing the wild progenitor species: a gap-free genome assembly of mandarin (*Citrus reticulata* 'Chachi'; CNGB accession: CNP0005504) (Zhu et al. 2024); a gap-free T2T assembly of pummelo (*Citrus maxima* 'Pingshan'; NCBI accessions: JBKJAM000000000) (Liu et al. 2025a); and a chromosome-scale genome assembly of citron (*Citrus medica*; GWH accession: GWHETLN00000000.1) (Feng et al. 2025). Transcriptome data for lemon consisted of RNA-seq libraries from two tissues leaf and bud.

To ensure consistent chromosome identification and orientation across the four species, we performed whole-genome synteny analysis using MUMmer (v4.0.0) with the nucmer program (Marçais et al. 2018). Based on collinear blocks, the chromosome numbering and orientation of the lemon, mandarin, pummelo, and citron genomes were systematically adjusted to match a common reference (the lemon T2T assembly). This standardisation guaranteed that all downstream analyses (ancestry assignment, Ka/Ks calculation, and expression bias comparison) were performed on correctly paired homologous chromosomes.

2.2. Standardized genome annotation for cross-haplotype comparability

To enable direct comparison across the five haplotypes (lemon hapA and hapB, mandarin, pummelo, and citron), we performed a unified annotation of repetitive elements and protein-coding genes, adapted from our Lablab purpureus protocol. Repeat annotation was conducted using EDTA (v2.2.1) (Ou et al. 2019), which combines homology-, structure-, and *de novo*-based methods. Repeat sequences were then masked with RepeatMasker (v4.1.5) prior to gene prediction.

Protein-coding gene annotation integrated three approaches. Homology-based prediction used GeMoMa (v1.9) with reference proteomes from six well-annotated plant species (Keilwagen et al. 2019). Transcriptome evidence incorporated RNA-seq reads (aligned with Hisat2 v2.2.1) (Kim et al. 2019) and,

where available, PacBio Iso-Seq data processed through the Iso-Seq3 pipeline (https://github.com/PacificBiosciences/IsoSeq); coding sequences were extracted with TransDecoder (v5.7.0) (https://github.com/TransDecoder/TransDecoder). *De novo* prediction employed Braker2 (v2.1.2) (Brůna et al. 2021), integrating Augustus (Stanke et al. 2006) and GeneMark-ET (Lomsadze et al. 2005), trained on RNA-seq alignments and Swiss-Prot proteins (Bairoch and Apweiler 1996). All predictions were merged using EVidenceModeler (v2.0.0) (Haas et al. 2008). Truncated and non-coding models were removed, yielding a high-confidence gene set per haplotype. Functional annotation was performed with eggNOG-mapper (v2.1.12) (Cantalapiedra et al. 2021) against the eggNOG database (v5.0.2) (Hernández-Plaza et al. 2026), assigning GO terms and KEGG Orthology identifiers.

2.3. Ancestral origin assignment for lemon haplotypes using a phylogenomic approach

To trace the parental ancestry of each lemon haplotype (*Citrus × limon*, Cli_hapA and Cli_hapB) at the gene level, we performed a best-reciprocal-hit assignment against the protein sequences of the three putative progenitor species: mandarin (*Citrus reticulata*, Cre), pummelo (*Citrus maxima*, Cma), and citron (*Citrus medica*, Cme). The protein sets of the three ancestral species were obtained from the standardized annotation described as above.

First, the sequence headers of the lemon T2T genome assembly were standardized to ensure consistent identification of gene loci. A local BLAST database was constructed from the concatenated ancestral proteomes. For rapid and sensitive homology searching, DIAMOND (v2.0.15) (Buchfink et al. 2015) was executed with an identity threshold of ≥ 95% and an e-value cutoff of 1e-5. For each lemon gene (from Cli_hapA and Cli_hapB separately), we performed a bidirectional best-hit strategy: the top hit in the ancestral database was retrieved, and then the corresponding ancestral protein sequence was searched back against the lemon proteome. If the top hit in both directions matched the same lemon–ancestor pair, the assignment was accepted. Genes that did not meet this reciprocal criterion or fell below the identity threshold were classified as Unsure (UN). Based on the best-hit ancestral species, each lemon gene was categorised into one of four classes: Cre-derived, Cma-derived, Cme-derived, or UN.

To visualise the distribution of ancestral origins along each chromosome and to identify mosaic regions, we applied two complementary sliding-window approaches: (1) physical-window analysis (1 Mb window): the proportion of genes belonging to each ancestry category was calculated for each window; (2) gene-count-based window analysis (500 genes): for each window, the relative proportions of Cre-, Cma-, Cme-derived, and UN genes were calculated. Based on the proportions, windows were further categorised into pure Cre, pure Cma, pure Cme, mosaic, unsure, or NA according to predefined thresholds (e.g., pure Cre: Cre ≥ 70% and Unsure < 10%). The resulting data were used to generate stacked area plots, which clearly visualised transitions in ancestry dominance along each chromosome.

All analyses were performed separately for lemon Cli_hapA and Cli_hapB. The resulting high-confidence, haplotype-resolved maps of parental ancestry served as the structural foundation for all subsequent functional interpretations, including expression bias and selection pressure analyses.

2.4. Sequence divergence and selection pressure (Ka/Ks) analysis

To assess the evolutionary divergence and selection pressures acting on the two lemon haplotypes (Cli_hapA and Cli_hapB), we performed pairwise synonymous (Ks) and non-synonymous (Ka) substitution rate calculations. Orthologous gene pairs between each lemon haplotype and the three progenitor genomes (mandarin Cre, pummelo Cma, and citron Cme) were identified based on the standardized annotations and synteny-adjusted chromosome coordinates described above. Collinearity relationships were established using the JCVI toolkit (v1.0.7) with default parameters (Tang et al. 2024).

For each orthologous pair, coding sequences (CDS) were aligned at the codon level using the 'para' module in TBtools (v1.12) (Chen et al. 2020). Ka and Ks values were then calculated using the YN00 model (Yang and Nielsen 2000). The following six pairwise comparisons were performed: (1) Cli_hapA vs. Cre, (2) Cli_hapA vs. Cma, (3) Cli_hapA vs. Cme, (4) Cli_hapB vs. Cre, (5) Cli_hapB vs. Cma, and (6) Cli_hapB vs. Cme.

For each comparison, we computed the global distributions of Ks and the Ka/Ks ratio using boxplots. Genes with Ka/Ks > 1 were considered to be under positive selection. Statistical significance of Ks differences between comparison groups (e.g., Cli_hapA vs. Cre versus Cli_hapA vs. Cma) was evaluated using the Wilcoxon rank-sum test.

2.5. Allele-specific expression and expression bias calling

RNA-seq reads from lemon leaves and buds were pre-processed to remove adapter sequences and low-quality bases using Trimmomatic (v0.39) with default parameters (Bolger et al. 2014). Cleaned reads were then aligned to the haplotype-resolved lemon T2T genome assembly (Cli_hapA and Cli_hapB concatenated) using HISAT2 (v2.2.1) in splice-aware mode (Kim et al. 2019). To ensure accurate allele-specific assignment, we applied the parameter '--outFilterMultimapNmax 1' to retain only uniquely mapping reads. The resulting BAM files were sorted and indexed with SAMtools (v1.10) (Danecek et al. 2021).

For each collinear gene pair between Cli_hapA and Cli_hapB (identified from synteny analysis), read counts mapping uniquely to each allele were extracted using featureCounts (v2.0.1) from the subread package (Liao et al. 2014), with the '-s 2' flag for strand-specific reverse strandedness. Only gene pairs with a combined read count ≥ 10 across all samples were retained.

Expression bias was defined as a significant deviation from the expected 1:1 allelic ratio. For each gene pair and each tissue, we applied a two-tailed binomial exact test to the allele-specific counts. The resulting p-values were adjusted for multiple testing using the Benjamini-Hochberg false discovery rate (FDR) procedure. Gene pairs with FDR < 0.05 were considered significantly biased. The direction of bias (Cli_hapA-biased or Cli_hapB-biased) was recorded, and biased genes were assigned to their parental origin (Cre-, Cma-, Cme-derived, or UN) using the ancestry classification from Section 3.

To compare expression levels between haplotypes for specific ancestry groups, we calculated transcripts per million (TPM) for each gene in each haplotype using the 'rsem-calculate-expression' module from RSEM (v1.3.3) (Li and Dewey 2011). For collinear gene pairs where the Cli_hapA gene was Cre-derived (or Cma-derived) and its Cli_hapB ortholog was Cme-derived, we compared their TPM values using the Wilcoxon signed-rank test (paired by gene pair).

2.6. Statistical analysis and visualisation

All statistical analyses were conducted in R (v4.2.0). The Wilcoxon rank-sum test was used for unpaired two-group comparisons (e.g., Ks distributions across different comparison groups in Fig. 3). The Wilcoxon signed-rank test was used for paired comparisons (e.g., expression levels of orthologous gene pairs in Fig. 5). Figures were generated using the 'ggplot2' package (v3.3.6). Boxplots, bar plots, and scatter plots were produced following the parameters detailed above.

3. Results

3.1. Standardized re-annotation of four citrus genomes enables cross-haplotype comparability

To enable precise haplotype-resolved ancestry and functional comparisons, we collected publicly available high-quality genome assemblies of four citrus species closely related to lemon's progenitors: mandarin (*Citrus reticulata* 'Chachi', gap-free T2T), pummelo (*Citrus maxima* 'Pingshan', gap-free T2T), citron (*Citrus medica* 'Tibetan', chromosome-scale), and lemon (*Citrus × limon* 'Perfume Lemon', haplotype-resolved T2T) (Bao et al. 2023; Zhu et al. 2024; Feng et al. 2025; Liu et al. 2025a). The lemon assembly consists of two fully separated haplotypes (Cli_hapA and Cli_hapB). The total assembly sizes range from 312.97 Mb (mandarin) to 633.0 Mb (lemon, sum of both haplotypes). Contig N50 values are high (32.18–37.51 Mb), reflecting the gap-free or near-gap-free nature of most assemblies. All genomes were retrieved from public databases and their chromosome numbering and orientation were unified using whole-genome synteny analysis using the lemon T2T assembly as the reference.

Using a unified annotation pipeline, we generated comparable gene sets for all five haplotypes (Table 1). Protein-coding gene numbers vary: mandarin (23,289), pummelo (23,657), citron (23,277), and lemon (43,966 for both haplotypes combined). BUSCO completeness for genome assemblies is excellent (> 98% for all species). Gene-level BUSCO completeness is also high: 98.0% (mandarin), 98.1% (pummelo), 98.2% (citron), and 96.6% (lemon). Because the lemon assembly contains two complete haplotypes, most BUSCO genes are present as duplicate copies (D). This standardised annotation framework provides a solid foundation for subsequent ancestry assignment, evolutionary divergence, and expression bias analyses.

Table 1
Genome statistics of four citrus reference genomes.

Name	Mandarin	Pummelo	Citron	Lemon
Species	*Citrus reticulata* 'Chachi'	*Citrus maxima* 'Pingshan'	*Citrus medica* 'Tibetan'	*Citrus × limon* 'Perfume Lemon'
Assembly level	Gap-free (T2T)	Gap-free (T2T)	Chromosome-scale	Haplotype-resolved T2T
Genome size (Mb)	312.97	371.17	402.23	633.0 (total of two haplotypes)
Contig N50 (Mb)	32.18	32.72	37.51	35.6
Genome BUSCO completeness (%)	98.4	98.3	99	98.7
Number of protein-coding genes	23,289	23,657	23,277	43,966
Gene BUSCO completeness (%)	C:98.0% [S:97.0%,D:1.0%]	98.1% [S:96.5%,D:1.6%]	C:98.2% [S:96.2%,D:2.0%]	96.6% [S:3.7%,D:92.9%]
Reference	Zhu et al. 2024	Li et al. 2025	Feng et al. 2025	Bao et al. 2023

3.2. Distinct ancestral mosaicism revealed in the two lemon haplotypes

To trace the parental origin of the two lemon haplotypes, we first reconstructed the known hybridisation history of lemon (Fig. 1a). Lemon is derived from a cross between sour orange (itself an F1 hybrid of pummelo and mandarin) and citron. Based on this phylogenetic framework, we developed a phylogenomic assignment pipeline (Fig. 1b): for each gene in Cli_hapA and Cli_hapB, we performed a best reciprocal BLAST search (DIAMOND, identity ≥ 95%) against the protein sets of the three progenitor species – mandarin (*C. reticulata*, Cre), pummelo (*C. maxima*, Cma), and citron (*C. medica*, Cme). Genes were classified as Cre-derived, Cma-derived, Cme-derived, or Unsure (UN). This approach allowed us to generate high-resolution, haplotype-resolved ancestry maps (Fig. 1c), revealing a strikingly mosaic genome architecture.

Based on these maps, we re-evaluated the previously assigned parental attributes of chromosomes A and B. Originally, chr01A and chr01B were tentatively designated based on generic synteny; however, our high-confidence phylogenetic assignments demonstrate that Cli_hapA is predominantly derived from the sour orange (Cre + Cma) parent, while Cli_hapB is derived from the citron (Cme) parent (Fig. 1c; Supplementary Table 1). Consequently, we redefined the haplotype identities: Cli_hapA is now designated as the sour-orange-like haplotype, and Cli_hapB as the citron-like haplotype. All subsequent analyses are based on this corrected haplotype assignment.

Across the entire genome, Cli_hapA consists of 38.3% Cre-derived, 23.5% Cma-derived, 1.5% Cme-derived, and 36.7% unsure (UN) genes (Fig. 1c; Supplementary Table 1). For Cli_hapB, the proportions are 5.3% Cre, 8.1% Cma, 47.4% Cme, and 39.2% UN. The high proportion of UN genes in both haplotypes (~ 37–39%) likely reflects gene loss or unusal sequence divergence.

To visualise the fine-scale distribution of ancestry along each chromosome, we performed a sliding-window analysis and plotted the ancestry proportions as stacked bar charts across all 18 pseudochromosomes (Fig. 2). The results strongly support the haplotype redefinition above. Cli_hapA (chromosomes A01–A09) displays a highly mixed pattern, with Cre and Cma dominating most windows, while Cme is nearly absent. In contrast, Cli_hapB (chromosomes B01–B09) is overwhelmingly dominated by Cme-derived sequences across all chromosomes, with only sporadic windows showing minor Cre or Cma signals.

Per-chromosome analysis further highlighted the mosaic nature. In Cli_hapA, chromosomes such as Chr04A and Chr09A are strongly Cre-biased (54.2% and 48.9% Cre, respectively), whereas Chr07A shows clear Cma dominance (47.1%) (Fig. 2). Some chromosomes (e.g., Chr05A) maintain a balanced mixture of Cre and Cma. Notably, Chr09A is almost completely devoid of Cma and Cme contributions, being nearly pure Cre-derived. In Cli_hapB, all chromosomes are predominantly Cme-derived, with proportions ranging from 41.1% (Chr09B) to 52.6% (Chr05B). The UN fraction remains high across all chromosomes, consistent with the genome-wide averages.

These fine-scale ancestry maps provide the structural foundation for linking sequence divergence, expression bias, and functional specialization to specific ancestral blocks in the following analyses.

3.3. Co-variation of sequence divergence and selection with ancestry boundaries

To assess the evolutionary divergence of the two lemon haplotypes, we calculated synonymous substitution rates (Ks) and Ka/Ks ratios for orthologous gene pairs between each lemon haplotype and the three progenitor genomes (mandarin Cre, pummelo Cma, citron Cme). The results strongly support the ancestry assignments from Section 2 (Fig. 3).

For Cli_hapA, the global Ks distribution (Fig. 3a) shows that orthologs between Cli_hapA and Cre have significantly lower Ks values than those between Cli_hapA and Cma (Wilcoxon rank-sum test, $p < 0.001$). Ks values for Cli_hapA vs. Cme are intermediate but closer to those of Cre. This pattern is consistent with Cli_hapA being predominantly derived from the sour orange parent (a Cre–Cma hybrid), with Cre contributing more to the ancestry than Cma. For Cli_hapB (Fig. 3c), the lowest Ks values are observed in the Cli_hapB vs. Cme comparison, significantly lower than those for Cli_hapB vs. Cre or vs. Cma ($p < 0.001$). This confirms that Cli_hapB is primarily derived from the citron progenitor.

The Ka/Ks distributions (Fig. 3b, d) for all comparisons are centered well below 1, indicating that most orthologous genes have evolved under purifying selection. No substantial differences in Ka/Ks were

observed among the different comparison groups. Using a threshold of Ka/Ks > 1, we identified positively selected genes in each comparison: 126 genes in Cli_hapA vs. Cre, 130 genes in Cli_hapA vs. Cma, and 113 genes in Cli_hapB vs. Cme. These candidate genes may represent targets of adaptive evolution in the two haplotypes.

3.4. Genome-wide allele-specific expression bias organised by ancestry

To investigate whether the two lemon haplotypes exhibit differential expression patterns, we performed allele-specific expression analysis using RNA-seq data from leaf and bud tissues. For each collinear gene pair between Cli_hapA and Cli_hapB, we calculated the $\log_2$ fold change (hapA vs. hapB) and assessed significance using a binomial exact test with FDR correction (FDR < 0.05).

A total of 7,515 gene pairs were identified as significantly biased out of 17,799 expressed gene pairs (Fig. 4a). Among these, 3,835 genes were biased toward Cli_hapB and 3,680 genes were biased toward Cli_hapA, indicating a roughly balanced overall bias direction, with a slight preference for hapB. The distribution of $\log_2$ fold changes (Fig. 4a) shows that most significant biased genes exhibit moderate bias intensities, while a smaller fraction show extreme skew toward one haplotype.

Comparison of expression levels (TPM) between haplotypes for the significantly biased genes (Fig. 4b) reveals that genes with higher expression in one haplotype tend to have lower expression in the other, consistent with true allele-specific expression rather than global haplotype dominance. The relationship between bias intensity ($|\log_2 FC|$) and average expression level (Fig. 4c) demonstrates that bias intensity is largely independent of overall expression magnitude: both highly and lowly expressed genes can exhibit strong allelic bias, with no systematic correlation.

These 7,515 significantly biased gene pairs were further stratified by parental origin (Cre-, Cma-, Cme-derived, or UN) based on the ancestry assignments from Section 2, providing a foundation for linking expression bias to ancestral contributions.

3.5. Ancestry-specific expression bias reveals higher activity of Cre- and Cma-derived genes

To determine whether the ancestral origin of a gene influences its expression level, we compared, within collinear gene pairs, the expression of Cli_hapA genes of defined ancestry (Cre-derived or Cma-derived) against their corresponding Cli_hapB orthologs, which are almost exclusively Cme-derived (Fig. 5). For Cre-derived Cli_hapA genes (n = 1,738), their median expression level (TPM) was significantly higher than that of their Cme-derived Cli_hapB counterparts (n = 1,755) ($P < 0.001$, Wilcoxon signed-rank test). Similarly, Cma-derived Cli_hapA genes (n = 1,015) also showed significantly higher expression than their Cme-derived Cli_hapB orthologs (n = 1,009) ($P < 0.001$). In contrast, no such elevation was observed for Cme-derived genes in Cli_hapB when compared to any Cli_hapA group (data not shown). These results indicate that genes originating from the mandarin (Cre) and pummelo (Cma) ancestors, which reside in

the sour-orange-like haplotype (Cli_hapA), are systematically more highly expressed than their citron-derived counterparts in the other haplotype. Thus, the Cre and Cma ancestral components contribute more actively to the overall transcriptional output of biased genes in the lemon genome.

To further prioritise functionally relevant candidates, we intersected the three sets of ancestry-specific biased genes (Cre-derived Cli_hapA-biasedc, and Cme-derived Cli_hapB-biased) with genes showing Ka/Ks > 1 in the corresponding divergence comparisons (Section 3). This filtering yielded a shortlist of genes that are both evolutionarily accelerated and differentially expressed between haplotypes (Supplementary Table 2).

The genes listed exhibit evolutionary signatures consistent with positive selection (Ka/Ks > 1), suggestive of adaptive roles in citrus growth, fruit development, metabolism, and stress tolerance. Among those from Cma and Cme origins, several contribute to primary metabolism and nutrient synthesis. The UDP-glycosyltransferase NM_gene06351 is a key flavonoid glycosyltransferase responsible for the glycosylation of secondary metabolites, which affects both fruit taste and stress responses (Chen et al. 2024). The MATE transporter NM_gene08678 is involved in the transport of flavonoids and citrate, thereby influencing fruit color and acidity (Liu et al. 2022). Meanwhile, the ABC transporter family member NM_gene03950 has been shown to mediate auxin and flavonoid transport, directly regulating fruit development(Zhang et al. 2024a). Regarding secondary metabolism regulation and signaling, NM_gene45857 (F-box protein) has been identified as a pollen S-determinant in self-incompatibility, functioning in the ubiquitination of non-self S-RNase to prevent inbreeding(Cao et al. 2024).

For genes from the Cre origin, multiple families play roles in cellular defense and stress responses. The molecular chaperone BiP (NM_gene04602) maintains endoplasmic reticulum homeostasis and is upregulated under drought, osmotic, and cold treatments, demonstrating its involvement in abiotic stress tolerance (Guimarães et al. 2018). The pentatricopeptide repeat (PPR) protein NM_gene30493 has been identified as a candidate restorer-of-fertility (Rf) gene in citrus cytoplasmic male sterility, likely influencing pollen grain number by affecting mitochondrial gene expression (Goto et al. 2023). Additionally, NM_gene29459 (thaumatin-like protein) contribute to transcriptional regulation and pathogen defense, further highlighting the broad adaptive significance of this gene set in citrus (Li et al. 2025).

Collectively, these positively selected, ancestry-biased genes represent a prioritized list for marker-assisted selection (MAS) and genome editing. For example, the MATE transporter (NM_gene08678) and ABC transporter (NM_gene03950) can be developed into haplotype-specific PCR markers to introgress the highly expressed mandarin/pummelo alleles into citron-based rootstocks. The PPR gene (NM_gene30493) is a candidate for restoring male fertility in hybrid breeding, and the BiP chaperone (NM_gene04602) offers a target for drought tolerance improvement. We provide these 55 candidate genes in Supplementary Table 2 as a breeder-ready resource.

4. Discussion

In this study, we performed a systematic, haplotype-resolved ancestral dissection of the lemon genome by re-analysing publicly available T2T assemblies and transcriptome data. We developed a phylogenomic pipeline to assign each gene in the two lemon haplotypes to one of three progenitor species – mandarin, pummelo, or citron – and integrated this with Ka/Ks analysis and allele-specific expression profiling. Our primary objective was to move beyond conventional “mixed” genome assemblies and to resolve the functional consequences of lemon’s complex hybrid origin at the level of individual haplotypes.

To our knowledge, this is the first study to apply a phylogenomic best-reciprocal-hit approach to directly map parental ancestry in a highly heterozygous, multi-parental (more than two progenitors) citrus genome. The genetic constitution of lemon involves three ancestral species (mandarin, pummelo, and citron), and its genome exhibits a striking mosaic architecture that fundamentally differs from classical biparental hybrid models. Importantly, such complex, multi-parental hybridisation patterns are widespread across many plant lineages, occurring frequently in genera where interspecific crossing and polyploidisation are common. The analytical framework presented here – integrating high-quality T2T genomes, phylogenomic ancestry assignment, Ka/Ks selection pressure analysis, and allele-specific expression profiling – provides a transferable paradigm for in-depth studies of such complex groups. Unlike traditional methods that rely on a few molecular markers or plastid loci, our framework fully exploits genome-wide sequence information and haplotype-resolved assemblies, enabling accurate discrimination of genomic segments derived from different parental species and assessment of their functional activity. This strategy is particularly well-suited for: (1) natural hybrid populations where progenitor species are unknown or unavailable; (2) analysis of reticulate evolutionary events involving three or more ancestral species; and (3) comparative functional studies of haplotypes in highly heterozygous genomes. Beyond elucidating the evolutionary history of complex hybrids, this framework provides a solid data foundation for mining and utilising favourable alleles from multiple progenitors, thus carrying substantial practical value for basic research and breeding applications in multi-parental hybrid progeny.

A key functional insight emerges from our expression analysis. Despite the roughly equal numbers of biased genes between the two haplotypes (3,680 hapA-biased vs. 3,835 hapB-biased), Cre- and Cma-derived genes residing in hapA are expressed at significantly higher levels than their Cme-derived orthologs in hapB. This indicates that the two ancestral components of the sour-orange parent (Cre and Cma) are transcriptionally more active than the citron-derived genes, at least for the set of collinear, ancestry-paired gene pairs. Thus, the functional mosaicism of the lemon genome is not merely a matter of which genes are present, but also of how actively they are expressed. The systematically higher expression of mandarin/pummelo-derived alleles in hapA does not automatically imply agronomic superiority; it may reflect a regulatory ‘loudness’ that could be either beneficial (e.g., higher disease resistance protein production) or detrimental (e.g., metabolic waste). Nevertheless, this asymmetry provides a molecular handle for breeders to select haplotypes with desired expression levels for target

genes. This finding adds an important layer of regulatory complexity to our understanding of hybrid genomes.

A notable proportion of genes in both haplotypes (~ 37–39%) could not be confidently assigned to any of the three progenitors (Unsure category). Several non-mutually exclusive explanations may account for this. First, our strict assignment criteria (identity ≥ 95%, bidirectional best hit) may fail to capture genes that have diverged substantially due to relaxed purifying selection or positive selection, a possibility supported by the existence of > 100 positively selected genes in each haplotype. Second, the lemon genome may contain introgressed material from other, more distantly related Citrus species or even from close genera (e.g., *Fortunella* or *Poncirus*), which are not represented in our reference proteome set. Third, gene loss in one or more of the reference progenitor genomes could break the reciprocal best-hit relationship. The high UN fraction highlights the limitations of using only three reference species to fully capture the genetic complexity of a highly admixed hybrid like lemon.

The haplotype-resolved ancestry and expression asymmetry have direct implications for citrus breeding. First, the sour-orange-like haplotype (Cli_hapA) carries mandarin/pummelo alleles that are transcriptionally more active; breeders can intentionally increase the dosage of hapA for genes where higher expression correlates with desirable traits (e.g., disease resistance, flavor compounds). Second, the 7,515 significantly biased gene pairs serve as natural eQTLs – their allelic expression imbalance can be exploited to predict hybrid performance without transcriptome sequencing. Third, the positively selected, ancestry-biased genes listed in Supplementary Table 2 can be converted into competitive allele-specific PCR (KASP) markers for routine MAS. For instance, the PPR gene (NM_gene30493) located on chromosome 7B is a strong candidate for restoring fertility in cytoplasmic male sterile lines, a trait of high importance in F1 hybrid breeding. We envision that the framework presented here will accelerate the transition from descriptive genomics to haplotype-aware breeding in citrus.

5. Conclusion

In conclusion, this study provides the first haplotype-resolved phylogenomic dissection of the lemon genome, tracing each gene copy to three progenitor species – mandarin, pummelo, and citron. Our key findings are threefold. First, the sour-orange-derived haplotype exhibits significantly higher transcriptional activity for its Cre/Cma alleles than the citron-derived alleles in the other haplotype, revealing regulatory asymmetry. Second, ~ 37–39% of genes could not be assigned to any reference progenitor, indicating limitations of current reference-based approaches and possible contributions from more distant relatives. Third, we identified positively selected, expression-biased genes (e.g., transporters, transcription factors) that serve as candidates for agronomic traits.

Our analytical framework – integrating T2T genomes, ancestry assignment, selection analysis, and allele-specific expression – is transferable to other multi-parental hybrid systems. While computational, this work establishes a foundation for future functional validation and provides a breeder-ready toolkit: (i) a set of ancestry-diagnostic SNP markers to track mandarin/pummelo vs. citron introgression; (ii) a

priority list of 7,515 expression-biased gene pairs for eQTL-assisted selection; and (iii) 120–130 positively selected candidate genes (e.g., transporters, PPR, BiP) for targeted mutagenesis or MAS. This roadmap enables breeders to adopt a haplotype-aware selection strategy, accelerating the development of lemon and other citrus varieties with optimized allelic combinations.

Declarations

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Author Contribution

H.L and X.L. conceived the idea for the article. F.Z. and H.L.supervised this work. X.L. and Y.W. conducted the literature search and data analysis. X.L. and H.L. drafted the manuscript. Y.W. and F.Z. revised the manuscript. All authors read and approved the final manuscript.

Acknowledgement

We thank Dr. Gaofeng Jia and Dr. Yangrong Cao for their valuable contributions and discussions during this study. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability

The codes for re-annotations are available in the GitHub repository (https://github.com/HaominLyu/HiGenAA).

References

1. Alger EI, Edger PP. One subgenome to rule them all: underlying mechanisms of subgenome dominance. Curr Opin Plant Biol. 2020;54:108–13. https://doi.org/10.1016/j.pbi.2020.03.004.
2. Avni R, Kamal N, Bitz L, et al. A pangenome and pantranscriptome of hexaploid oat. Nature. 2026;649:131–9. https://doi.org/10.1038/s41586-025-09676-7.
3. Bairoch A, Apweiler R. The SWISS-PROT Protein Sequence Data Bank and Its New Supplement TREMBL. Nucleic Acids Res. 1996;24:21–5. https://doi.org/10.1093/nar/24.1.21.
4. Bao Y, Zeng Z, Yao W, et al. A gap-free and haplotype-resolved lemon genome provides insights into flavor synthesis and huanglongbing (HLB) tolerance. Hortic Res. 2023;10:uhad020.

https://doi.org/10.1093/hr/uhad020.

5. Bird KA, VanBuren R, Puzey JR, Edger PP. The causes and consequences of subgenome dominance in hybrids and recent polyploids. New Phytol. 2018;220:87–93. https://doi.org/10.1111/nph.15256.
6. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20. https://doi.org/10.1093/bioinformatics/btu170.
7. Brůna T, Hoff KJ, Lomsadze A, et al. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP + and AUGUSTUS supported by a protein database. NAR Genom Bioinform. 2021;3:lqaa108. https://doi.org/10.1093/nargab/lqaa108.
8. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60. https://doi.org/10.1038/nmeth.3176.
9. Cantalapiedra CP, Hernández-Plaza A, Letunic I, et al. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol. 2021;38:5825–9. https://doi.org/10.1093/molbev/msab293.
10. Cao Z-H, Song D, Hu Y, et al. An *S* -locus F-box protein as pollen *S* determinant targets non-self S-RNase underlying self-incompatibility in *Citrus*. J Exp Bot. 2024;75:3891–902. https://doi.org/10.1093/jxb/erae110.
11. Castle WE. Mendel's Law of Heredity. Science. 1903;18:396–406. https://doi.org/10.1126/science.18.456.396.
12. Chen B-J, Wang J-Y, Niu S-C, et al. Comparative genomics analyses reveal genomic variation and evolutionary adaptation in Dendrobium orchids. Nat Commun. 2025;17:63. https://doi.org/10.1038/s41467-025-66688-7.
13. Chen C, Chen H, Zhang Y, et al. TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. Mol Plant. 2020;13:1194–202. https://doi.org/10.1016/j.molp.2020.06.009.
14. Chen J, Qiu X, Sun Z, et al. Genome-wide analysis of UDP-glycosyltransferase family in *Citrus sinensis* and characterization of a UGT gene encoding flavonoid 1–2 rhamnosyltransferase. Int J Biol Macromol. 2024;280:135752. https://doi.org/10.1016/j.ijbiomac.2024.135752.
15. Chen ZJ. Genomic and epigenetic insights into the molecular bases of heterosis. Nat Rev Genet. 2013;14:471–82. https://doi.org/10.1038/nrg3503.
16. Cheng F, Wu J, Cai X, et al. Gene retention, fractionation and subgenome differences in polyploid plants. Nat Plants. 2018;4:258–68. https://doi.org/10.1038/s41477-018-0136-7.
17. Conant GC, Wolfe KH. Turning a hobby into a job: How duplicated genes find new functions. Nat Rev Genet. 2008;9:938–50. https://doi.org/10.1038/nrg2482.
18. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10:giab008. https://doi.org/10.1093/gigascience/giab008.
19. Debray K, Le Paslier M-C, Bérard A, et al. Unveiling the Patterns of Reticulated Evolutionary Processes with Phylogenomics: Hybridization and Polyploidy in the Genus Rosa. Syst Biol. 2022;71:547–69. https://doi.org/10.1093/sysbio/syab064.

20. Droc G, Giraud D, Belser C, et al. A Super-Pangenome for Cultivated Citrus Reveals Evolutive Features During the Allopatric Phase of Their Reticulate Evolution. Plant Biotechnol J. 2026;24:3345–67. https://doi.org/10.1111/pbi.70553.
21. Feng D, Liu S, Chen M, et al. Volatile content and genetic variation of citron in Tibet and Yunnan. Plant Physiol. 2025;197:kiae634. https://doi.org/10.1093/plphys/kiae634.
22. Freudenstein JV. Orchid phylogenetics and evolution: history, current status and prospects. Ann Bot. 2025;135:805–22. https://doi.org/10.1093/aob/mcae202.
23. Givnish TJ, Spalink D, Ames M et al. (2015) Orchid phylogenomics and multiple drivers of their extraordinary diversification. Proceedings of the Royal Society B: Biological Sciences 282:20151553. https://doi.org/10.1098/rspb.2015.1553
24. Goto S, Fujii H, Hamada H, et al. Allelic haplotype combinations at the MS-P1 region, including P-class pentatricopeptide repeat family genes, influence wide phenotypic variation in pollen grain number through a cytoplasmic male sterility model in citrus. Front Plant Sci. 2023;14. https://doi.org/10.3389/fpls.2023.1163358.
25. Guimarães RF, Camillo LR, Sousa AO, et al. Genome-wide identification and expression analysis of the molecular chaperone binding protein BiP genes in Citrus. Tree Genet Genomes. 2018;14:90–90.
26. Haas BJ, Salzberg SL, Zhu W, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008;9:R7. https://doi.org/10.1186/gb-2008-9-1-r7.
27. Hernández-Plaza A, Deng Z, Robledo-Yagüe F, et al. eggNOG v7: phylogeny-based orthology predictions and functional annotations. Nucleic Acids Res. 2026;54:D402–8. https://doi.org/10.1093/nar/gkaf1249.
28. Jiao Y, Wickett NJ, Ayyampalayam S, et al. Ancestral polyploidy in seed plants and angiosperms. Nature. 2011;473:97–100. https://doi.org/10.1038/nature09916.
29. Kamal N, Tsardakas Renhuldt N, Bentzer J, et al. The mosaic oat genome gives insights into a uniquely healthy cereal crop. Nature. 2022;606:113–9. https://doi.org/10.1038/s41586-022-04732-y.
30. Keilwagen J, Hartung F, Grau J. GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. In: Kollmar M, editor. Gene Prediction: Methods and Protocols. New York, NY: Springer New York; 2019. pp. 161–77.
31. Kim D, Paggi JM, Park C, et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15. https://doi.org/10.1038/s41587-019-0201-4.
32. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323. https://doi.org/10.1186/1471-2105-12-323.
33. Li F, Hou Z, Xu S, et al. Haplotype-resolved genomes of octoploid species in Phyllanthaceae family reveal a critical role for polyploidization and hybridization in speciation. Plant J. 2024;119:348–63. https://doi.org/10.1111/tpj.16767.
34. Li X, Fan L, Liu C, et al. Genome-Wide Identification and Biotic Stress Responses of TLP Gene Family in Citrus sinensis. Int J Mol Sci. 2025;26:10133. https://doi.org/10.3390/ijms262010133.

35. Li Y, Zhang S, Zhang B, et al. Single-cell and haplotype-resolved transcriptomics reveal molecular signatures of orchid floral structures. Mol Plant. 2026. https://doi.org/10.1016/j.molp.2026.03.016.
36. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30. https://doi.org/10.1093/bioinformatics/btt656.
37. Liu S, Bei X, Li Y, et al. CgABR1-CgFAD negatively regulates the fruit neck length in citrus. Plant Biotechnol J. 2025a;23:3083–93. https://doi.org/10.1111/pbi.70111.
38. Liu S, Li Y, Fang H, et al. Genome-wide identification and expression analysis of *MATE* gene family in citrus fruit (*Citrus clementina*). Genomics. 2022;114:110446. https://doi.org/10.1016/j.ygeno.2022.110446.
39. Liu S, Xu Y, Yang K, et al. Origin and de novo domestication of sweet orange. Nat Genet. 2025b;57:754–62. https://doi.org/10.1038/s41588-025-02122-4.
40. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 2005;33:6494–506. https://doi.org/10.1093/nar/gki937.
41. Marçais G, Delcher AL, Phillippy AM, et al. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol. 2018;14:e1005944. https://doi.org/10.1371/journal.pcbi.1005944.
42. Ou S, Su W, Liao Y, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20:275. https://doi.org/10.1186/s13059-019-1905-y.
43. Pérez-Escobar OA, Bogarín D, Przelomska NAS, et al. The origin and speciation of orchids. New Phytol. 2024;242:700–16. https://doi.org/10.1111/nph.19580.
44. Rao MJ, Zuo H, Xu Q. Genomic insights into citrus domestication and its important agronomic traits. Plant Commun. 2021;2:100138. https://doi.org/10.1016/j.xplc.2020.100138.
45. Stanke M, Keller O, Gunduz I, et al. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34:W435–9. https://doi.org/10.1093/nar/gkl200.
46. Tang H, Krishnakumar V, Zeng X, et al. JCVI: A versatile toolkit for comparative genomics analysis. iMeta. 2024;3:e211. https://doi.org/10.1002/imt2.211.
47. Van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 2009;10:725–32. https://doi.org/10.1038/nrg2600.
48. Wang J, Zhou Y, Zhang M, et al. Resolving floral development dynamics using genome and single-cell temporal transcriptome of Dendrobium devonianum. Plant Biotechnol J. 2025;23:2997–3011. https://doi.org/10.1111/pbi.70094.
49. Wang L, He F, Huang Y, et al. Genome of Wild Mandarin and Domestication History of Mandarin. Mol Plant. 2018;11:1024–37. https://doi.org/10.1016/j.molp.2018.06.001.
50. Wu B, Yu Q, Deng Z et al. (2023) A chromosome-level phased genome enabling allele-level studies in sweet orange: a case study on citrus Huanglongbing tolerance. Horticulture Research 10.

51. Wu GA, Prochnik S, Jenkins J, et al. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nat Biotechnol. 2014;32:656–62. https://doi.org/10.1038/nbt.2906.
52. Wu GA, Terol J, Ibanez V, et al. Genomics of the origin and evolution of Citrus. Nature. 2018;554:311–6. https://doi.org/10.1038/nature25447.
53. Yang Z, Nielsen R. Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. Mol Biol Evol. 2000;17:32–43. https://doi.org/10.1093/oxfordjournals.molbev.a026236.
54. Zhang M, Zhao Y, Nan T, et al. Genome-wide analysis of *Citrus medica* ABC transporters reveals the regulation of fruit development by CmABCB19 and CmABCC10. Plant Physiol Biochem. 2024a;215:109027. https://doi.org/10.1016/j.plaphy.2024.109027.
55. Zhang X, Lan L, Yang Y, et al. Pangenomic analyses of rose uncover widespread structure variation and empower genomics-directed breeding. Nat Genet. 2026;58:1164–75. https://doi.org/10.1038/s41588-026-02569-z.
56. Zhang Z, Yang T, Liu Y, et al. Haplotype-resolved genome assembly and resequencing provide insights into the origin and breeding of modern rose. Nat Plants. 2024b;10:1659–71. https://doi.org/10.1038/s41477-024-01820-x.
57. Zhu C, You C, Wu P, et al. The gap-free genome and multi-omics analysis of Citrus reticulata 'Chachi' reveal the dynamics of fruit flavonoid biosynthesis. Hortic Res. 2024;11:uhae177. https://doi.org/10.1093/hr/uhae177.

Figures

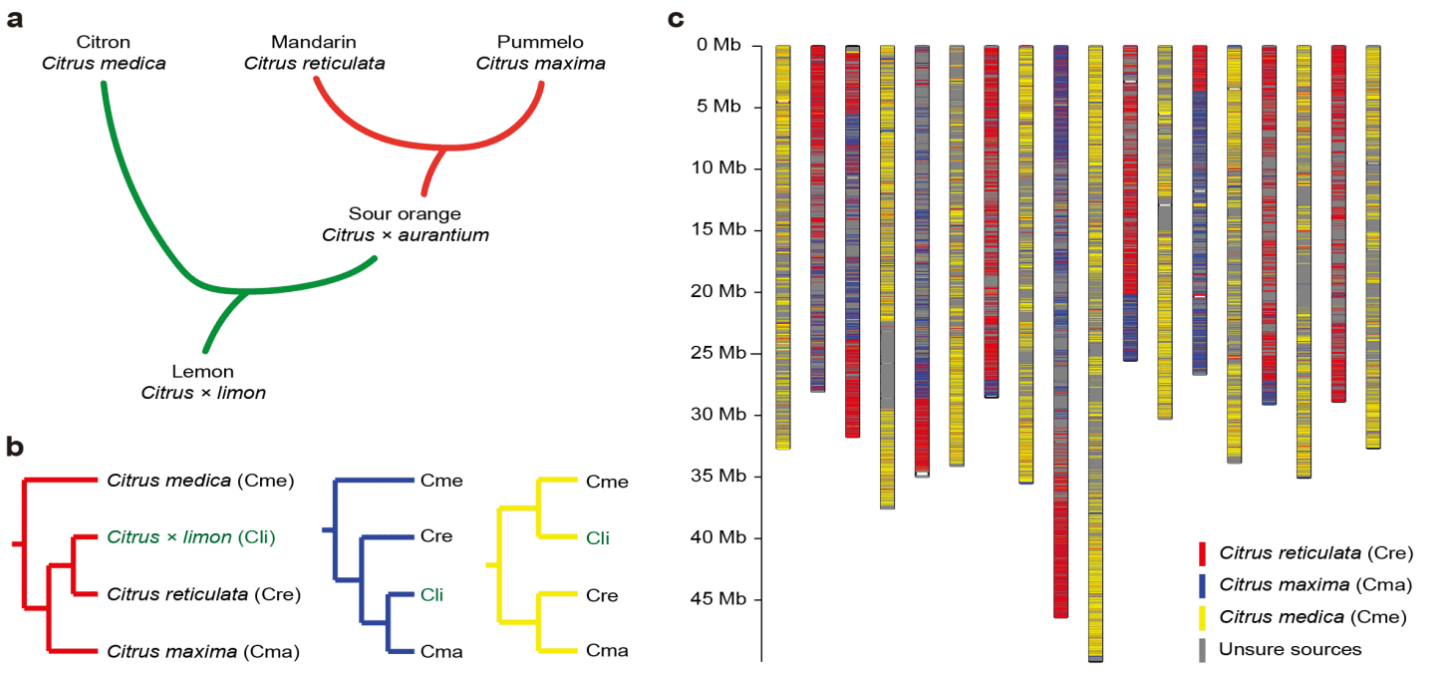

Figure 1

Haplotype-resolved ancestral origin of the lemon genome. (a) Schematic illustration of the proposed hybrid origin of lemon (*Citrus × limon*). The three progenitor species are indicated: mandarin (*C. reticulata*, Cre), pummelo (*C. maxima*, Cma), and citron (*C. medica*, Cme). (b) Workflow of the phylogenomics-based ancestry assignment method. Genes are classified as Cre-derived (red), Cma-derived (blue), Cme-derived (yellow), or Unsure (UN). (c) Chromosome-scale distribution of ancestral origins along a representative lemon chromosome. Each coloured bar or segment indicates the predominant ancestry of genes with the colours: red (Cre), blue (Cma), yellow (Cme), grey (unsure source).

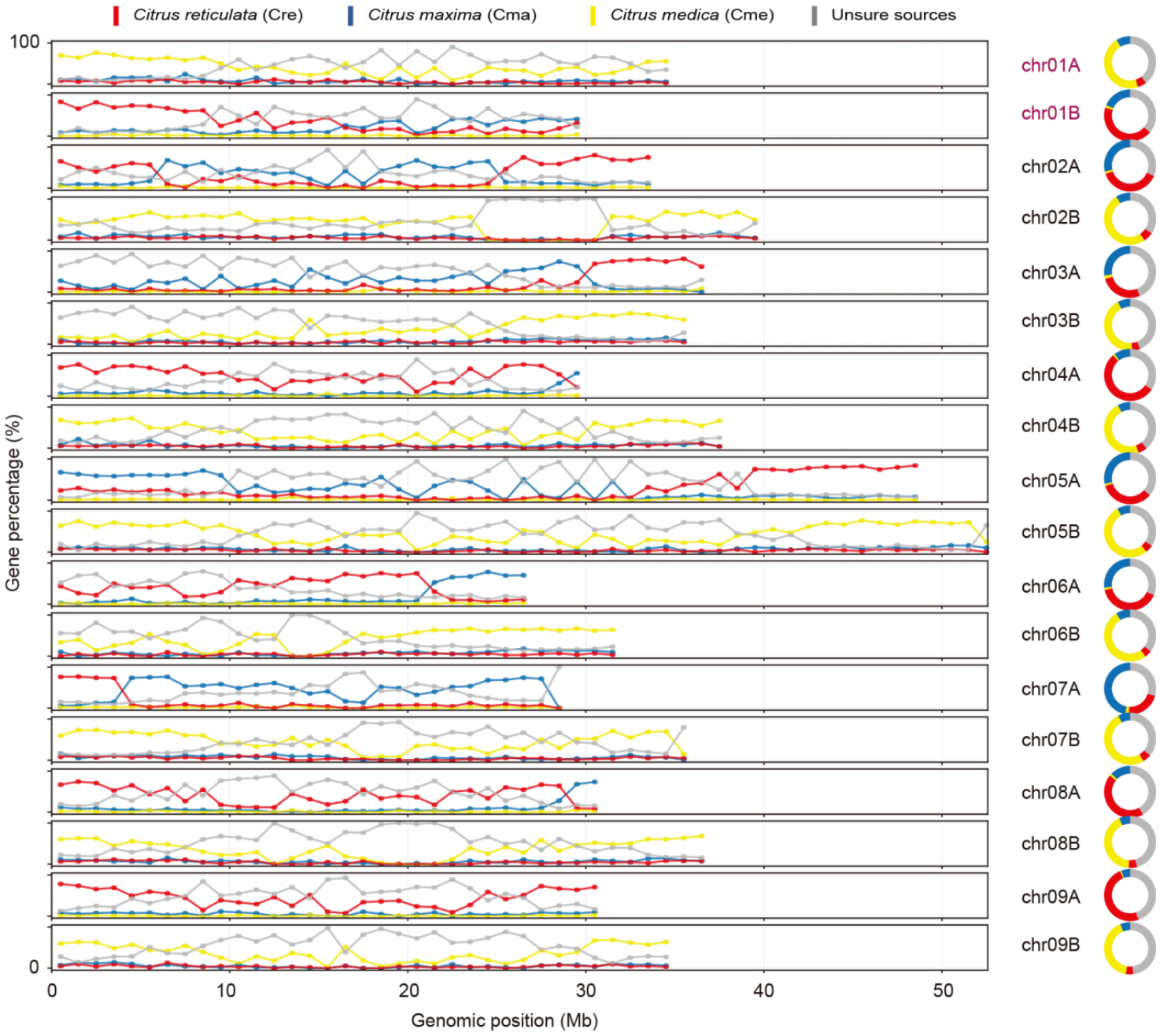

Figure 2

Ancestral origin composition along lemon chromosomes. For each lemon haplotype (Cli_hapA and Cli_hapB, chromosomes 01–09), sliding-window analysis (1 Mb window) was used to calculate the proportion of genes derived from each of the three progenitor species: mandarin (*Citrus reticulata*, Cre; red), pummelo (*C. maxima*, Cma; blue), and citron (*C. medica*, Cme; yellow). The chr01A and chr01B have been revised and assigned to the right haplotypes. Genes with no confident assignment are shown as "Unsure sources" (grey). The left panel displays, for each chromosome, the ancestry proportions as stacked bars along the physical position (x-axis, in Mb). The right-hand circular plots summarise, per chromosome, the overall percentage of each ancestry category, providing a global view of the mosaic ancestry distribution across the lemon genome.

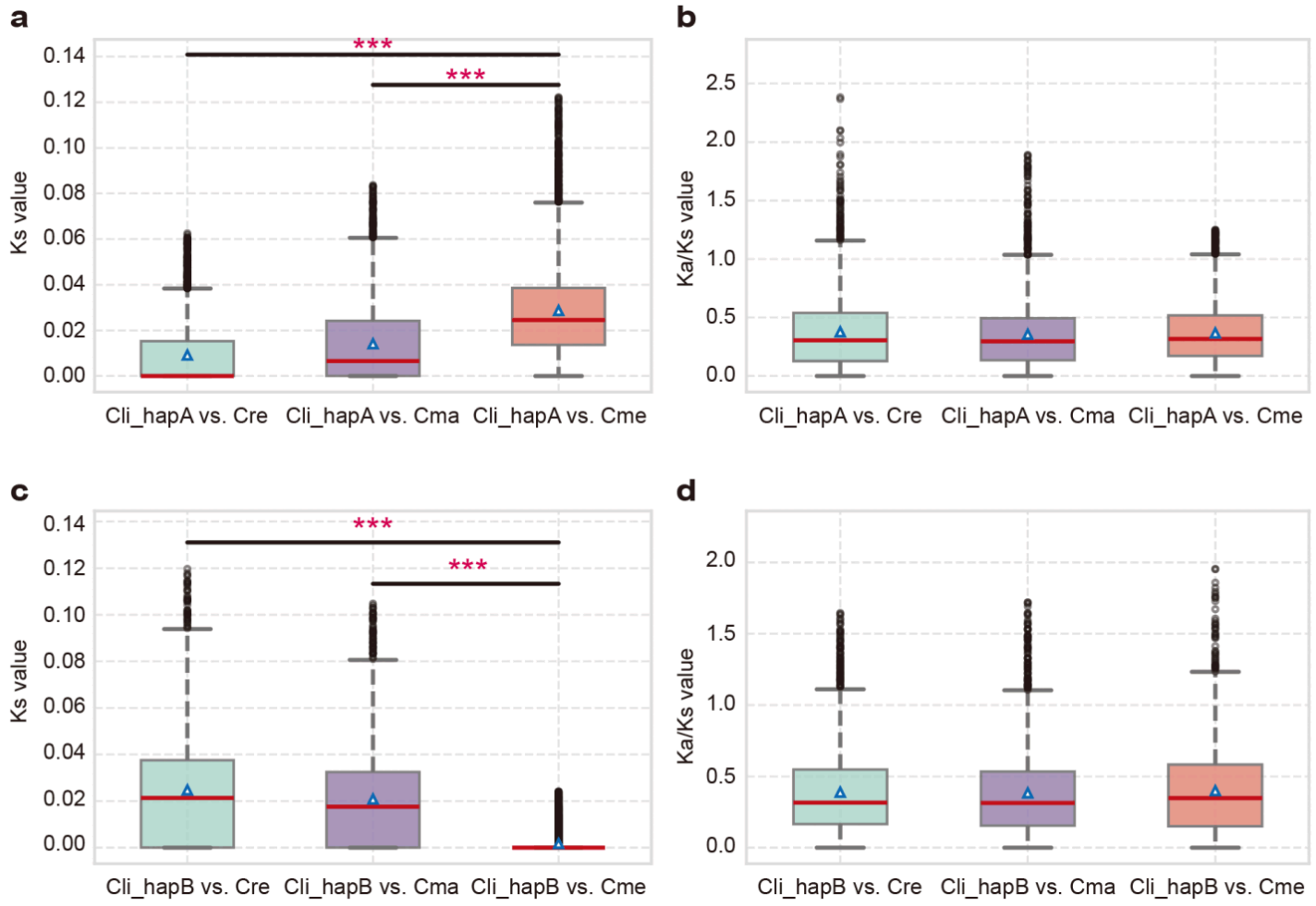

Figure 3

Sequence divergence and selection pressure between lemon haplotypes and progenitor genomes. (a) Distribution of synonymous substitution rates (Ks) for orthologous gene pairs between Cli_hapA and each of the three progenitor species: mandarin (Cre), pummelo (Cma), and citron (Cme). (b) Distribution of Ka/Ks ratios for the same comparisons as in (a). (c) Distribution of Ks values for orthologous gene pairs between Cli_hapB and Cre, Cma, and Cme. (d) Distribution of Ka/Ks ratios for the same comparisons as in (c).

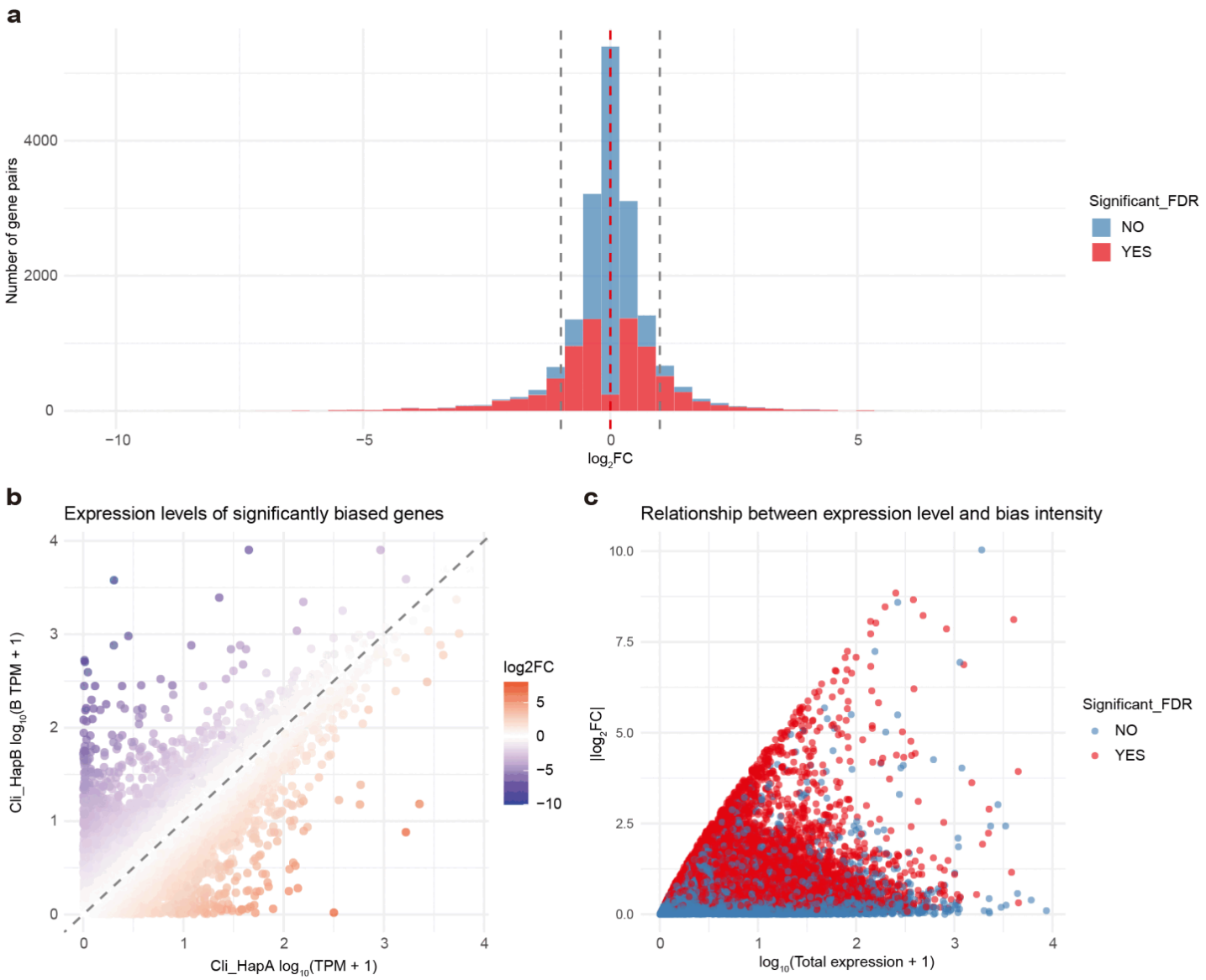

Figure 4

Identification and expression analysis of allele-specific biased genes in lemon haplotypes. (a) Distribution of log_2(fold change) (Cli_hapA vs. Cli_hapB) for all expressed gene pairs. Red bars represent significantly biased genes (FDR < 0.05, binomial exact test), while blue bars indicate non-significant genes. (b) Scatter plot comparing expression levels of significantly biased genes between Cli_hapA (x-axis) and Cli_hapB (y-axis). Each dot represents a gene pair. Points above the diagonal indicate hapA-biased expression; points below indicate hapB-biased expression. (c) Relationship between expression level (average TPM of Cli_hapA and Cli_hapB, x-axis) and bias intensity ($|log_2FC|$, y-axis). Red dots denote significantly biased genes; blue dots denote non-significant genes. This panel illustrates that bias intensity is largely independent of overall expression magnitude.

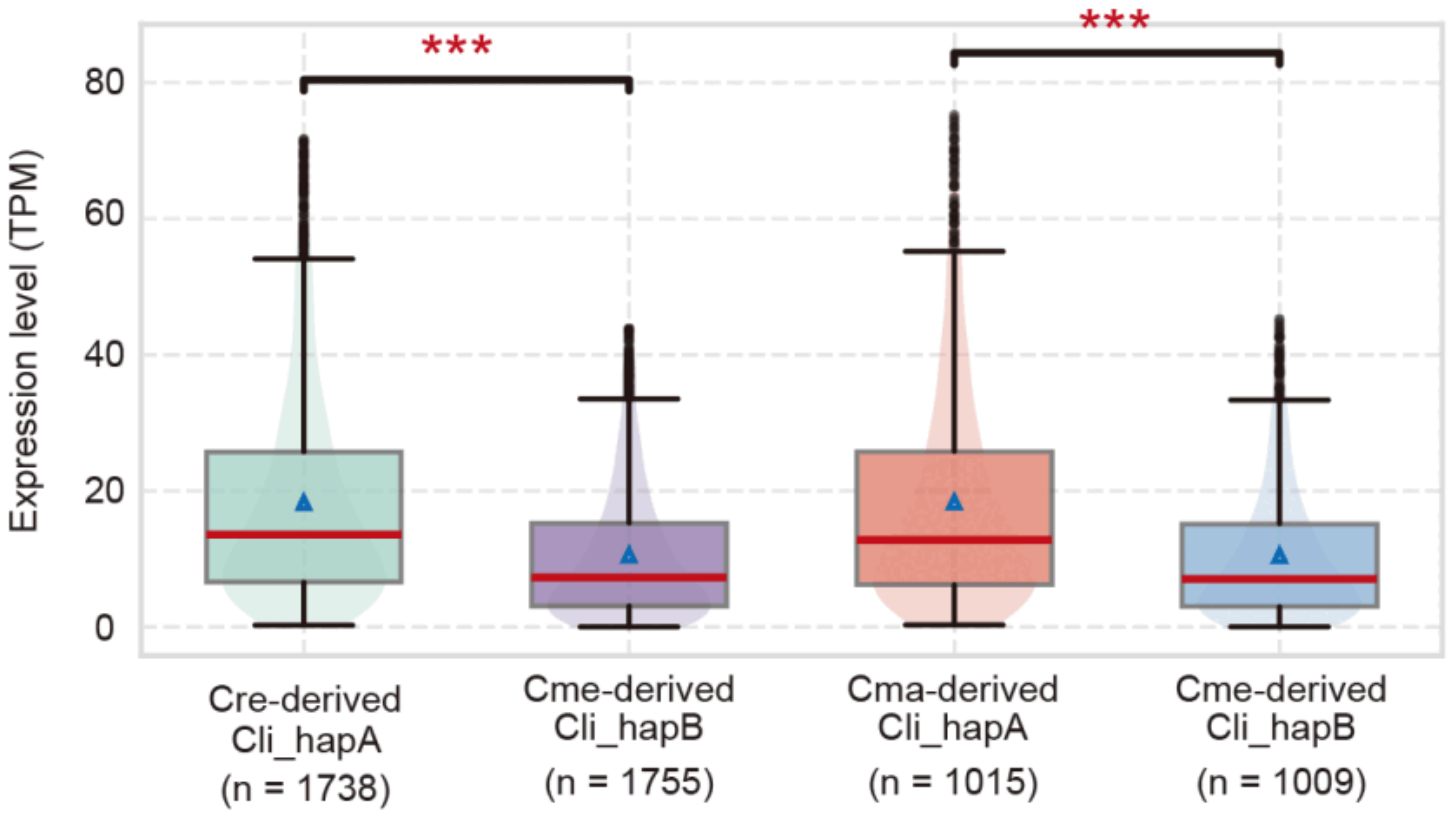

Figure 5

Ancestry-specific expression comparison between collinear gene pairs of lemon haplotypes. Only collinear gene pairs between Cli_hapA and Cli_hapB are considered. Expression levels (TPM) are compared between Cli_hapA genes of defined ancestry and their corresponding Cli_hapB orthologs derived from citron (Cme). Left pair: Cre-derived genes in Cli_hapA (n = 1,738) versus their Cme-derived orthologs in Cli_hapB (n = 1,755). Right pair: Cma-derived genes in Cli_hapA (n = 1,015) versus their Cme-derived orthologs in Cli_hapB (n = 1,009). *** indicates $P < 0.001$ using Wilcoxon signed-rank test. Bars represent median TPM; error bars indicate standard error of the mean.

Supplementary Files

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

- SupplTables.xlsx