RESEARCH

Open Access

The haplotype-resolved genome assembly of an ancient citrus variety provides insights into the domestication history and fruit trait formation of loose-skin mandarins

Minqiang Yin[1†], Xiaochan Song[1†], Chao He[1], Xiyuan Li[1], Mengyuan Li[1], Jiangbo Li[2], Hao Wu[3], Chuanwu Chen[4], Li Zhang[1], Zhenmei Cai[1], Liqing Lu[1], Yanhui Xu[1], Xin Wang[1], Hualin Yi[1*] and Juxun Wu[1*]

†Minqiang Yin and Xiaochan Song are co-first authors.

*Correspondence:
yihualin@mail.hzau.edu.cn;
wjxun@mail.hzau.edu.cn

[1] National Key Laboratory for Germplasm Innovation & Utilization of Horticultural Crops, Huazhong Agricultural University, Wuhan 430070, China
[2] Fuzhou Agricultural and Rural Industry Development Service Center, Fuzhou 344100, China
[3] Fuzhou Institute of Agricultural Sciences, Fuzhou 344100, China
[4] Guangxi Key Laboratory of Germplasm Innovation and Utilization of Specialty Commercial Crops in North Guangxi, Guangxi Academy of Specialty Crops, Guilin 541004, China

Abstract

Background: Loose-skin mandarins (LSMs) are among the oldest domesticated horticultural crops, yet their domestication history and the genetic basis underlying the formation of key selected traits remain unclear.

Results: We provide a chromosome-scale and haplotype-resolved assembly for the ancient Chinese citrus variety Nanfengmiju tangerine. Through the integration of 77 resequenced and 114 published citrus germplasm genomes, we categorize LSMs into 12 distinct groups based on population genomic analyses. We infer that the ancestors of modern cultivated mandarins diverged from wild mandarins in Daoxian approximately 500,000 years ago, when they entered the Yangtze and Pearl River Basins. There, they were domesticated into four ancient cultivation groups, forming the cornerstone of modern Chinese LSM cultivation. We identify selective sweeps underlying quantitative trait loci and genes related to important fruit quality traits, including sweetness and size. We reveal that the co-selection of sugar transporter and metabolism genes are associated with increased fruit sweetness. Significant alterations in the auxin and gibberellin signaling networks may contribute to the enlargement of LSM fruits. We also provide a comprehensive, high-spatiotemporal-resolution atlas of allelic gene expression during citrus fruit development. We detect 5890 allele pairs showing specific expression patterns and a significant increase in variation levels.

Conclusions: Our study provides valuable genomic resources and further revises the origin and domestication history of LSMs, offering insights for genetic improvement of citrus plants.

Keywords: *Citrus reticulata*, Genome, Haplotypes, Evolution, Domestication and improvement, Fruit development

Background

Loose-skin mandarin (LSM) is the general term for tangerine and mandarin (*Citrus reticulata*), which are favored by consumers for their ease of peeling. In 2022, China's citrus production exceeded 60 million tons, with loose-skin mandarins (LSMs) making up 71% of the total production (Data source: National Bureau of Statistics and China Agriculture Research System-Citrus). As a result, LSMs hold significant importance in China. The domestication and improvement of LSM have driven remarkable changes in key traits such as sugar-acid balance and fruit size [1–3]. However, the unclear domestication history of LSMs have hindered the identification of potential selection signals.

The complex genetic background, disordered classification, and lack of fossil resources of LSMs have led to long-standing controversy regarding their origin and domestication history. Recent studies, based on extensive citrus genomic data, have inferred that LSMs originated is the Nanling Mountains in southern China, with the Mangshan wild mandarin (*Citrus reticulata* "Mangshan") possibly being the oldest ancestor [1, 3]. Wild and semi-domesticated LSMs (Daoxian wild mandarin and Huapiju mandarin) have been found in Daoxian, an ancient human settlement within the Nanling [4]. Wang et al. suggested that Daoxian may be the cradle of LSM domestication, speculating that the Daoxian wild mandarin underwent two independent domestication events to form the modern cultivated mandarin [1]. Additionally, Wu et al. proposed that *Citrus ryukyuensis* found in the Ryukyu Islands, the mainland Asian Mangshan wild mandarin, and the common mandarin are the sources of mandarin diversity. They believe that the Daoxian wild mandarin is a hybrid of the common mandarin and the Mangshan wild mandarin [5]. Overall, the specific domestication processes and the connections among modern LSMs remain unclear. Fortunately, over 4000 years of citrus cultivation history in China has led to the preservation of many ancient cultivated citrus varieties, including Hongju (红橘, red tangerine), Nanju (南橘), Jiangan (建柑), Nianju (年橘), and Ruju tangerine (乳橘, also named Zhengan in "Chu Lu") [2], which will provide clues for further deciphering the LSM domestication history.

In the domestication and improvement of horticultural crops, traits beneficial to humans are typically selected, including larger edible organs, enhanced aroma, and increased sweetness [6]. Heightened sweetness not only augments the fruit's palatability but also boosts its energy value, serving as a target of domestication in many horticultural crops. For example, cultivated varieties of watermelon, peach, and kiwifruit exhibit significantly greater sweetness than wild species [7–9]. In citrus, sweetness is determined by the concentrations of sucrose, glucose, and fructose in the fruit, but their regulation is poorly understood. We have noted a substantial variability in sugar content among different LSMs, with some cultivated varieties displaying significantly higher sugar levels than the wild types, highlighting potential selective events that have been overlooked [1]. Bigger fruit size is another typical trait targeted curing crop domestication and improvement [10]. The extensive introgression from *Malus sylvestris* during apple domestication has contributed to increased fruit size [11]. In the domestication and improvement of citrus, a positive correlation exists between the proportion of pummelo introgression and fruit size [12], but the exact reasons for this correlation remain unknown.

To enhance the understanding of the domestication and modern breeding of LSMs, we present a chromosome-scale and haplotype-resolved genome assembly of Nanfengmiju tangerine as a reference. Nanfengmiju tangerine, has been cultivated in China for more than 1300 years [2], is particularly prized by consumers for its seedlessness, high sugar content, low acid content, and unique flavor. Compared to Clementine tangerine, Nanfengmiju has a purer genetic background with almost no pummelo introgression [1, 5]. Furthermore, it exhibits a lower degree of genetic divergence from other LSMs than Mangshan wild mandarin [1], making it more suitable for exploring genetic variations in LSMs. Subsequently, we resequenced the genomes of 77 broadly cultivated LSM germplasms (including two ancient cultivated varieties, Nanfengmiju tangerine and Shatangju tangerine, as well as key parental cultivars used in current citrus hybrid breeding, such as Satsuma mandarin and Ponkan tangerine) in southern China and analyzed them with previously sequenced wild and domesticated LSMs. Our study clarified classification and phylogenetic relationships of LSMs and further revises their domestication and improvement history. Additionally, we identified several potential domestication and improvement selection signals that could accelerate the genetic enhancement of important citrus traits. Finally, by employing the haplotype-resolved assembled genome of Nanfengmiju and transcriptome data of Guiyuehong tangerine (GYH, a cultivar of Nanfengmiju) from three fruit tissues across five developmental stages, we constructed an allelic expression atlas of citrus fruit development.

Results

Genome sequencing, assembly, and annotation for Nanfengmiju tangerine

For generating high-quality genome assemblies, the genomic sequences of 120-year-old Nanfengmiju tangerine (NFMJ-120y) were sequenced using a combination of sequencing platforms (see “Methods”). Before de novo assembly, we first produced ~ 60 × Illumina paired-end short reads (150 bp) to investigate the overall genome characteristics of NFMJ-120y. The genome size was estimated to be 298.9 Mb, with a heterozygosity ratio of 0.935% according to the K-mer distribution assessment ($K = 17$) (Additional file 2: Table S1). A total of 17.6 Gb HiFi reads (~ 60 × , subread N50 = 16,227) and 30.6 Gb Hi-C reads (~ 100 ×) were combined to generate a preliminary contig-level assembly by Hifiasm, resulting in one monoploid (Mono) and two phased haplotype assemblies (Hap1 and Hap2) of NFMJ-120y to represent the Nanfengmiju genome. The contig N50 length of all initial assemblies was greater than 28.1 Mb (Table 1), which was greatly improved compared with that of the previously published citrus genomes [13, 14]. Then, the resulting three contig-level assemblies were corrected and linked into 9 pseudo-chromosomes that anchored 299.73, 294.98, and 297.67 Mb of preliminary assemblies using the 3D-DNA chromosome construction pipeline, which was very close to the estimated genome size (Table 1 and Additional file 1: Fig. S1). Finally, 98.4%, 98.1%, and 98.4% of the complete BUSCO genes were found in Mono, Hap1, and Hap2, respectively (Table 1 and Additional file 2: Table S2), indicating high completeness of the three genomic gene regions. Furthermore, the LAIs of the all three genomes exceeded 20 (Table 1), indicating the high continuity and completeness of intergenic and repetitive sequence assembly.

In Mono, a total of 201.8 Mb of repetitive sequences were found, accounting for 57.16% of the assembly, of which transposable elements (TEs) accounted for 54.6%.

Table 1 Statistics for the genome assembly of Nanfengmiju tangerine

	Monoploid	Haplotype 1	Haplotype 2
Size of the assembled scaffold (Mb)	352.86	345.02	336.00
Chromosome size (Mb)	299.73	294.98	297.67
Percent of estimated genome size (%)	100.3	98.7	99.6
Number of contigs	243	324	180
Largest contig (Mb)	51.03	33.46	49.63
contig N50 (Mb)	31.83	28.11	30.30
contig N90 (Mb)	11.01	6.15	2.70
Largest scaffold (Mb)	53.33	50.60	50.68
scaffold N50 (Mb)	32.11	30.34	31.70
scaffold N90 (Mb)	29.07	29.11	29.06
GC (%)	38.2	37.8	37.3
Number of protein coding genes	29,872	29,845	30,482
Mean gene length (bp)	3211.5	3246.5	3194.7
Mean coding sequence length (bp)	1148.0	1148.1	1136.4
Mean exon per gene	6.7	6.8	6.5
Mean exon length (bp)	293	291	285
Percentage of TEs (%)	50.6	48.9	48.3
Number of genes with annotated alleles	-	21,031	
BUSCO completeness of assembly (%)	98.4	98.1	98.4
LTR Assembly Index (LAI)	21.5	20.4	20.7

However, fewer repetitive sequences and TEs were detected in Hap1 and Hap2 (Table 1, Fig. 1a and Additional file 2: Table S3). Moreover, differences were observed in the types, number, and distribution of TEs in the two haplotype assemblies. For example, 132,976 LTR sequences were annotated and classified into 8 types in Hap2, whereas in Hap1, only 103,904 LTR sequences were annotated into 6 types (Additional file 2: Table S3). This discrepancy may be attributed to various factors, such as homologous chromosome recombination, TE insertion preference, and postintegration selection processes [15], which could lead to differences in the gene structure and composition of the cis-regulatory elements of alleles between two haplotypes, resulting in divergent transcriptional regulation levels [16].

In total, 29,872 putative protein-coding gene models were predicted in Mono by integrating ab initio gene prediction, homology-based prediction, and transcript evidence from the RNA-seq dataset. Similarly, 29,845 and 30,482 putative protein-coding genes were annotated in Hap1 and Hap2, respectively (Table 1).

Genomic variations between the two haplotypes

Genomic variation leads to the generation of new genes and the diversification of gene function, which plays a pivotal role in speciation and adaptive divergence. To identify genomic variations, we initially aligned the assembly of Hap2 to that of Hap1 using MUMmer. Genomic regions of 282.3 and 284.6 Mb could be successfully aligned between Hap1 and Hap2, respectively. Among these, 22,618 were one-to-one alignments with cumulative lengths of 260.3 Mb (Additional file 2: Table S4). Next, we identified 13,404 and 14,154 chromosomal structural variations in Hap1 and Hap2, respectively by SyRI, including 130 inversions, 4248 translocations, and more than 9000 duplications

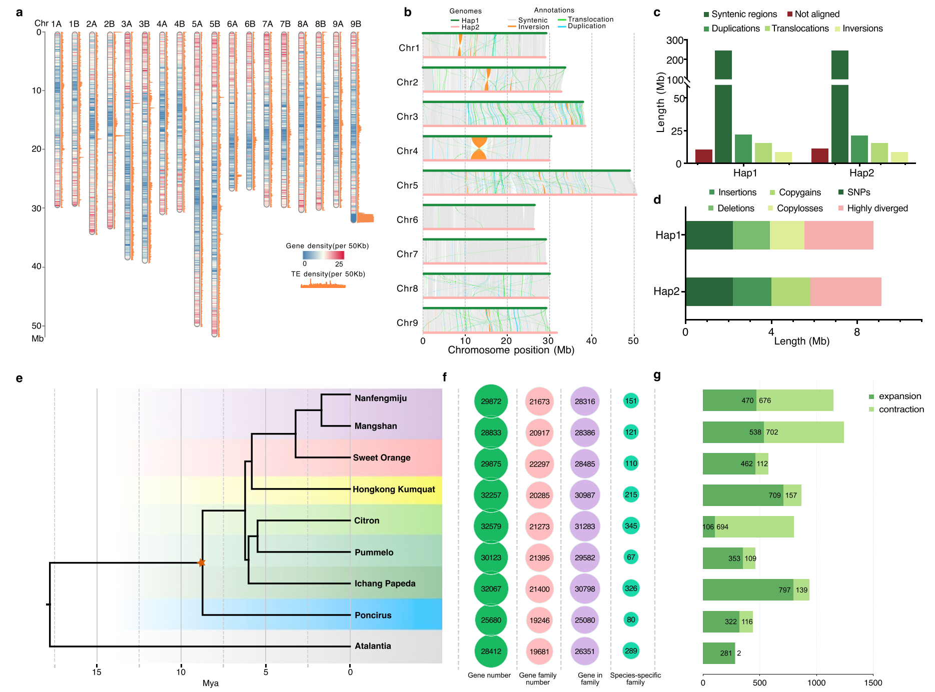

Fig. 1 Haplotype diversity and comparative genomic analysis of Nanfengmiju tangerine. **a** Distribution of genomic elements of Haplotype1 (Hap1) and Haplotype2 (Hap2) in Nanfengmiju tangerine. The chromosomes of Hap1 and Hap 2 were numbered as A and B, respectively. The densities of genes and transposable elements (TEs) were calculated per 50 kb. **b** The collinearity analysis of Hap1 and Hap2. Hap1 as a reference genome. **c** Lengths of syntenic and unaligned regions in Hap1 and Hap2 genomes. **d** Statistical of variation between Hap1 and Hap2. **e** Phylogenetic relationships and divergence times between Nanfengmiju tangerine and citrus-related species. Clementina: Citrus clementina, Mangshan: Citrus reticulata cv. Mangshan, Sweet Orange: Citrus sinensis, Hongkong Kumquat: Fortunella hindsii, Citron: Citrus medica , Pummelo: Citrus maxima; Ichang Papeda: Citrus ichangensis, Poncirus: Poncirus trifoliata, Atalantia: Atalantia buxifolia. Atalantia as outgroup. The asterisk represents the correction of time point using the speciation time of Poncirus trifoliata and Citrus. **f** Number of gene families identified by OrthoFinder. **g** Number of expansion and contraction gene families ($P < 0.05$) identified by CAFE5

(9026 in Hap1 and 9776 in Hap2), but rare in in Chr6 and Chr7 (Fig. 1b, c and Additional file 2: Table S5). Notably, there was a large 3.8 Mb inversion on chromosome 4 (Chr4A:11,493,167–15,301,309, Chr4B:11,320,326–15,048,651), which was supported by Hi-C data (Fig. 1b and Additional file 1: Fig. S1d). Additionally, an abundance of inversions existed on all chromosomes, ranging in size from 100 bp to 987 kb. These inversions may be an important reason for adaptive divergence of Nanfengmiju.

Furthermore, we detected 2,203,686 SNPs, 377,400 InDel (~ 1–50 bp), and 1942 SVs (> 50 bp) between the two haplotypes (Fig. 1d and Additional file 2: Table S6). However, only approximately 2.23% SNPs, 0.87% InDels, and 2.78% SVs might have induced deleterious variations (Additional file 2: Table S6). Overall, the intragenomic diversity was estimated to be ~ 1.88% (Additional file 2: Table S5), a level lower than the diversity of heterozygous diploid potato [17].

Comparative genomic analyses among Citrinae species

To infer the phylogenetic relationship between Nanfengmiju and other Citrinae species, we compared the Nanfengmiju Mono assembly with the genomes of 9 representative

Citrinae species (Additional file 2: Table S7) and constructed a high-confidence phylogenetic tree using the obtained 5789 single-copy orthologs (Fig. 1e and Additional file 2: Table S8). Our results showed that LSMs (Nanfengmiju and Mangshan) diverged from other Citrinae species approximately 3.22 (1.42–4.71) million years ago (MYA). Nanfengmiju was estimated to have diverged from Mangshan approximately 1.69 (0.73–2.49) MYA (Fig. 1e).

We further performed gene family clustering analysis on 9 species. A total of 28,316 genes in Nanfengmiju clustered into 21,673 gene families, 13,824 of which were shared by all 9 species, while only 151 gene families (462 genes) were specific to Nanfengmiju (Fig. 1f and Additional file 2: Table S8). Analyses of gene family expansion and contraction revealed 470 (2280 genes) and 676 (208 genes) gene families that exhibited significant ($P < 0.05$) expansion and contraction in Nanfengmiju, respectively (Fig. 1g). KEGG pathway analysis revealed that the contracting gene families were significantly enriched in the ubiquinone and other terpenoid – quinone biosynthesis and pentose and glucuronate interconversions (Additional file 1: Fig. S2a). In addition, the expanding gene families were mainly enriched in phenylpropanoid biosynthesis, photosynthesis, and carbon fixation (Additional file 1: Fig. S2b).

Population structure and phylogeny of 191 citrus accessions

To gain further insight into the genetic history of LSMs origin and domestication, we resequenced 77 accessions of the most widely cultivated LSMs (including Nanfengmiju tangerine, Shatangju tangerine, Satsuma mandarin, and Ponkan tangerine) in China. Along with 114 published resequenced citrus samples, a total of 191 citrus accessions were collected, including 122 *Citrus reticulata* (LSM), 24 *Citrus sinensis* (SWO), 20 *Citrus maxima* (PU), 11 *Citrus ichangensis* (IC), 8 *Citrus medica* (CI), and 6 *Citrus aurantium* (SOO) samples (Additional file 2: Table S9). A total of 14.2 billion clean reads were used for analysis, and 91.5% of the paired reads on average were properly mapped onto the Mono assembly. We identified 47.93 million variations across accessions, among which 6.76 million high-quality biallelic SNPs were further analyzed. Nucleotide diversity (π) analysis indicated that the genetic diversity of LSM was greater than that of IC, CI, and PU but lower than that of SWO and SOO (Additional file 1: Fig. S3).

To assess the population genetic structure of the 191 citrus accessions, we performed principal component analysis (PCA) and inferred a phylogenetic tree based on the 192,855 LD pruning SNP dataset. These analyses accurately separated LSM, IC, CI, and PU and demonstrated the high intraspecific diversity of LSM (Fig. 2a and b). Surprisingly, we previously divided 122 LSMs into five classified groups (wild mandarin, Nanfengmiju tangerine, Ponkan tangerine, Shatangju tangerine, and Satsuma mandarin) and one unclassified group, but there were obvious separations within some of the classified groups. For example, Nanfengmiju tangerine was divided into two groups, similar to Shatangju tangerine (Fig. 2a and b), indicating that the high and complex intraspecific diversity of LSMs. Nonetheless, the deep divergence within groups has been overlooked in most previous studies [1, 5, 12], potentially resulting in potential domestication events being overlooked.

To further investigate subdivision and classification within LSMs, we utilized ADMIXTURE to estimate the ancestry composition of LSMs. Similarly, ADMIXTURE

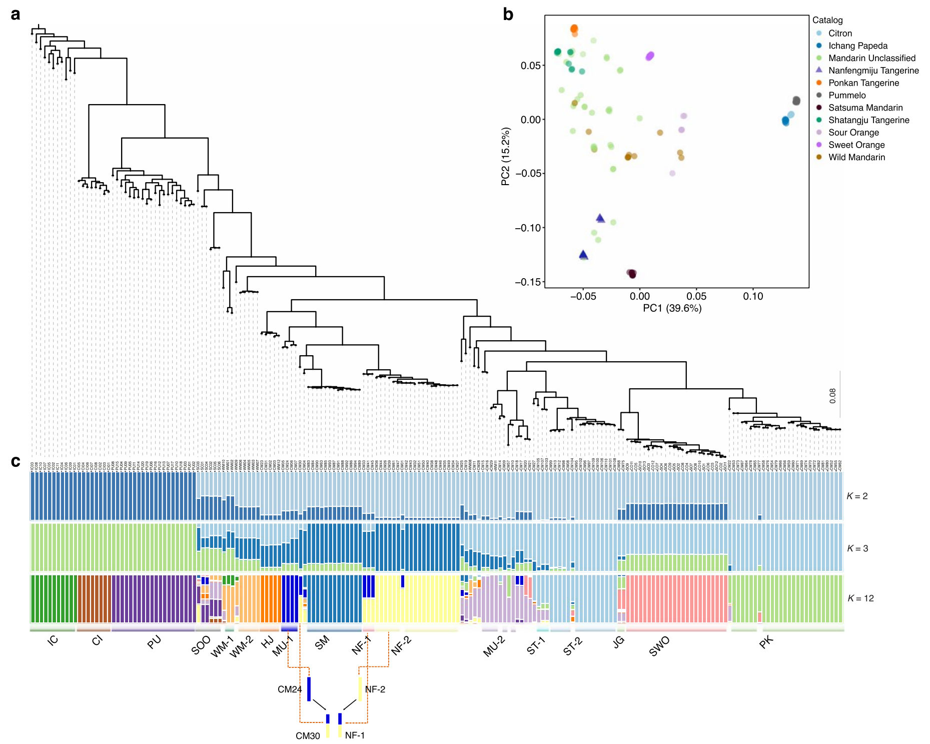

Fig. 2 Population genetic structure and phylogenetic relationship of 191 citrus accessions. **a** A maximum-likelihood phylogenetic tree was inferred by LD pruning SNPs. **b** Principal-component analyses (PCA) by Plink. **c** ADMIXTURE clustering of the accessions and sorting by phylogenetic tree. IC: Ichang papeda, CI: Citron, PU: Pummelo, SOO: Sour orange, WM-1: Mangshan wild mandarin. WM-2: Represented by Daoxian wild mandarin. HJ: Represented by Hongju tangerine. MU-1: Mandarin Unclassified group 1, represented by Jiangan tangerine. SM: Satsuma Mandarin. NF-1: Nanfengmiju tangerine group 1. NF-2: Nanfengmiju tangerine group 2. MU-2: Mandarin Unclassified group 2, represented by Nianju tangerine. ST-1: Shatangju tangerine group 1. ST-2: Shatangju tangerine group 2. JG: Jiaogan (*Citrus reticulata* ' Tankan'). SWO: Sweet orange. PK: Ponkan. Two similar mixed events are indicated below the Admixture plot (CM24, Huoju; CM30, Xiaoyeguangju). Two inferred similar hybridization events are indicated below the Admixture plot (CM24, Huoju; CM30, Xiaoyeguangju)

at $K=2$ and 3 could segregate LSMs from IC, CI, and PU and divide LSMs into two distinct groups (Fig. 2c). From $K=4$ to 12, new subpopulations arose from LSMs, and PCA confirmed the existence of such subpopulations (Fig. 2b, c and Additional file 1: Fig. S4). At $K=12$, hierarchical clustering of ancestry components revealed 12 LSM groups of 102 LSM accessions. Within these categories, wild mandarins were delineated into WM-1 (represented by the Mangshan wild mandarin) and WM-2 (represented by the Daoxian wild mandarin No.2) groups, Nanfengmiju was categorized as NF-1 and NF-2, and Shatangju was segmented into ST-1 and ST-2. Seventeen of the 32 unclassified mandarins were grouped into HJ, MU-1, NF-2, MU-2, ST-2, and JG (Fig. 2c, Fig. S4c and Additional file 2: Table S9). A high degree of intersubgroup population divergence (average Fst ~ 0.32) was maintained, while the intrasubgroup kinship coefficient (average kinship coefficient ~ 0.35) was high (Fig. 3a and Additional file 2: Table S10).

HJ, MU-1, and MU-2, represented by Hongju, Jiangan, and Nianju, respectively, are archaic cultivated varieties indigenous to China [2]. Originating from an early divergence from wild mandarins, these cultivars suggest a primordial domestication process

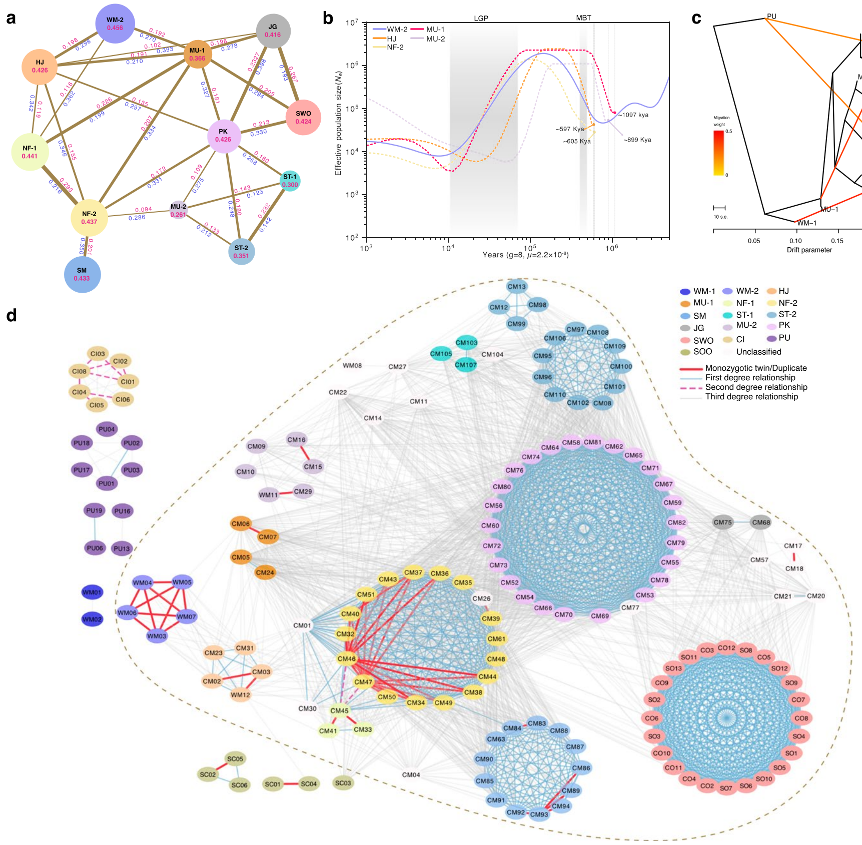

Fig. 3 Inference of genetic relationship of LSMs. **a** The Kinship coefficient and Fst within 12 groups. Green represents the average *F*st value. **b** Effective population size and split time inferred by SMC++. MBT: mid-Brunhes climate transition, LGP: Last Glacial Period. **c** Migration events between different citrus groups were inferred by TreeMix. A line with an arrow corresponds to the event and direction of the migration. Outgroup is set as IC. **d** Kinship relationships estimated for 167 accessions by KING software. Different lines represent different kinship levels, and different colors of ellipses represent different groups

(Fig. 2a), underscoring their pioneering role in the domestication of cultivated citrus in China.

Origin and domestication history of LSMs

We found that WM-1 was highly differentiated from the other groups (average *F*st ~ 0.35) (Additional file 2: Table S10), and Wu et al. proposed that it be recognized as a subspecies [5]. Furthermore, there was a distinct internal genetic divergence within WM-1 and an evident lack of genetic relationship with other LSMs (Fig. 3d). Concurrently, we noted that WM-1 exhibited lower nucleotide diversity ($\pi = 2.00 \times 10^{-3}$) than did the other LSMs (Additional file 1: Fig. S5), which could be attributed to a severe population bottleneck event, geographic isolation, or the absence of pummelo gene introgression [1].

In contrast, although WM-2 was also highly differentiated from the other groups (*F*st ~ 0.33) (Additional file 2: Table S10), it was more closely related to the cultivated varieties (Fig. 3a and d). Hence, we performed identity-by-descent analysis with KING software [18]. Among them, WM-2 showed a third-degree relationship with CM02, CM03, and CM31 in HJ. Similarly, WM05 and WM07 in WM-2 exhibited a third-degree relationship with CM06 in MU-1 (Fig. 3d and Additional file 2: Table S11). These findings suggest that compared with WM-1, WM-2 may be the earliest direct ancestor of modern cultivated LSMs.

To further infer the divergence of WM-2 from the other groups, we used SMC + + to infer the population dynamic history and split time of the different groups. The results showed that MU-1, HJ, NF-2, and MU-2 diverged from WM-2 during the first bottleneck of WM-2 approximately 2076–680 kilo years ago (KYA), with MU-1 and MU-2 splitting early (~ 1100–900 KYA) and HJ and NF-2 parting later (~ 605–597 KYA) in the bottleneck period (Fig. 3b). Following separation from WM-2, the effective population sizes (*Ne*) of HJ, MU-1, NF-2, and MU-2 gradually recovered from ~ 600 KYA to reach a maximum (Fig. 3b). This timeframe aligns closely with the mid-Brunhes climate transition (MBT) [19]. A warmer interglacial period may have caused the ancestors of these ancient LSMs to disperse with glacial meltwater from the Nanling Mountains into the Yangtze River Basin (HJ, MU-1, and NF-2) and the Pearl River Basin (MU-2) (Additional file 1: Fig. S6). This geographical dispersion likely led to further differentiation and independent domestication between the two regional populations. During this period, ST-1 diverged from MU-2 at ~ 480 KYA, and both experienced a significant population bottleneck at ~ 110–180 KYA (Additional file 1: Fig. S7). Additionally, at ~ 100–10 KYA, HJ, MU-1, and NF-2 underwent a substantial reduction in their effective population size (*Ne*), possibly due to the effects of the Last Glacial Period (LGP) (Fig. 3b) [20].

After initial domestication, artificial selection accelerated citrus improvement. We observed that NF-1 contained mixed genetic material, with a substantial contribution of genetic material originating from or similar to that of NF-2 and MU-1, indicating that NF-1 may be a hybrid offspring. We found that CM24 from MU-1, along with NF-1 and NF-2, was cultivated in Nanfeng County, Jiangxi Province. Moreover, CM30 displayed a niche and genetic constitution similar to that of NF-1 (Fig. 2c). Furthermore, kinship inference revealed a first- and second-degree relationship between NF-2 and NF-1 (Fig. 3a and c). And we detected a TreeMix migration edge from MU-1 to NF-1, further supporting our hypothesis (Fig. 3c). We also believe that NF-2 may serve as a parent to Satsuma mandarin (SM, a group widely cultivated in China, known for its excellent fruit quality and strong cold resistance, and also a commonly used parent in citrus breeding), with a kinship coefficient of 0.201 (Fig. 3a). A parent–offspring relationship was identified between CM46 (a centenarian NF-2) and CM84 (SM) by KING (Fig. 3d and Additional file 2: Table S11). Then, evidence of gene flow from NF-2 and PU into SM was detected using TreeMix and the *fb* statistic (Fig. 3c and Additional file 1: Fig. S8d). These findings suggest that NF-2 is the parent of SM and has experienced introgression from PU, which is consistent with the hypothesis proposed by Wu et al. [12]. Furthermore, it also indicates that NF-2 has very high breeding value. On the other hand, compared to other cultivated citrus varieties, ST-2 exhibited the lowest nucleotide diversity ($\pi = 2.200 \times 10^{-3}$) (Additional file 1: Fig. S5). A third-degree relationship and low *F*st (~ 0.142) were maintained between ST-2 and ST-1, indicating that ST-2 might have undergone further domestication from ST-1 (Fig. 3a and b).

These findings indicate that cultivated mandarins in China can be clearly divided into two large groups, the Yangtze River group (YZ) and the Pearl River group (PR), according to geographical region. YZ comprises varieties such as MU-1, HJ, NF-1, and NF-2, while PR includes MU-2, ST-1, and ST-2. The average intragroup kinship coefficients for YZ and PR were approximately 0.195 and 0.169, respectively (Additional file 2: Table S10 and 11). Furthermore, there was a notable absence of kinship and gene flow between these two groups, which provides support for the hypothesis of independent domestication between them [1].

Intriguingly, we discovered that Ponkan (PK) acts as a crucial connector between YZ and PR, each sharing third-level kinship with PK, as evidenced by kinship coefficients of 0.163 and 0.150, respectively (Fig. 3a and d). The detection of a TreeMix migration edge from ST-2 to PK suggested that the gene flow originated from ST-2 (Fig. 3c). This assertion was corroborated by *F*st estimates, which revealed minimal differentiation between PK and ST-2 ($F\text{st}=0.248$) within PK-related groups (Fig. 3a and Additional file 2: Table S10). Furthermore, the *D statistics* revealed slight PU introgression in PK ($D\text{-}statistic=0.08$, $|Z|=4.64$, $P=3.41\times10^{-6}$) (Additional file 2: Table S12), potentially contributing to its larger fruit size. As the most widely cultivated citrus in China [2] and a key parental cultivar in citrus hybrid breeding, the strong adaptability and breeding potential of PK is likely attributed to its mixed genetic composition. Additionally, SWO exhibited a third-degree kinship with both PK and MU-1 (Fig. 3a and d). Furthermore, the TreeMix migration edge from PU to SWO revealed significant gene flow from PU, reinforcing the conclusion that SWO is a descendant of LSM and PU (Fig. 3c) [12].

Notably, the kinship coefficients of JG with SWO, PK, and MU-1 were 0.267, 0.233, and 0.198 (Fig. 3a and d), respectively. This further substantiates the inference that JG is a natural hybrid offspring of sweet orange and LSMs [21]. Phylogenetic analysis of the chloroplast genome revealed that the chloroplast of JG is similar to that of LSMs (Additional file 1: Fig. S8e), indicating that LSM is the maternal parent of JG. Further investigation revealed that the ecological niche of JG is close to that of PK but significantly distant from that of MU-1 [2], lending credibility to the belief that PK is the maternal parent of JG.

Identification of potential domestication and improvement selection regions through selective sweep analysis

The domestication and improvement of mandarin involved key transitions in several traits, including a significant increase in fruit sweetness in the early domesticated group (NF-2) compared to that in the wild mandarins and a further increase in fruit size in the recently improved elite cultivar (Satsuma mandarin). We measured the sugar contents of WM02 and NF-HZ, which are representative varieties of the WM-2 and NF-2 groups, respectively. The results showed that the contents of sucrose, fructose, and glucose in NF-HZ were significantly higher than those in WM02 (Fig. 4b). In addition, compared with NF-2, the average fruit weight and vertical and transverse diameters of SM increased 3.7-, 1.7-, and 1.5-fold (Fig. 4c, Additional file 1: Fig. S9 and Additional file 2: Table S13), respectively. To identify the genomic regions selected during domestication and improvement, we used the NF-2 vs. WM-2 and SM vs. NF-2 pairs for scanning. Selection sweeps between NF-2 and WM-2 were postulated to be involved in domestication, while those between SM and NF-2 were associated with breeding improvements (see “Methods”). In total, we identified 1100 domestication and 863

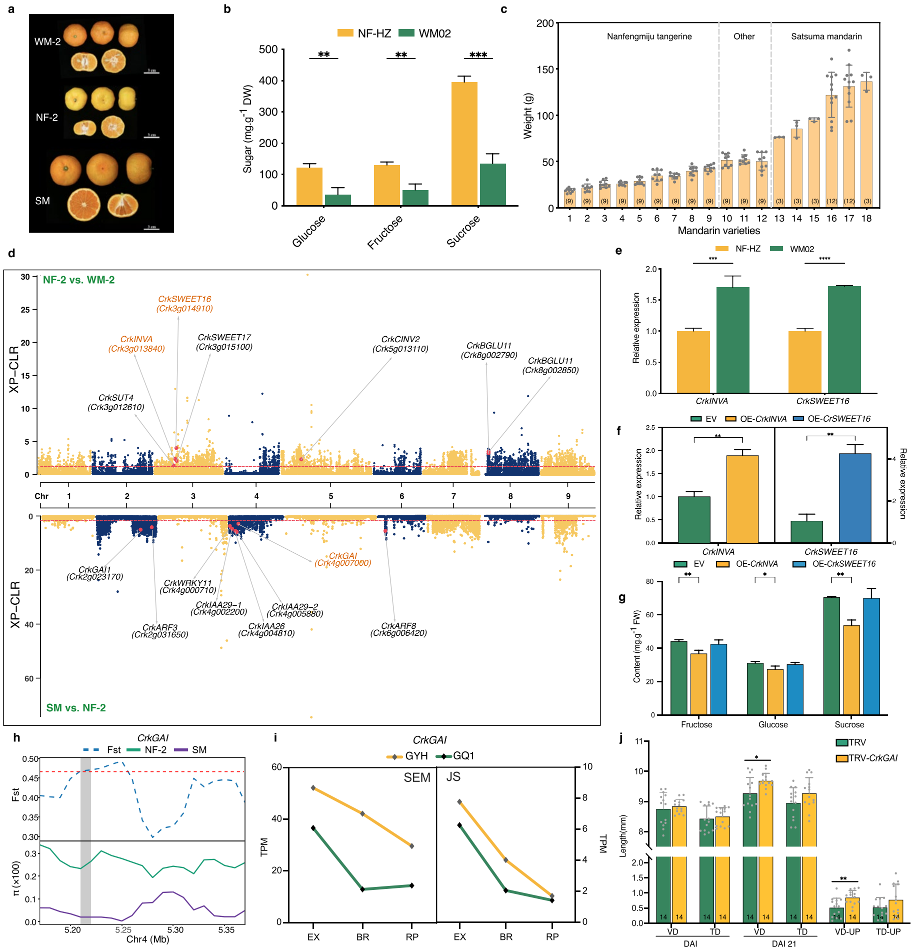

Fig. 4 LSM domestication and improvement selection sweep region. **a** Phenotypes of WM-2, NF-2 and SM. The representative varieties Daoxian wild mandarin2 (WM02), Nanfengmiju tangerine "Guiyuehong" (GYH) and Satsuma mandarin "Oita No.4" (ON4) of WM-2, NF-2, and SM are shown in the figure. **b** Soluble sugar content in the NF-HZ and WM02. The asterisk indicates significance two-sided Student's t test, The *P* values for glucose, fructose, and sucrose are 0.0040, 0.0029, and 0.0002. **c** Fruit weight of Nanfengmiju tangerine and Satsuma mandarin. The LSM varieties corresponding to the numbers below the bar plot are provided in Additional file 2: Table S13. **d** Genome-wide selective signals (XP-CLR score) of NF-2 vs. WM-2, and SM vs. NF-2. Regions with both XP-CLR and fixation index (*Fst*) values in the top 10% were regarded as having selective signals, while the red dashed lines are the threshold of the top 10% XP-CLR score. **e** Relative expression of the candidate genes of NF-2 vs. WM-2. The asterisk indicates significance of the two-sided Student's t test. The *P* values for *CrkINVA* and *CrkSWEET16* are 0.0003 and <0.0001. **f** Relative expression of the *CrkINVA* and *CrkSWEET16* in OE-*CrkINVA* and OE-*CrkSWEET16* lines (detached fruits). The asterisk indicates significance two-sided Student's t test, The *P* values from left to right are 0.0034 and 0.0016, respectively. **g** Fruit soluble sugar content of OE-*CrkINVA* and OE-*CrkSWEET16* (detached fruits). The asterisk indicates significance of the two-sided Student's t test. The *P* values from left to right are 0.0042, 0.0400, and 0.0009, respectively. **h** Signals of artificial selection in the CrkGAI gene. Purple and green solid lines indicate π value statistics in SM and NF-2, respectively. The blue dashed line indicates the *Fst* value between SM and NF-2, while the red dashed line is the threshold of the top 10% *Fst*. **i** Expression patterns of *CrkGAI* during the development of the GYH and GQ1 fruit segment membrane (SEM) and juice sac (JS). EX: the expansion period, CP: the breaker period, RP: the ripening period. **j** Fruit size of TRV and TRV-*CrkGAI*. VD and TD: vertical and transverse diameters of fruit. DAI: the day after injection. UP: length of increase at 21 days after injection. The asterisk indicates significance two-sided Student's t test, The *P* values for DAI21-VD and VD-UP are 0.0126 and 0.0037

improvement-selective sweep regions, containing 2396 and 2333 protein-coding genes, respectively (Additional file 2: Table S14 and 15).

Within the domestication-selective sweep regions, several genes involved in sucrose catabolism and sugar transport were identified (Fig. 4d and Additional file 2: Table S14). Among them, *CrkSUT4* (Crk3g012610), *CrkINVA* (Crk3g013840), *CrkSWEET16* (Crk3g014910), and *CrkSWEET17* (Crk3g015100) were concentrated in a 1.7-Mb region on chromosome 3 (Additional file 1: Fig. S10A). Notably, the expression levels of *CrkSWEET16* and *CrkINVA* in mature NF-HZ fruit were significantly lower than those in WM02 fruit (Fig. 4e). The haplotype network analysis revealed that modern LSMs share the domesticated haplotypes H7 and H4 of *CrkINVA* and *CrkSWEET16*, respectively, while the haplotypes of wild mandarins are more ancient (H3 and H1), suggesting their functional differentiation in wild and modern LSMs (Additional file 1: Fig. S10b). Furthermore, we examined the expression patterns of these genes during the fruit development and ripening of GYH (a high-sugar citrus cultivar from the NF-2 group). The results showed that the expression levels of *CrkSWEET16* and *CrkINVA* increased at the ripening stages of citrus fruits (EX, BR, and RP) (Additional file 1: Fig. S10c), suggesting that these genes can regulate sugar accumulation throughout the fruit ripening process. To verify our hypothesis, we overexpressed *CrkSWEET16* and *CrkINVA* in detached mature Nanfengmiju fruits and measured the sugar content after 7 days. The results revealed that, compared to EV, the average contents of sucrose, fructose, and glucose decreased by 23.91%, 16.85%, and 11.9%, respectively, in the OE-*CrkINVA* line (Fig. 4f and g). However, overexpression of *CrkSWEET16* did not significantly impact the sugar content (Fig. 4f and g). To further substantiate the function of *CrkINVA*, we overexpressed this gene in mature Nanfengmiju fruits on the tree, and the sucrose content was also significantly decreased (Additional file 1: Fig. S10d and e). These results suggested that *CrkINVA* plays an important role in regulating sugar content in citrus fruits.

In addition, we found multiple homologous genes involved in hormone signal transduction regulation (Fig. 4d). Among them, two DELLA protein-coding genes, *CrkGAI* (Crk4g007000) and *CrkGAI1* (Crk2g023170), were identified on chromosomes 2 and 4; these genes exhibited significantly decreased π values in elite SM and a high *F*st score above the threshold (Fig. 4h and Additional file 1: Fig. S12). These compounds have been reported to inhibit plant growth by inhibiting the expression of genes involved in the gibberellin (GA) pathway [22, 23]. Interestingly, by comparing the transcriptomes of Satsuma mandarin "Guoqing No. 1" (GQ1) and GYH pulp (SEM and JS) across 4 developmental periods, we found that the expression level of *CrkGAI* in GQ1 was lower than that in GYH from the fruit expansion stage to the ripening stage and showed a continuous downward trend, indicating that it may inhibit citrus fruit growth (Fig. 4i). Additionally, compared to other LSMs, SM shares the more recently evolved haplotype H4 with SWO (Additional file 1: Fig. S11a). To further investigate the function of *CrkGAI* in fruit development, we specifically silenced *CrkGAI* in Hongkong kumquat fruit (*Fortunella hindsii*) via a virus-induced gene silencing (VIGS) approach. At 21 days after injection (DAI 21), both the vertical and transverse diameters of the fruits and their growth amount of TRV-*CrkGAI* increased compared with the TRV control (Fig. 4j). Moreover, the weight of the fruits of the TRV-*CrkGAI* lines increased by 15.4% (Additional file 1: Fig. S11c, d). This result demonstrated that *CrkGAI* could significantly negatively regulate citrus fruit size. In addition, we found that multiple homologous genes involved in auxin signal transduction regulation, including *CrkARF8*, *CrkWRKY11 CrkIAA26*, *CrkIAA29*, and *CrkARF3*, showed strong artificial selection supported by a high *F*st score and significantly decreased π value in SM (Additional file 1: Fig. S12), suggesting a change in the auxin signaling pathway.

Allelic expression pattern of Nanfengmiju tangerine during fruit development

The fully phased diploid genome provides a platform for obtaining a comprehensive and accurate understanding of allele expression. We obtained 21,031 pairs of allelic genes using BLAST (see "Methods"). Next, we performed the transcriptomes profile of the Nanfengmiju tangerine GYH fruit at 5 developmental stages from 3 tissues (Additional file 2: Table S16). PCA revealed that the fruit transcriptional landscape was mainly affected by genome-wide allele-specific expression (ASE), followed by gene expression in different tissues (Fig. 5a).

Across 15 samples of GYH, we detected 5809 (27.6% of total pairs of allelic genes) ASE genes (ASEGs) and 5185 allele-equivalent expression (AEE) genes (AEEGs), with an average of 4531 ASEGs in each sample (Fig. 5b, Additional file 2: Table S17 and 18). The proportion of ASEGs was much greater than that in diploid potato (16%) and apple (19%) (Sun et al., 2020; Zhou et al., 2020). Among these genes, approximately 60% of the ASEGs were globally expressed during the development of three fruit tissues, while only approximately 7.8% of the ASEGs were expressed in one sample (Fig. 5c and Additional file 1: Fig. S13a).

In addition, the ASEGs and AEEGs were enriched in distinct biological process (BP) and molecular function (MF) terms. For instance, the ASEGs were significantly enriched in the terms "regulation of transcription by polymerase II," "mRNA splicing via spliceosome" (BP), and "ubiquitin-protein transferase activity" (MF), while the AEEGs were significantly enriched in "response to abscisic acid," "chloroplast organization" (BP), and "glycosyltransferase activity" (MF) (Additional file 1: Fig. S14 and Additional file 2: Table S19). These findings indicate dissimilar functional roles for ASE and AEEGs during fruit development.

Genetic variants can lead to ASE by altering transcription factor binding sites, chromatin accessibility, and posttranscriptional processes (Cleary and Seoighe, 2021). We detected 725,589 genetic variations in 5079 ASEGs (87.4% of all ASEGs), and 736,232 genetic variations in 5182 AEEGs (99.9% of all AEEGs) (Additional file 2: Table S20). Nevertheless, we detected 4.23 deleterious, 4.63 missense, and 1.97 splice region variations in each ASEG, which was significantly greater than the AEEGs (Fig. 5d). Additionally, compared with AEEGs, the 2-kb upstream flanking regions (up2k), exons, introns, and 2-kb downstream flanking regions (down2k) of the ASEGs had greater SNP density (Fig. 5e). These abundant allelic variations may be an important cause of ASE. For instance, SWEET15 is an important transporter protein for unloading sugars in fruit pulp (including segment membrane (SEM) and juice sac (JS) tissues) and may be a key factor in sugar accumulation in citrus fruits (Feng et al., 2021). We observed that the two allelic genes encoding SWEET15 in GYH, *H1_Crk7g001990* and *H2_Crk7g001990*, showed almost identical expression levels in the early developmental stages of the SEM and JS tissues. However, as fruit development progressed, the difference in expression

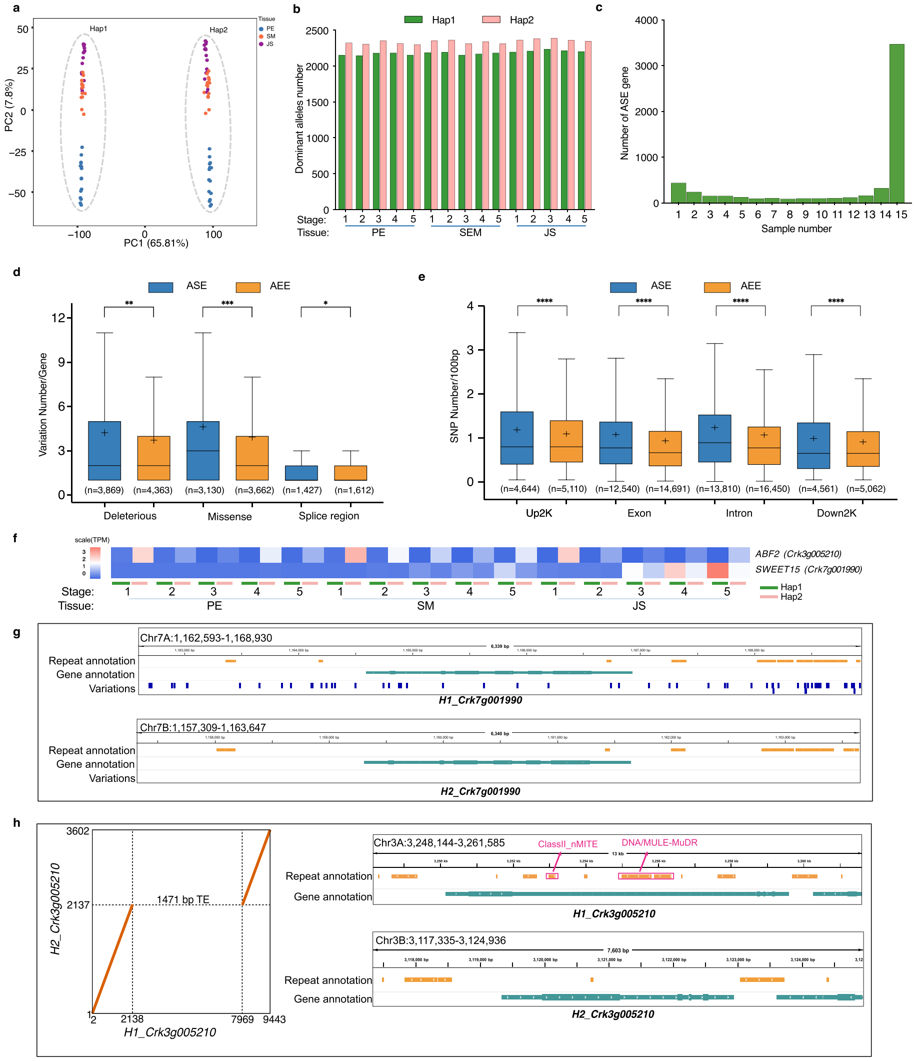

Fig. 5 Allele-specific gene expression of Nanfengmiju tangerine. **a** Principal component analysis (PCA) of GYH fruit allele expression profiles at 5 different development stages in 3 tissues. PE: peel, SEM: segment membrane, JS: juice sac. **b** Dominance expression pattern of allele-specific expressions (ASEs) at 5 different development stages in 3 tissues of GYH fruit. **c** Distribution of genes with allele-specific expression (ASE genes) in 15 GYH samples. **d** Statistical of deleterious, missense, and splice region variations in ASE genes and genes with allele equivalent expression (AEE genes). The high and moderate variations have been classified as deleterious variations. The asterisk indicates significance two-sided Student's t test, The *P* values for deleterious, missense, and splice region are 0.0026, 0.0004, and 0.0169. "+" indicates means. "n" represents the number of genes. **e** SNP density of ASE genes and AEE genes. The asterisk indicates significance two-sided Student's t test. The *P* values for up2k (2Kb upstream), exon, intron, and down2K (2Kb downstream) of gene are <0.0001, <0.0001, < 0.0001, and <0.0001. "+" indicates means. "n" represents the number of elements. **f** Expression patterns and variations of two haplotypes of *CrkABF2* and *CrkSWEET15*. **g** Allelic variations around *CrkSWEET15* between the two haplotypes. h Structure variation around *CrkABF2* between the two haplotypes

levels between the two allelic genes gradually increased, with *H1_Crk7g001990* becoming the dominant allele in the later stages of development (Fig. 5f). Examination of their sequences revealed numerous variations within the gene and up2k regions, which likely led to the allele dominance of *H1_Crk7g001990* during fruit development (Fig. 5g). Furthermore, transposon insertion may also be a significant reason for ASE. For example, *ABF2* (Crk3g005210) exhibited strong ASE, with the expression level of its dominant allele, *H2_Crk3g005210*, far exceeding that of *H1_Crk3g005210* in all developmental stages and tissues (Fig. 4f). Interestingly, we identified 1471 bp transposon insertion (including two DNA/MULE-MuDR and one ClassII_nMITE transposon) within the intron of *H1_Crk3g005210* (Fig. 5h), suggesting that the insertion of these transposons might significantly affect the transcription of *H1_Crk3g005210.*

Discussion

Here, based on PacBio HiFi reads and Hi-C data, we describe high-quality de novo chromosome-level and haplotype-resolved genome assemblies for Nanfengmiju tangerine. Compared with the monoploid genome, the haplotype-solved genome contains the genetic information of each chromosome of polyploid species, which provides rich information for exploring the genetic content, allele expression, and regulation in heterozygous genomes.

Allele-specific expression has been reported in several horticultural crop species, such as kiwifruit [9], apple [24], sugarcane [25], litchi [26], and potato [17]. Our study revealed high-resolution profiles of global allele expression during citrus fruit development. We detected a large amount of allele-specific expression during the developmental stages of Nanfengmiju fruit. This pattern of allele expression dominance demonstrated marked temporal and spatial consistency (Additional file 1: Fig. S13), reflecting stringent transcriptional regulation and playing an important role in citrus development. Such regulation is integral to the developmental processes of citrus fruits. Analysis revealed an increased prevalence of sequence variations within different elements of ASEGs. These variations impact both the transcriptional regulation and post-transcriptional processing (such as RNA splicing) of alleles, potentially leading to differences in allele expression or triggering non-sense mRNA-mediated decay (NMD) [27]. In addition, the insertion of transposons in genes may also be a cause of ASE. The specific functional roles of these genetic variants, however, remain unclear.

In this study, we present a large-scale survey of LSM genomes from southern China. Utilizing genome data, we categorized the LSMs into 12 distinct subgroups. Based on the detailed classification, we rearranged and refined the history of LSM domestication. The Mangshan wild mandarin (WM-1) is considered the oldest LSM. But, the origins and genetic relationships of the two wild species, Mangshan wild mandarin and Daoxian wild mandarin, remain unclear. Wu et al. suggested that Daoxian wild mandarin is a hybrid offspring of Mangshan wild mandarin and common mandarins [5], while Wang et al. proposed they are the typical mandarins [1]. The population structure inferred by ADMIXTURE showed that WM-1 was mixed (Fig. 2c), and WM-2 was the ancestral population of WM-1, which seems to contradict the findings of previous studies. Nonetheless, Lawson et al. suggested that when STRUCTURE/ADMIXTURE is used to infer the ancestor composition of a group, the offspring of a small group generated by the extreme population bottleneck of the ancestor may be listed separately, resulting in a mixed situation of real ancestors [28]. Notably, Wang et al. consistently identified an extreme population bottleneck in WM-1 [1]. Furthermore, considering the observed gene flow from WM-1 to WM-2 in the absence of pummelo introgression (Fig. 3c and Additional file 1: Fig. S8d) [1, 5], we propose a two-step formation process for WM-2:

initial divergence from the WM-1 ancestors through an extreme bottleneck event, followed by WM-1 introgression that increased its nucleotide diversity (Fig. 6).

We believe that there are two independent domestication groups, YZ and PR, in the Chinese LSMs, and WM-2 is the most direct ancestor, which is similar to the hypothesis of Wang et al. [1]. However, a notable divergence from their hypothesis is that YZ and PR were not directly domesticated from WM-2 by human intervention. Instead, their ancestors diverged from WM-2 between 500 and 1000 KYA, much earlier than the time of human activity in the region [4]. Based on the analysis, it is hypothesized that the

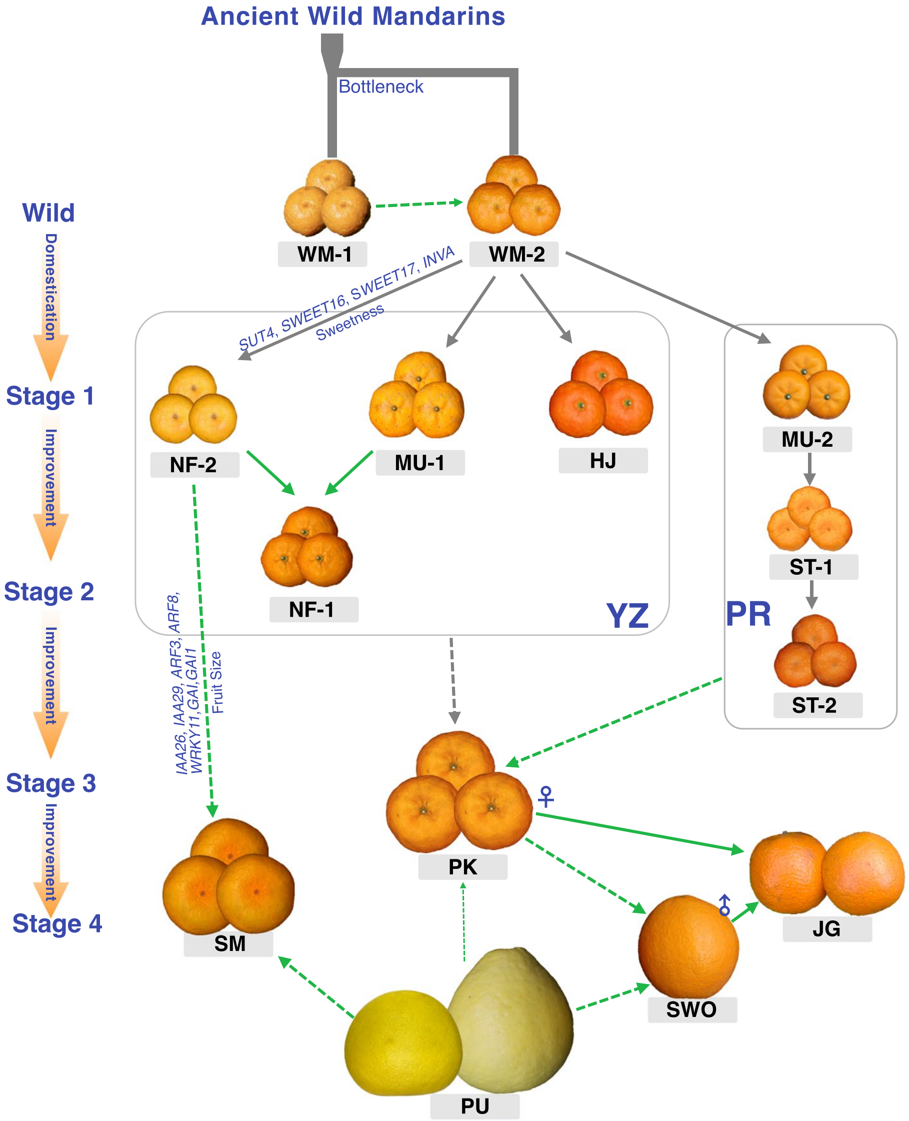

Fig. 6 Proposed model of loose-skin mandarin speciation and breeding history. The gray solid line represents the possible direction of species formation and domestication, while the green line indicates the direction of introgression processes. Green solid lines represent confirmed parent–offspring relationships, while green dashed lines suggest possible parent–offspring relationships. The gray dashed line represents suggest possible parent–offspring relationships, but no evidence of introgression was detected. The thicker the green line, the more significant the gene introgression. The selected loci and candidate genes in this study were displayed above and below the arrows, respectively

domestication of Chinese LSMs involved at least four distinct stages (Fig. 6). Initially, the progenitors of ancient Chinese LSMs, specifically HJ, MU-1, MU-2, and NF-2, likely diverged from WM-2 between approximately 500 and 1000 Kya (Fig. 3b). Concurrently, during a warm interglacial period, the melting glaciers contributed to their dispersion into the Yangtze River and Pearl River Basins (ST-1 was splitted from MU-2 during this period (~480 KYA)). Prolonged geographical isolation facilitated their respective domestication, resulting in the formation of distinct varieties.

After the initial domestication, we divided the development of the remaining LSMs into three stages based on the degree of pummelo gene introgression, with artificial selection playing a key role in significantly accelerating this process. In the second stage, there was little to no pummelo gene introgression, but clear evidence of artificial selection. The mixture of NF-2 and MU-1 led to the emergence of NF-1, while ST-1 underwent further domestication, giving rise to ST-2. The third stage involved the development of Ponkan (PK), which demonstrated a close genetic relationship with PR and YZ. Functioning as a genetic bridge between PR and YZ, it is conjectured that PK contains a mixture of the genetic information of PR and YZ with slight introgression of pummelo (PU) genes, thereby enhancing its adaptability, fruit size, and breeding potential. The final stage involved significant introgression of PU genes, culminating in the creation of elite varieties such as Satsuma mandarin (SM), Sweet Orange (SWO), and Jiaogan (JG) (Fig. 6). These hypotheses offer a more nuanced understanding compared to previous studies, linking the diversification of LSMs with paleoclimatic events. However, the history of LSM domestication remains complex and intriguing. The existing evidence is insufficient for a comprehensive reconstruction of the entire domestication process, necessitating further investigation through fossil and genomic studies. It is worth noting that we found that NF-2 is a parent of NF-1 and SM (Figs. 2c and Fig. 6), indicating that NF-2 has very high breeding value. In the future, using NF-2 as a parent, there is potential to breed new loose-skin mandarin varieties with high quality and strong adaptability.

Fruit size, acidity, and sweetness are key traits of domestication and are prominently targeted in current citrus breeding programs. Previous studies have shown that a decrease in acidity is a sign of citrus domestication, while the role of sugar selection has been traditionally underestimated [1, 3]. However, we observed that the sugar content of NF-2 was significantly greater than that of WM-2, suggesting that increased sugar levels also indicate citrus domestication. Citrus fruit sweetness is primarily determined by the accumulation of sucrose, fructose, and glucose in the JS [29]. We identified several genes involved in sugar transport and sucrose metabolism within the domestication selection region, including *SUT4*, *SWEET16*, *SWEET17*, *CINV2*, and *INVA*. Notably, *SUT4*, *SWEET16*, *SWEET17*, and *INVA* were clustered within a 1.7-Mb segment on chromosome 3, suggesting simultaneous selection during domestication. In the SEM and JS developmental transcriptomes of GYH, the expression levels of *INVA* and *SWEET16* increased suddenly from BR to RP, suggesting that these genes play important roles in the late stage of fruit ripening. SWEET16 is located on the vacuolar membrane and can transport sucrose, fructose, and glucose bi-directionally [30]. INVA, a mitochondrial enzyme, catalyzes the conversion of sucrose to glucose and fructose, supplying substrates for mitochondria-associated hexokinase [31]. Interestingly, lower expression levels and domesticated haplotypes of *SWEET16* and *INVA* were detected in mature NF-2 fruit, suggesting that the high expression of these genes at maturity may lead to decreased fruit sweetness. Specifically, the ability of *INVA* to align with this hypothesis was validated through subsequent experimental analysis. Although the precise functions of the other genes remain to be fully determined, their apparent co-selection and insights from prior studies underscore their substantial role in the domestication of citrus sweetness.

In addition, we explored genetic changes during the process of SM improvement. Multiple gibberellin (*GAI* and *GAI1*) and auxin (*IAA26*, *IAA29*, *ARF3*, *ARF8*, and *WRKY11*) response genes were annotated in the improved selection sweep region. The expression patterns of GQ1 and GYH were similar during fruit development, but the expression levels significantly differed. *GAI* and *GAI1* are genes encoding DELLA proteins, which are key inhibitors in the gibberellin signaling pathway. Their expression levels continuously decrease after EX, potentially facilitating cell expansion by releasing gibberellin signals and thus promoting fruit growth [32]. Notably, in comparison to NF-2, SM shares a newer *GAI* haplotype with SWO, and *GAI* exhibits a lower expression level during the maturation process of SM fruits, suggesting its regulatory role in fruit size determination. This hypothesis was further substantiated through VIGS experiments. In addition, the transient silencing of *GAI* led to accelerated fruit coloration, demonstrating its potential function in fruit ripening regulation, which is worthy of further study. Auxin, gibberellin, and auxin-gibberellin interactions are key factors affecting fruit growth [32], with the ARF and Della protein complex serving as a critical junction in the crosstalk between auxin and GA signals [33]. During the process of improving SM, the genetic regulatory networks of auxin and gibberellin were inadvertently altered, leading to changes in the size of citrus fruits. These findings provide actionable targets for high-quality precision breeding and fruit size improvement in citrus.

Conclusions

Based on the genome of ancient citrus variety, we clarified the evolutionary relationships among loose-skin mandarins, and found that the artificial selection of *INVA* and *GAI* changed the sweetness and size of fruits, respectively. Furthermore, by constructing the haplotype-resolved genome, we provided an allelic gene expression atlas of Nanfengmiju tangerine. Taken together, our study provides valuable genomic resources and further revises the origin and domestication history of LSMs, offering insights for genetic improvement of citrus plants.

Methods

Sample collection and DNA sequencing

The NFMJ-120y sample were collected in Nanfeng County, Jiangxi Province, and used for PacBio sequencing and de novo genome assembly. In addition, we collected 76 accessions of the most widely cultivated LSMs in China. Young leaves of 77 accessions were washed with sterile water, frozen immediately with liquid nitrogen, and stored at − 80 °C. Total DNA was extracted from young leaves, and whole-genome resequencing was performed using the Illumina NovaSeq platform with a read length of 150 bp.

Reads of approximately 570 Gb were sequenced on the Illumina NovaSeq platform with the 150-bp paired-end sequencing model at Novogene (Beijing, China).

Genome assembly

The Illumina reads of NFMJ-120y (Additional file 2: Table S1) were first used to estimate genome size and heterozygosity with Jellyfish [34] and GenomeScope 2.0 [35]. We assembled the NFMJ-120y genome by incorporating PacBio single-molecule real-time long-read sequences and sequences from high-throughput chromatin conformation capture (Hi-C) technology. A total of 20 Gb of PacBio HiFi reads and 32 Gb of Hi-C reads were generated from the PacBio Sequel II platform and Illumina NovaSeq platform, respectively, at Novogene (Beijing, China). These HiFi and Hi-C reads were assembled using Hifiasm v1.6 (Hi-C Integrated Assembly) [36] to generate one primary genome (Monoploid) and two haplotype draft contig genomes (Haplotype1 and Haplotype2). For each primary assembly, we mapped the Hi-C data to the corresponding contigs using Juicer v1.9.9 [37] and built primary scaffolds with 3D-DNA v180922 [38]. Juicebox Assembly Tools v1.11.08 was used to visualize and manually curate the assembly and obtain the final assembly [39]. The final assembly was aligned to the *Citrus sinensis* v3.0 [13] genome with NUCmer [40] to identify pseudomolecules corresponding to chromosome IDs in *Citrus sinensis* v3.0. Finally, Benchmarking Universal Single-Copy Orthologous (BUSCO v5.2.2) gene analysis [41] and QUAST v5.0.2 [42] were used to evaluate the completeness and continuity of the assembled genome, respectively.

Genome annotation

For each assembly, LTRharvest [43], LTR_Finder v1.07 [44], and MITE-Hunter [45] are used to initially identify long-terminal repeat retrotransposons in the genome. LTR_retriever v2.8 [46] was used to integrate the above results and calculate LTR Assembly Index (LAI). Following the integration of LTRs, RepeatMasker (http://www.repeatmasker.org/) is used to mask these elements in the genome, and RepeatModeler (http://www.repeatmasker.org/) annotates the remaining repetitive sequences. For the unclassified repetitive sequences, further comparison and classification are carried out through protExcluder [47] and DeepTE [48]. Finally, all predicted results are integrated, and RepeatMasker is utilized for masking and statistical analysis of repetitive sequences, with manual correction based on the classification methods of Wicker et al. [49] and EDTA (https://github.com/oushujun/EDTA/blob/master/util/TE_Sequence_Ontology.txt).

The gene model of each assembly is annotated using the MAKER genome annotation pipeline [50]. The inputs for MAKER annotation primarily include the following: (1) Gene models trained using Genemark [51] and SNAP [52] based on RNA-seq data (derived from leaves, stems, roots, and fruits of NFMJ-120y). (2) Plant protein sequences from the SwissProt database (https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/uniprot_sprot_plants.dat.gz) annotated Citrinae species protein sequences (http://citrus.hzau.edu.cn/index.php), and transcriptome assembly obtained with Trinity [53], serving as homology evidence. (3) Repetitive sequences identified in the previous step. Additionally, MAKER incorporates the ab initio gene predictions software AUGUSTUS [54]. We conducted three iterations of

MAKER to achieve more refined annotations. Finally, the annotations from MAKER were further refined using PASA [55].

The gene functions are annotated using InterProScan v5.55–88.0 with default parameters [56]. Additionally, a database is established using plant protein sequences from the SwissProt database, and homology searches are conducted using BLAST + v2.9 [57] with parameters " -evalue 1e-10 -outfmt 6 -num_alignments 1 -max_hsps 1", retaining the best matches for annotation purposes. Genes that do not match are then compared against plant protein sequences from the UniProtKB TrEMBL database. Finally, the results from InterProScan and BLAST + are consolidated.

Variation detection of two haplotypes

The assemblies of Haplotype1 and Haplotype2 were aligned using Nucmer alignment tool from Mummer v4.0 [40] with default parameters. Next, filtered using delta-filter with parameters "-m -i 90 -l 100". The alignments were then used for variation detection with the SyRI [58] and visualization using plotsr. Finally, annotation of variations was performed with snpEff v4.3 [59].

Comparative genomics analyses

Non-redundant protein sequences from 9 species were prepared for ortholog analyses (Additional file 2: Table S7). Gene family clusters and single-copy ortholog sequences within the protein set were identified using OrthoFinder v2.3.8 [60]. The identified single-copy orthologous genes were utilized to construct a maximum-likelihood phylogenetic tree. Protein sequences were aligned using MUSCLE v3.8 [61], and conserved sites were extracted with Gblocks [62]. The phylogenetic tree, with Atalantia as the outgroup, was then constructed using RAxML v8.2.12 [63]. Species divergence times were estimated through mcmctree in the PAML toolkit [64], and performed time calibration based on the citrus fossil *C.linczangensis* [65], and set the divergence time of *Citrus* and *Poncirus trifoliata* to 7–10 million years (Mya). Gene family expansion and contraction analyses were conducted using CAFE5 [66], and KEGG pathway annotations were performed with KOBAS v3.0 [67].

Mapping and variant calling

A total of 191 citrus resequencing datasets, including 77 new datasets and 114 previously published datasets, were used for population genetic and phylogenetic analyses. Raw data filtering was performed with TrimGalore v0.6.6 with the parameters "-q 25". The quality-controlled data were aligned to the monoploid assembly using BWA-MEM [68]. Format conversion, sorting, and statistical analysis were carried out with SAMtools [69]. MarkDuplicates from GATK4 [70] was used for PCR duplicate removal, followed by the use of HaplotypeCaller call individual-specific gvcf files. These were then merged using CombineGVCFs. Finally, GenotypeGVCFs was used for joint calling of SNPs. SelectVariants was utilized to extract SNPs and InDels. After quality control of the SNPs using BCFtools, the SNPs were hard filtered using GATK VariantFiltration (QD < 2.0 || MQ < 40.0 || FS > 60.0 || SOR > 3.0 || MQRankSum < − 12.5 || ReadPosRankSum < − 8.0), and only biallelic SNPs were selected for further analysis.

Principal component analysis, phylogenetic analysis, and ADMIXTURE analysis

Before analysis, SNPs were filtered using VCFtools v0.1.16 [71] to ensure that the minor allele frequency (MAF) < 0.05, the missing rate < 0.2, the minor allele count (MAC) > 3, and the minimum depth (minDP) > 3. Subsequently, SNPs with $R^2 > 0.1$ were removed using Plink v1.9 [72] with the parameters "indep-pairwise 50 5 0.1". Finally, 192,855 SNPs were retained.

PCA was performed using Plink software. Using IC03 (Ichang papeda) as the outgroup, the population SNP phylogenetic tree was estimated through SNPhylo v1.0 [73] with the parameters "-B 1000 -A -b -r -P". The resulting fasta files were then input into RAxML, employing the GTRGAMMA model along with 1000 bootstrap replicates for the construction of the phylogenetic tree. The ancestral genetic component of each accession was inferred by ADMIXTURE [74] with fivefold cross-validation by increasing K (the number of clusters or groups) from 2 to 14 and plotted using Pophelper packages [75] in R. We selected $K = 12$ as the optimal value based on the following observations: (1) Beyond $K = 12$, the cross-validation (CV) error changed by only 0.008 or less, indicating minimal improvement in model fit with increased complexity. (2) When $K = 12$, the population structure has stabilized, and the basic citrus species *Citrus reticulata* (LSM), *Citrus maxima* (PU), *Citrus ichangensis* (IC), and *Citrus medica* (CI) can be clearly distinguished, reflecting the lowest model complexity (Fig. S4). (3) When $K = 13$ or larger, the population structure becomes more fragmented and does not provide additional biological information.

Fst and π estimates

Groups were formed based on phylogenetic and population structure analyses. The pairwise fixation index (*F*st) values of inter-group were calculated using VCFtools with a 50-kb window and a 10-kb sliding window. Nucleotide diversity (π) for each group was calculated using VCFtools with a 50-kb window and a 10-kb sliding window.

Kinship inference

The KING v2.2.7 [18] was employed for the inference of kinship relationships between pairs of species, using SNPs that had not undergone LD purging as input (auther recommended). The command "–related –build –degree 2" was executed, and the results were visualized through Cytoscape v3.1.0 [76].

Demographic analysis and Split time estimation by SMC + +

We inferred historical population sizes for each group with SMC + + v1.15.4 [77] based on a constant generation time of 8 year and a per-generation mutation rate of 2.2e-8 with parameters "smc + + estimate –timepoints 1 10,000,000 –em-iterations 20 –knots 8 –spline cubic". The "smc + + split" commond fitted two-population clean split models using marginal estimates produced by "smc + + estimate".

Introgression analyses

To detect gene flow events between groups, we employed TreeMix v1.13 (Pickrell and Pritchard, 2012) for gene flow analysis (TreeMix recommends removing SNPs with $R^2 > 0.1$). The parameters were set to "-k 500 -bootstrap -global -se -global -m -root PU", with m (migration edges) ranging from 0 to 12 and the analysis being repeated 12 times. The optimal number of migration edges was determined using the R package OptM [78]. Prior to the analysis, to minimize the impact of sample size, we randomly selected three samples from each group for analysis (WM-1 and JG had only two samples, which were all retained). Additionally, we used Dsuite v0.5 r52 (Malinsky, Matschiner, and Svardal, 2021) to perform the ABBA-BABA test (*D* statistic) and *fb* statistic, with IC as the outgroup, to further confirm gene flow events.

Chloroplast genome assembly

Chloroplast genomes were assembled using GetOrganelle v1.7.7.0 [79], based on Illumina reads. The complete assembled genomes were aligned using MAFFT v7.453 [80], and a phylogenetic tree was constructed through IQ-TREE v1.6.12 [81].

Inference of selective sweeps

The study of selective sweep patterns associated with artificial selection was based on the cross-population composite likelihood ratio (XP-CLR) and the population fixation index (*F*st). Two group comparisons were utilized to infer signals of artificial selection, namely, NF-2 vs. WM-2 and SM vs. NF-2. To mitigate the impact of sample size differences, in the NF-2 vs. WM-2 comparison, five samples each from NF-2 and WM-2 were used, while in the SM vs. NF-2 comparison, 12 and 14 samples of SM and NF-2, respectively, were employed. The pairwise fixation index (*F*st) values between groups were calculated using VCFtools with a 50-kb window and a 10-kb sliding window, retaining the top 10% of regions. Similarly, potential selective sweeps were screened using XP-CLR (python version, https://github.com/hardingnj/xpclr) with the same window parameters, considering the top 10% of the XP-CLR score regions as candidate signals for selective sweeps. Regions exhibiting both *F*st values and XP-CLR scores in the top 10% were selected as the final candidate regions for selective sweeps.

Haplotype network analysis of candidate selective gene

Haplotype networks were constructed using SNPs within the fragment spanning each gene and its 2 kb upstream and downstream regions. The genotypes of each gene were processed generate haplotypes and transformed to PHYLIP formats by vcf2phylip (https://github.com/edgardomortiz/vcf2phylip). Then, the haplotype network was constructed using fastHaN [82] by templeton-crandall-sing (TCS) method and visualized using tcsBU [83].

RNA-seq and transcriptome analysis

Fifteen samples of the three fruit tissues across 5 developmental stages of Nanfengmiju tangerine “Guiyuehong” (GYH) and 4 samples of fruit SEMs across 4 developmental stages of Satsuma mandarin “Guoqing No. 1” (GQ1) were harvested, and three biological replicates were harvested for each sample. A total of 57 transcriptome profiles were obtained by RNA-seq using the Illumina NovaSeq 6000 platform at Novogene (Beijing, China). Then, the clean transcriptomic reads were mapped to the monoploid using HISAT2 v2.2.1 [84] and counted by featureCounts [85]. The transcripts per million (TPM) values were calculated by R.

ASE gene identification

The non-redundant CDSs of Haplotype 1 and Haplotype 2 were aligned against the non-redundant CDS of the monoploid using BLAST +. The best matches located on the same pseudochromosome were extracted. Subsequently, genes of Haplotype 1 and Haplotype 2 with consistent alignment results were identified as alleles. A metagenome was constructed by merging the assembly and annotation of Haplotype 1 and Haplotype 2. The clean reads from 45 GYH transcriptomes were aligned to the metagenome using HISAT2. Gene counting was performed with featureCounts, and genes with counts greater than 10 in at least 15 samples were retained. Subsequently, we compared the expression of alleles from Hap1 and Hap2 within the same sample using DESeq2 (Love, Huber, and Anders, 2014). ASEGs were determined if the log fold change in the count between two alleles was greater than 2 with an adjusted p value < 0.05. Then, if the allele has a higher expression in Hap1, the dominant alleles of the ASEGs are defined to be from Hap1, and vice versa. The remaining genes were defined as genes with allele-equivalent expression (AEEGs).

Gene cloning, vector construction, and transiently transformation

For overexpression, the full-length CDS of *CrkINVA* and *CrkSWEET16* were amplified from Nanfengmiju tangerine pulp cDNA using Phanta Super-Fidelity DNA Polymerase (Vazyme). The primers are listed in Supplemental Table 20. The PCR products were first cloned into p-TOPO (Aidlab, China) vectors, and then recombined into pK7WG2D vector through the gateway BP (11,789,100, Thermo Fisher, CA, USA) and LR (11,791,020, Thermo Fisher, CA, USA) reaction. The constructed vectors were transformed into ripening fruit of Nanfengmiju tangerine through *Agrobacterium tumefaciens*-mediated (GV3101) genetic transformation as described previously [86]. After injected 7 days, we picked the injected fruits, the fruits were cut off at 1 cm^3 near the injection site and stored in a − 80 ℃ refrigerator for subsequent gene expression analysis and sugar content measurements.

For VIGS-mediated gene silencing, 220 bp CDS fragments of *CrkGAI* were cloned into TRV2 vector (Tobacco Rattle Virus-based 2). The TRV2, TRV1, and the fusion constructs (TRV2-*GAI*) were transformed into A. *tumefaciens* strain GV3101, respectively. The suspension strains mixed with TRV1 and TRV2 or TRV2-*GAI* (1:1) were injected into the Hongkong kumquat (F. hindsii) as described previously [87]. Notably, prior to injection, the longitudinal and transverse diameters of each fruit were measured to ensure they were between 8 and 9 mm, indicating that the fruits were nearing the fruit expansion period. After injected 7 days, we picked a few injected fruits for gene expression analysis. After 21 days, a sufficient number of fruits were collected to assess fruit size and for photographic documentation.

RNA extraction and gene expression analysis

Total RNA was extracted from Nanfengmiju tangerine and Daoxian wild mandarin pulp, and Hongkong kumquat fruit using a FastPure Universal Plant Total RNA Isolation Kit (RC411; Vazyme Biotech) as described previously [88]. RNA quality was evaluated using Denovix 2017 (Bio-SUN). The mRNA was reverse transcribed using the HiScriptIIQ

RT SuperMix for qPCR (+gDNA wiper) kit. *CsACTIN* and *CrkACTIN* were used as the housekeeping gene. For gene expression analysis, qRT-PCR was carried out on the QuantStudio 6 Flex real-time PCR system (Applied Biosystems, USA) with 384-well plates. All data were analyzed using the $2^{-\Delta\Delta Ct}$ analysis method as described previously [89]. All primers are listed in Additional file 2: Table S21.

Measurement of sugar content

The sucrose, glucose, and fructose contents were measured according to Yu et al. [90] with minor modification. Briefly, 0.1 g freeze-dried sample or 1 g fresh sample were extracted in 5 mL 80% (v/v) ice-cold methanol. After holding in a 75 °C water bath for 30 min and ultrasonically extracting at ice-cold water for 90 min, the samples were centrifuged at 4000 rpm for 10 min. The supernatant was pooled and mixed with an internal standard (methyl-a-D-glucopyranoside, Sigma, USA). The mixture was dried using Speed Vac (Eppendorf, Hamburg, Germany) and derivatized using hydroxylamine hydro-chloride: hexamethyldisilazane (HMDS) and trimethylchlorosilane (TMCS). A 0.5-μL aliquot was analyzed using GC on an Agilent 6890N system (Santa Clara, CA, USA) equipped with a flame ionization detector. The analysis employed a capillary column (HP-5, 5%-phenyl-methyl polysiloxane, 30 m × 320 μm i.d. × 0.25 μm). Nitrogen was used as the carrier gas at a flow rate of 45 mL/min, with hydrogen and air flow rates set at 40 mL/min and 450 mL/min, respectively. Sugar identification was achieved by comparing retention times with standard compounds supplied by Sigma (St. Louis, MO, USA).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03535-4.

Additional file 1: Fig S1-S14. Fig. S1. The NFMJ_120y fruit phenotype and heat maps of Mono, Hap1 and Hap2 showing the distribution of Hi-C interaction signals in a 500 kb resolution. a. The NF-1and NF-2 fruit phenotype.: Genome-wide Hi-C interaction heat map.The juicerbox was used to correct the inversion at Chr4B: 11,320,32615,048,651, but the Hi-C signal at both ends of the chromatin was weakened after correction, indicating that the position does not need to be corrected, and the inversion really exists. High densities of interactions are indicated in red. Fig. S2. KEGG pathway analysis of expanding and contracting gene families. Fig. S3. Nucleotide diversity of different citrus species SOO: Sour orange, SWO: Sweet orange, LSM: Loose-skin mandarin, PU: Pummelo, CI: Citron, IC: Ichang papeda. Fig. S4. The population structure and ancestry composition estimated by Admixture in mandarin population. a Cross-validation error plot for the unsupervised ADMIXTURE analysis. b. ADMIXTURE clustering of core accessions from*K*=2 to 14. c. Principal component analysiswas performed after 12 groups were divided. Fig. S5. Nucleotide diversityof 16 groups. Fig. S6. Distribution of LSMs in southern China. YZ: Yangtze River group; PR: Pearl River group; PK: Ponkan; WM: Wild mandarin. Fig. S7. Effective population size and split time beteewn MU-2 and ST-1 inferred by SMC++. Fig. S8. Analysis of introgression in different citrus groups.The mean and standard deviationacross 12 iterations for the composite likelihood Land proportion of variance explained.The second-order rate of changeacross values of m, the optimal m number shown by the red circle.The residual fit plots from the maximum likelihood tree in Supplementary Figure11D.Branch introgressionstatistic. Values of the branch introgressionstatistic represent the potential signal between groups.Phylogenetic tree constructed by chloroplast genomes. Fig. S9. Fruit vertical and transverse diameters of Nanfengmiju tangerine and Satsuma mandarin. Fig. S10. The candidate selective gene of NF-2 vs. WM-2. Signals of artificial selection in the candidate gene of NF-2 vs. WM-2. Purple and green solid lines indicate π value statistics in WM-2 and NF-2, respectively. The blue dashed line indicates the *Fst* value between NF-2 and WM-2, while the red dashed line is the threshold of the top 10% *Fst*.The haplotype network analysis of *CrkINVA* and *CrkSWEET16*. Each circle represents one haplotype and the circle area is proportional to the frequency of each haplotype. Different colors represent different populations. Small uncoloured circles represent non-observed haplotypes; each line connecting haplotypes represents a single mutational change.Expression patterns of *CrkINVA* and *CrkSWEET16* of NF-2 vs. WM-2 in the JS during development of GYH. FF: the fruit-falling period, EEX, the Early expansion period. Relative expression of the *CrkINVA* in OE-*CrkINVA* lines. The asterisk indicates significance two-sided Student's t-test, The *P* value is 0.000004.Fruit soluble sugar content of OE-*CrkINVA* lines. The asterisk indicates significance two-sided Student's t-test, The *P* value for sucrose is 0.0135. Fig. S11. The haplotype network analysis and transient silence of *CrkGAI* by VIGS in Hongkong kumquat fruits.The haplotype network analysis of *CrkGAI*. Each circle represents one haplotype and the circle area is proportional to the frequency of each haplotype. Different colors represent different populations. Small uncolored circles represent non-observed haplotypes; each line connecting haplotypes represents a single mutational change.The fruits phenotype of TRV and TRV-*CrkGAI*.Relative expression of the *CrkGAI* in TRV and TRV-*CrkGAI*. The asterisk indicates significance with two-sided Student's t-test, the *P* values is 0.0014.The fruit weight of TRV and TRV-*CrkGAI*. The asterisk indicates significance with two-sided Student's t-test, the *P* value is 0.0088. Fig. S12. Signals of artificial selection in the candidate genes of SM vs. NF-2. Purple and green solid lines indicate π value statistics in SM and NF-2, respectively. The blue dashed line indicates the *F*st value between SM and NF-2, while the red dashed line is the threshold of the top 10% *F*st. Fig. S13. Allele-specific gene expression during GYH fruit development.UpSet plot showing overlap of ASEGs in different stages and tissues. Only top 50 categories are shown. H1, the dominant allele was from Hap1. H2, the dominant allele was from Hap2. Sample information is shown in Supplemental Table 7.The heatmap shows allele-specific expression pattern in 15 samples. The number of dominant alleles of ASEGs. Fig. S14. GO enrichment analysis of ASE and AEE genes The results of the GO termenrichment analysis visualized by the 'treemap' view of REVIGO. Each rectangle is a single cluster representative. The representatives are joined into 'superclusters' of loosely related terms, visualized with different colors. The size of the rectangles is adjusted to reflect the *P*-value.

Additional file 2: Table S1-S21. Table S1. Genome survey analysis of Nanfengmiju tangerine. Table S2 BUSCOs analysis of monoploid and haplotypes. Table S3. The result of NFMJ-120y repeat sequence annotation. Table S4. The result of aligning Hap2 to the Hap1 using NUCmer. Table S5. Statistics of variation between two haplotypes. Table S6. Annotation of sequence variations by snpEff v4.3. Table S7. Species used in phylogenetic analysis. Table S8. Summary of gene family clustering. Table S9. Statistics of genome sequence data of Citrus accessions used in this study. Group, grouping by population structure and phylogeny. Table S10. The summary table of *Fst* and Kinship Coefficient between different mandarin groups. Table S11. Kinship relationships estimated for different citrus cultivars and wild citrus species. Table S12. The result of ABBA-BABA test by Dsuite. Table S13. Statistics of different mandarins fruit size. Table S14. Population divergence across the genome.. Table S15. Population divergence across the genome.. Table S16. Statistical results of RNA-Seq data alignment to the metagenome from 3 tissues at 5 different developmental stages of the GYH fruit. DAF, days after full-bloom. Table S17. Gene expression levelswith ASE in GYH fruit. ASEor DEG were defined under the cutoff of adjused *P* value < 0.05 and log2 fold change≥ 2 or ≤ -2. Table S18. Gene expression levelswith AEE in GYH fruit. Table S19. GO enrichment analysis of ASE and AEE genes. Table S20.The counts of number of variants affecting each ASEGs and AEEGs. Table S21 List of primer sequences used in this study.

Acknowledgements
We thank Ningxian Li for his help in collecting samples. We thank Qiang Xu and Yongzhong Liu from the National Key Laboratory for Germplasm Innovation and Utilization of Horticultural Crops, Huazhong Agricultural University for providing Daoxian wild mandarin and Nanfengmiju tangerine. The computations in this paper were run on the bioinformatics computing platform of the National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University.

Peer review information
Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contribution
J.W. and H.Y. conceived the project and designed the research. M.Y. performed data collection and bioinformatics analysis. X.S. performed the transgenic experiments. M.Y. and X.S. performed measurement of sugar content. M.Y., C.H., L.L., L.Z., Z.C., Y.X., X.L., M.L., J.L., H.W. and C.C. carried out samples collection and conducted the experiments. M.Y., J.W., H.Y. and X.W. wrote the paper. All authors read and approved the final manuscript.

Funding
This research was supported by the National Key Research and Development Project (2024YFD1200501), the China Agriculture Research System-Citrus (CARS-26), the 'Unveiling and Leading' Project of Fuzhou City (XMBH00091), the National Natural Science Foundation of China (NSFC) (32172524), the Fruit Tree Project of Research Institutes and Enterprises Consortium of Chongqing (KYLX20231000025), and the Hubei Province Technology Innovation Plan Project (2024BBB085).

Data availability
The raw resequencing reads, Hi-C data, and RNA-seq data, as well as the assemblies, have been deposited at the China National Genomics Data Center with Bioproject ID PRJCA025346 [91]. The published resequencing data information is in Additional file 2: Table S9, including NCBI BioProject PRJDB5882 [92], PRJNA225965 [93], PRJNA320985 [94], PRJNA414519 [95], PRJNA482734 [96], PRJNA687608 [97], PRJNA745525 [98], PRJNA746063 [99], PRJNA762220 [100]. The assembled genomes and annotations are also available at Figshare (https://doi.org/10.6084/m9.figshare.28504274) [101]. No other scripts and software were used other than those mentioned in the "Methods" section.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 30 April 2024 Accepted: 10 March 2025
Published: 17 March 2025

References

1. Wang L, He F, Huang Y, He J, Yang S, Zeng J, Deng C, Jiang X, Fang Y, Wen S, et al. Genome of wild mandarin and domestication history of mandarin. Mol Plant. 2018;11:1024–37.
2. Deng X. Citrus varieties in China. 2nd ed. Beijing: China Agriculture Press; 2023.
3. Huang Y, He J, Xu Y, Zheng W, Wang S, Chen P, Zeng B, Yang S, Jiang X, Liu Z, et al. Pangenome analysis provides insight into the evolution of the orange subfamily and a key gene for citric acid accumulation in citrus fruits. Nat Genet. 1964;2023:55.
4. Liu W, Martinon-Torres M, Cai YJ, Xing S, Tong HW, Pei SW, Sier MJ, Wu XH, Edwards RL, Cheng H, et al. The earliest unequivocally modern humans in southern China. Nature. 2015;526:696–9.
5. Wu GA, Sugimoto C, Kinjo H, Azama C, Mitsube F, Talon M, Gmitter FG Jr, Rokhsar DS. Diversification of mandarin citrus by hybrid speciation and apomixis. Nat Commun. 2021;12:4377.
6. Yang Z, Li G, Tieman D, Zhu G. Genomics approaches to domestication studies of horticultural crops. Horticult Plant J. 2019;5:240–6.
7. Guo S, Zhao S, Sun H, Wang X, Wu S, Lin T, Ren Y, Gao L, Deng Y, Zhang J, et al. Resequencing of 414 cultivated and wild watermelon accessions identifies selection for fruit quality traits. Nat Gen. 2019;51:1616-+.
8. Yu Y, Guan J, Xu Y, Ren F, Zhang Z, Yan J, Fu J, Guo J, Shen Z, Zhao J, et al. Population-scale peach genome analyses unravel selection patterns and biochemical basis underlying fruit flavor. Nat Commun. 2021;12:3604.
9. Han X, Zhang Y, Zhang Q, Ma N, Liu X, Tao W, Lou Z, Zhong C, Deng XW, Li D, He H. Two haplotype-resolved, gap-free genome assemblies for Actinidia latifolia and Actinidia chinensis shed light on the regulatory mechanisms of vitamin C and sucrose metabolism in kiwifruit. Mol Plant. 2023;16:452–70.
10. Wang R, Li X, Sun M, Xue C, Korban SS, Wu J. Genomic insights into domestication and genetic improvement of fruit crops. Plant Physiol. 2023;192:2604–27.
11. Duan N, Bai Y, Sun H, Wang N, Ma Y, Li M, Wang X, Jiao C, Legall N, Mao L, et al. Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. Nat Commun. 2017;8:249.
12. Wu GA, Terol J, Ibanez V, Lopez-Garcia A, Perez-Roman E, Borreda C, Domingo C, Tadeo FR, Carbonell-Caballero J, Alonso R, et al. Genomics of the origin and evolution of citrus. Nature. 2018;554:311–6.
13. Wang L, Huang Y, Liu ZA, He JX, Jiang XL, He F, Lu ZH, Yang SZ, Chen P, Yu HW, et al. Somatic variations led to the selection of acidic and acidless orange cultivars. Nat Plants. 2021;7:954-+.
14. Lu Z, Huang Y, Mao S, Wu F, Liu Y, Mao X, Adhikari PB, Xu Y, Wang L, Zuo H, et al. The high-quality genome of pummelo provides insights into the tissue-specific regulation of citric acid and anthocyanin during domestication. Hortic Res. 2022;9:uhac175.
15. Sultana T, Zamborlini A, Cristofari G, Lesage P. Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet. 2017;18:292–308.
16. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvak Z, Levin HL, Macfarlan TS, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19:199.
17. Zhou Q, Tang D, Huang W, Yang ZM, Zhang Y, Hamilton JP, Visser RGF, Bachem CWB, Buell CR, Zhang ZH, et al. Haplotype-resolved genome analyses of a heterozygous diploid potato. Nat Gen. 2020;52:1018-+.
18. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26:2867–73.
19. Ao H, Rohling EJ, Stringer C, Roberts AP, Dekkers MJ, Dupont-Nivet G, Yu J, Liu Q, Zhang P, Liu Z, et al. Two-stage mid-Brunhes climate transition and mid-Pleistocene human diversification. Earth-Sci Rev. 2020;210:103354.
20. Corrick EC, Drysdale RN, Hellstrom JC, Capron E, Rasmussen SO, Zhang X, Fleitmann D, Couchoud I, Wolff E. Synchronous timing of abrupt climate changes during the last glacial period. Science. 2020;369:963-+.
21. Lu XX, Yu CH. The origin and taxonomic position of jiaogan (Citrus tankan Tankan). Acta Hortic Sin. 1995;22:105–109. (Chinese).
22. Peng JR, Carol P, Richards DE, King KE, Cowling RJ, Murphy GP, Harberd NP. The Arabidopsis GAI gene defines a signaling pathway that negatively regulates gibberellin responses. Genes Dev. 1997;11:3194–205.
23. Liu Q, Wu K, Harberd NP, Fu XD. Green revolution DELLAs: from translational reinitiation to future sustainable agriculture. Mol Plant. 2021;14:547–9.
24. Sun X, Jiao C, Schwaninger H, Chao CT, Ma Y, Duan N, Khan A, Ban S, Xu K, Cheng L, et al. Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. Nat Genet. 2020;52:1423–32.
25. Zhang J, Zhang X, Tang H, Zhang Q, Hua X, Ma X, Zhu F, Jones T, Zhu X, Bowers J, et al. Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. Nat Genet. 2018;50:1565–73.
26. Hu G, Feng J, Xiang X, Wang J, Salojarvi J, Liu C, Wu Z, Zhang J, Liang X, Jiang Z, et al. Two divergent haplotypes from a highly heterozygous lychee genome suggest independent domestication events for early and late-maturing cultivars. Nat Genet. 2022;54:73–83.
27. Cleary S, Seoighe C. Perspectives on allele-specific expression. Annu Rev Biomed Data Sci. 2021;4:101–22.
28. Lawson DJ, van Dorp L, Falush D. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. Nat Commun. 2018;9:3258.
29. Feng G, Wu J, Xu Y, Lu L, Yi H. High-spatiotemporal-resolution transcriptomes provide insights into fruit development and ripening in Citrus sinensis. Plant Biotechnol J. 2021;19:1337.
30. Klemens PA, Patzke K, Deitmer J, Spinner L, Le Hir R, Bellini C, Bedu M, Chardon F, Krapp A, Neuhaus HE. Overexpression of the vacuolar sugar carrier AtSWEET16 modifies germination, growth, and stress tolerance in Arabidopsis. Plant Physiol. 2013;163:1338–52.

31. Xiang L, Le Roy K, Bolouri-Moghaddam MR, Vanhaecke M, Lammens W, Rolland F, Van den Ende W. Exploring the neutral invertase-oxidative stress defence connection in Arabidopsis thaliana. J Exp Bot. 2011;62:3849–62.
32. Fenn MA, Giovannoni JJ. Phytohormones in fruit development and maturation. Plant J. 2021;105:446–58.
33. He H, Yamamuro C. Interplays between auxin and GA signaling coordinate early fruit development. Hortic Res. 2022;9:uhab078.
34. Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764–70.
35. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020;11:1432.
36. Cheng HY, Concepcion GT, Feng XW, Zhang HW, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods. 2021;18:170-+.
37. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016;3:95–8.
38. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356:92–5.
39. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 2016;3:99–101.
40. Marcais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. Plos Comput Biol. 2018;14:e1005944.
41. Seppey M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. Methods Mol Biol. 2019;1962:227–45.
42. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34:i142–50.
43. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008;9: 18.
44. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35:W265-268.
45. Han Y, Wessler SR. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. 2010;38: e199.
46. Ou SJ, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 2018;176:1410–22.
47. Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, Lei J, Achawanantakun R, Jiao D, Lawrence CJ, et al. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 2014;164:513–24.
48. Yan H, Bombarely A, Li S. DeepTE: a computational method for de novo classification of transposons with convolutional neural network. Bioinformatics. 2020;36:4269–75.
49. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.
50. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18:188–96.
51. 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.
52. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5: 59.
53. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng QD, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644-U130.
54. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34:W435-439.
55. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31:5654–66.
56. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.
57. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421.
58. Goel M, Sun H, Jiao WB, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 2019;20:277.
59. Cingolani P, Platts A, le Wang L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6:80–92.
60. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238.
61. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5: 113.
62. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17:540–52.
63. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
64. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91.

65. Xie S, Manchester SR, Liu K, Wang Y, Sun B. Citrus linczangensis sp. n., a Leaf Fossil of Rutaceae from the Late Miocene of Yunnan, China. Int J Plant Sci. 2013;174:1201–7.
66. Mendes FK, Vanderpool D, Fulton B, Hahn MW. CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics. 2021;36:5516–8.
67. Bu D, Luo H, Huo P, Wang Z, Zhang S, He Z, Wu Y, Zhao L, Liu J, Guo J, et al. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. Nucleic Acids Res. 2021;49:W317–25.
68. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
69. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. Genome Project data processing S: The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
70. Auwera GAVD, O'Connor BD. Genomics in the cloud: using docker, GATK, and WDL in Terra. California: O'Reilly Media; 2020.
71. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
72. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
73. Lee TH, Guo H, Wang X, Kim C, Paterson AH. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics. 2014;15: 162.
74. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics. 2011;12: 246.
75. Francis RM. pophelper: an R package and web app to analyse and visualize population structure. Mol Ecol Resour. 2017;17:27–32.
76. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
77. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet. 2017;49:303–9.
78. Fitak RR. OptM: estimating the optimal number of migration edges on population trees using Treemix. Biol Methods Protoc. 2021;6:bpab017.
79. Jin JJ, Yu WB, Yang JB, Song Y, dePamphilis CW, Yi TS, Li DZ. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 2020;21:241.
80. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
81. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.
82. Chi L, Zhang X, Xue Y, Chen H. fastHaN: a fast and scalable program for constructing haplotype network for large-sample sequences. Mol Ecol Resour. 2023;00:1–5.
83. Murias dos Santos A, Cabezas MP, Tavares AI, Xavier R, Branco M. tcsBU: a tool to extend TCS network layout and visualization. Bioinformatics. 2016;32:627–8.
84. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.
85. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.
86. Zhu CQ, Zheng XJ, Huang Y, Ye JL, Chen P, Zhang CL, Zhao F, Xie ZZ, Zhang SQ, Wang N, et al. Genome sequencing and CRISPR/Cas9 gene editing of an early flowering Mini-Citrus (*Fortunella hindsii*). Plant Biotechnol J. 2019;17:2199–210.
87. Zhang Y, Zhu J, Khan M, Wang Y, Xiao W, Fang T, Qu J, Xiao P, Li C, Liu JH. Transcription factors ABF4 and ABR1 synergistically regulate amylase-mediated starch catabolism in drought tolerance. Plant Physiol. 2023;191:591–609.
88. Liu Q, Xu J, Liu Y, Zhao X, Deng X, Guo L, Gu J. A novel bud mutation that confers abnormal patterns of lycopene accumulation in sweet orange fruit (Citrus sinensis L. Osbeck). J Exp Bot. 2007;58:4161–71.
89. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, et al. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin Chem. 2009;55:611–22.
90. Yu K, Xu Q, Da X, Guo F, Ding Y, Deng X. Transcriptome changes during fruit development and ripening of sweet orange (Citrus sinensis). BMC Genomics. 2012;13: 10.
91. Yin MQ, Song XC, He C, Li XY, Li MY, Li JB, Wu H, Chen CW, Zhang L, Cai ZM, Lu LQ, Xu YH, Wang X, Yi HL, Wu JX. The haplotype-resolved genome assembly of an ancient citrus variety provides insights into the domestication history and fruit trait formation of loose-skin mandarins. Datasets. Genome sequence archive. 2025. https://bigd.big.ac.cn/gsa/browse/CRA015998.
92. Shimizu T, Tanizawa Y, Mochizuki T, Nagasaki H, Yoshioka T, Toyoda A, Fujiyama A, Kaminuma E, Nakamura Y. Draft sequencing of the heterozygous diploid genome of satsuma (Marc.) using a hybrid assembly approach. Datasets. Genome sequence archive. 2017. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJDB5882.
93. Wu GA, Terol J, Ibanez V, Lopez-Garcia A, Perez-Roman E, Borreda C, Domingo C, Tadeo FR, Carbonell-Caballero J, Alonso R, et al. Genomics of the origin and evolution of Citrus. Datasets. Genome sequence archive. 2013. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA225965.
94. Wang X, Xu Y, Zhang S, Cao L, Huang Y, Cheng J, Wu G, Tian S, Chen C, Liu Y, et al. Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. 2016. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA320985.
95. Wu GA, Terol J, Ibanez V, Lopez-Garcia A, Perez-Roman E, Borreda C, Domingo C, Tadeo FR, Carbonell-Caballero J, Alonso R, et al. Genomics of the origin and evolution of Citrus. Datasets. Genome sequence archive. 2017. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA414519.

96. Li YP. Datasets. Genome sequence archive. 2018. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA482734.

97. Feng GZ, Ai X, Yi HL, Guo WW, Wu JX. Genomic and transcriptomic analyses of Citrus sinensis varieties provide insights into Valencia orange fruit mastication trait formation. Datasets. Genome sequence archive. 2020. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA687608.

98. Subtropical Horticulture Research Institute. Datasets. Genome sequence archive. 2021. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA745525.

99. Subtropical Horticulture Research Institute. Datasets. Genome sequence archive. 2021. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA746063.

100. Perez-Roman E, Borredá C, Usach ALG, Talon M. Single-nucleotide mosaicism in citrus: estimations of somatic mutation rates and total number of variants. Datasets. Genome sequence archive. 2021. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA762220.

101. Yin MQ, Song XC, He C, Li XY, Li MY, Li JB, Wu H, Chen CW, Zhang L, Cai ZM, Lu LQ, Xu YH, Wang X, Yi HL, Wu JX. The haplotype-resolved genome assembly of an ancient citrus variety provides insights into the domestication history and fruit trait formation of loose-skin mandarins. Datasets. Genome sequence archive. 2025. https://doi.org/10.6084/m9.figshare.28504274.

Publisher's Note