Horticulture Research

Horticulture Research, 2022, **9**: uhac012
https://doi.org/10.1093/hr/uhac012

Article

Chromosome-scale assembly and population diversity analyses provide insights into the evolution of *Sapindus mukorossi*

Ting Xue[1,2], Duo Chen[2], Tianyu Zhang[3], Youqiang Chen[2], Huihua Fan[4], Yunpeng Huang[4], Quanlin Zhong[1] and Baoyin Li[1,*]

[1]Fujian Provincial Key Laboratory for Plant Eco-physiology, State Key Laboratory for Subtropical Mountain Ecology of the Ministry of Science and Technology and Fujian Province, College of Geographical Sciences, Fujian Normal University, Fuzhou 350007, China
[2]College of Life Sciences, Fujian Normal University, Fuzhou 350117, China
[3]Shunchang County Forestry Science and Technology Center of Fujian Province, Forestry Bureau of Shunchang, Shunchang 353200, China
[4]Research Institute of Forestry, Fujian Research Institute of Forestry, Fuzhou 350000, China
*Corresponding author: E-mail: liby@fjnu.edu.cn

Abstract

***Sapindus mukorossi* is an environmentally friendly plant and renewable energy source whose fruit has been widely used for biomedicine, biodiesel, and biological chemicals due to its richness in saponin and oil contents. Here, we report the first chromosome-scale genome assembly of *S. mukorossi* (covering ~391 Mb with a scaffold N50 of 24.66 Mb) and characterize its genetic architecture and evolution by resequencing 104 *S. mukorossi* accessions. Population genetic analyses showed that genetic diversity in the southwestern distribution area was relatively higher than that in the northeastern distribution area. Gene flow events indicated that southwest species may be the donor population for the distribution areas in China. Genome-wide selective sweep analysis showed that a large number of genes are involved in defense responses, growth and development, including *SmRPS2*, *SmRPS4*, *SmRPS7*, *SmNAC2*, *SmNAC23*, *SmNAC102*, *SmWRKY6*, *SmWRKY26*, and *SmWRKY33*. We also identified several candidate genes controlling six agronomic traits by genome-wide association studies, including *SmPCBP2*, *SmbHLH1*, *SmCSLD1*, *SmPP2C*, *SmLRR-RKs*, and *SmAHP*. Our study not only provides a rich genomic resource for further basic research on Sapindaceae woody trees but also identifies several economically significant genes for genomics-enabled improvements in molecular breeding.**

Introduction

Sapindus mukorossi (Fig. 1), also known as soapberry tree, is a deciduous tree of the Sapindaceae family and is widely distributed in Southeast Asia, Japan, and southern and southwestern China [1]. Li Shizhen wrote 'By using *S. mukorossi*, shampoo can go to head wind, and washing can whiten' in *The Compendium of Materia Medica*, highlighting that *S. mukorossi* is a natural excellent detergent with good foaming and decontamination abilities [2]. As recorded in *Pu Ji Fang*, '*S. mukorossi* can also treat toothache, sore throat, acute gastroenteritis and so on', showing that *S. mukorossi* is a very important Chinese herbal medicine with multiple biomedical functions [3]. The fruit peel of *S. mukorossi* is rich in saponins (up to 10.76%) with physiological activities, such as antifungal, antidandruff and anti-itching activities [4, 5]. The seeds of *S. mukorossi* are rich in oil (up to 42.70%) with a high proportion of unsaturated fatty acids (up to 86.63%), rendering them ideal materials for developing biodiesel and coping with the energy crisis [6, 7]. With its richness in saponins in the peel and oil content in kernels, *S. mukorossi* has become an important industrial tree species that can provide multifunctional raw material for producing biological chemicals, biodiesel, and biomedicine [8, 9]. Additionally, with a beautiful crown shape, strong adaptability, and resistance to sulfur dioxide and industrial dust pollution, *S. mukorossi* is regarded as a new type of low-carbon, environmentally friendly renewable energy source for ecological restoration of barren mountains and rural greening [10, 11].

Based on a large number of studies, research on *S. mukorossi* mainly focuses on genetic and ecological characteristics, seed germination, landscaping, active ingredient extraction, etc. [12–16]. For example, Liu *et al.* comprehensively evaluated superior germplasms from natural germplasm resources of three species and one variety of *Sapindus* in China and Vietnam and divided them into three categories, each with 10 selected superior germplasm accessions for oil and saponins [17]. Phenotypic and molecular markers (inter simple

Received: 27 August 2021; **Accepted:** 3 December 2021; **Published:** 17 February 2022; **Corrected and Typeset:** 17 February 2022
© The Author(s) 2022. Published by Oxford University Press on behalf of Nanjing Agricultural University. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited..

Downloaded from https://academic.oup.com/hr/article/doi/10.1093/hr/uhac012/6529164 by guest on 30 July 2026

Figure 1. Representative photographs of *S. mukorossi*. **a** Mature *S. mukorossi* tree in its natural habitat. **b** Flowers. **c, d** Fruits. Bar =1 cm.

sequence repeat, simple sequence repeat and random amplified polymorphic DNA) were used to analyze the diversity of natural germplasm resources in India and China in different geographical regions. The results showed that the variation between individuals within populations is higher than that between populations [18–22]. Fan *et al*. investigated 122 accessions from nine provinces by measuring the main economic traits (saponins, oil, age, height, altitude, etc.), which showed that various economic traits of trees of different origins have obvious differences, and the main economic traits of different trees from the same area are also different [11]. However, previous studies have focused on small samples, resulting in only partial representation of *S. mukorossi* based on fruit and seed traits, and no studies have explored the related omics and molecular breeding of *S. mukorossi*. The abundance of germplasm resources is a prerequisite for the selection of improved varieties, and the genetic diversity of germplasm resources is the material basis for biological evolution and breeding [23]. Population resequencing has become an important strategy for further crop domestication, improvement, and breeding, as recently reported in jujube [24], soybean [25], chickpea [26], and peach [27]. Population resequencing studies have been performed to examine the genetic diversity and genetic basis of evolution and adaptation in species such as *Salix brachista* [28] and tea [29]. In theory, different species have experienced reproductive or geographical isolation. Due to the long evolutionary history of *S. mukorossi*, after long-term natural selection various morphological variations and biological genetic diversity in geographical areas are objectively present in nature. Moreover, different *S. mukorossi* varieties are easily cultivated by hybridization, complicating accurate tracing of the origin of the offspring. Although previous studies have attempted to explore this issue, the lack of genome data and annotated gene sets for *S. mukorossi* has limited functional studies and evolutionary analyses.

To gain a better understanding of the origin and evolution of *S. mukorossi*, we used PacBio and Hi-C technologies to assemble a genome, providing a good resource for investigating genetic diversity and evolution. We also collected 104 *S. mukorossi* resources for resequencing from nine provinces of China covering nearly all *S. mukorossi* planting areas at the whole-genome level, including 47 accessions from Northeast China (Anhui Province, AH; Jiangxi Province, JX; Fujian Province, FJ; Zhejiang Province, ZJ), 38 accessions from

Downloaded from https://academic.oup.com/hr/article/doi/10.1093/hr/uhac012/6529164 by guest on 30 July 2026

South China (Hubei Province, HB; Hunan Province, HN; FJ, JX, Guangdong Province, GD; Guangxi Province, GX), and 18 accessions from Southwest China (Yunnan Province, YN) (Supplementary Table S1). Using these sequencing data, we demonstrated the evolutionary history of *S. mukorossi* and revealed the phylogenetic relationship between these populations by analyzing genomic characteristics, population structures, and genetic diversity. To identify potential candidate genes associated with fruit weight, peel-to-fruit ratio, saponin contents, seed-to-fruit ratio, kernel-to-fruit ratio, and oil contents, we also performed genome-wide association studies (GWAS) using resequencing data from 57 accessions (Supplementary Table S2). Our results provide a new resource for further molecular breeding and studies of *S. mukorossi* biology. However, genome-wide association analyses of economic traits (saponins, oil, etc.) remain to be conducted in the future.

Results

Sequencing and assembly of the *S. mukorossi* genome

Approximately 21 Gb (~47×) of Illumina paired-end sequences were generated and used to estimate genome size and heterozygosity using Jellyfish software, which indicated that the size of the *S. mukorossi* genome was 432.29 Mb with 1.52% heterozygosity (Supplementary Fig. S1; Supplementary Table S3). A total of 22.96 Gb of HiFi circular consensus sequencing (CCS) data were obtained, yielding a *de novo* contig-level assembly of 391.55 Mb with a GC content of 33.20% and a contig N50 of 2.88 Mb, which was close to the size estimated according to flow cytometry (Fig. 2a; Supplementary Fig. S2; Supplementary Table S4) and genome survey analyses (Supplementary Fig. S1). The assembled genome size was smaller than those of other closely related species in Sapindales, such as *Dimocarpus longan* (~471 Mb) and *Acer yangbiense* (~666 Mb) [30, 31]. Regarding the chromosome number, *S. mukorossi* is a diploid species, which has a karyotype of $2n = 2x = 28$ homologous pairs of chromosomes based on karyotyping analysis (Supplementary Fig. S3). To obtain a chromosome-scale assembly of *S. mukorossi*, a total of 40.96 Gb of clean sequences were generated by Hi-C sequencing. Using the contact matrix and the agglomerative hierarchical clustering method in LACHESIS, 90.87% (~352.74 Mb) of the assembled genome was successfully anchored into 14 pseudochromosomes (the length of the pseudochromosomes ranged from 15.76 to 44.80 Mb) with a scaffold N50 value of 24.66 Mb, which is 8.56-fold higher than that of the *de novo* assembled genome (a contig N50 of 2.88 Mb) (Fig. 2a; Supplementary Fig. S4; Supplementary Table S5). Expressed sequence tags (EST; ~94.09% of transcriptome data) and Benchmarking Universal Single-Copy Orthologs (BUSCO; ~83.80% of complete BUSCOs) analyses showed high accuracy and good completeness of the assembly (Supplementary Tables S6 and S7).

Gene prediction and annotation

A total of 31 853 protein-coding genes were identified in the *S. mukorossi* genome, which is higher than the numbers in other closely related species, such as *D. longan* (31 007 proteins) and *A. yangbiense* (28 320 proteins). The average gene length was 2978 bp and the average coding sequence (CDS) length was 1524 bp. Among the 31 853 predicted proteins, 27 509 (86.36%) were functionally annotated from five known databases: NR (27 335 genes, 85.82%), Swiss-Prot (23 766 genes, 74.61%), GO (Gene Ontology; 18 615 genes, 58.44%), KOG (Eukaryotic Orthologous Groups of Proteins; 14 452 genes, 45.37%), and KEGG (Kyoto Encyclopedia of Genes and Genomes; 11 025 genes, 34.61%) (Supplementary Fig. S5; Supplementary Table S8). We identified 2140 transcription factors (TFs) distributed in 35 families, representing 6.72% of genes in the *S. mukorossi* genome, which is ~1.33-fold higher than that in the *D. longan* genome (1611, 4.1%) (Supplementary Table S9). The identification of these TFs in the *S. mukorossi* genome will help elucidate major secondary metabolite biosynthesis in the future. We further identified a total length of 57.16 Mb of repetitive elements, representing 14.60% of the assembled genome (Supplementary Table S10), which is significantly lower than the amount reported in *D. longan* (52.87%, 261.88 Mb), indicating that the sizes of the *S. mukorossi* and *D. longan* genomes are different due to the number of repetitive elements that they contain. Various fruit tree genomes have been reported to contain relatively high percentages of repetitive elements, including *D. longan* (52.87%), kiwifruit (36%), pineapple (38.3%), *Vitis vinifera* (41.4%), *Ziziphus jujuba* (49.49%), *Citrus sinensis* (51.9%), pear (53.1%), and apple (67.4%) [30, 32–38]. We speculated that these large numbers of repetitive sequences are specific to fruit species, which may be an important reason for the edibility of their pulp. Among the identified repetitive elements, 8.91% (34 883 414 bp) were long terminal repeats (LTRs) and 0.36% (1 402 605 bp) were long interspersed nuclear elements (LINEs) (Supplementary Table S10). In addition, 4018 *C. sinensis* tRNAs, 88 miRNAs, 3035 rRNAs, 572 snRNAs, 2675 introns, and 1 sRNA were detected (Supplementary Table S11).

Genome evolution

Evolutionary analysis was performed using OrthoFinder (version 2.3.12) among 23 species representing five orders, including Myrtales, Rosales, Malpighiales, Rutales, and Sapindales. The phylogenetic tree showed that *S. mukorossi* had a closer relationship to *D. longan* and *A. yangbiense* and phylogenetically diverged from the common ancestor (Sapindales) (Fig. 2b). We identified 24 761 gene families and analyzed their evolution by comparing 21 plant species, and 852 expanded and 4580 contracted gene families were detected in the *S. mukorossi* genome. Interestingly, we found that the numbers of contracted and expanded gene families in *S. mukorossi* were 2.71- and 0.56-fold those in *D. longan* (1692 contracted and

Downloaded from https://academic.oup.com/hr/article/doi/10.1093/hr/uhac012/6529164 by guest on 30 July 2026

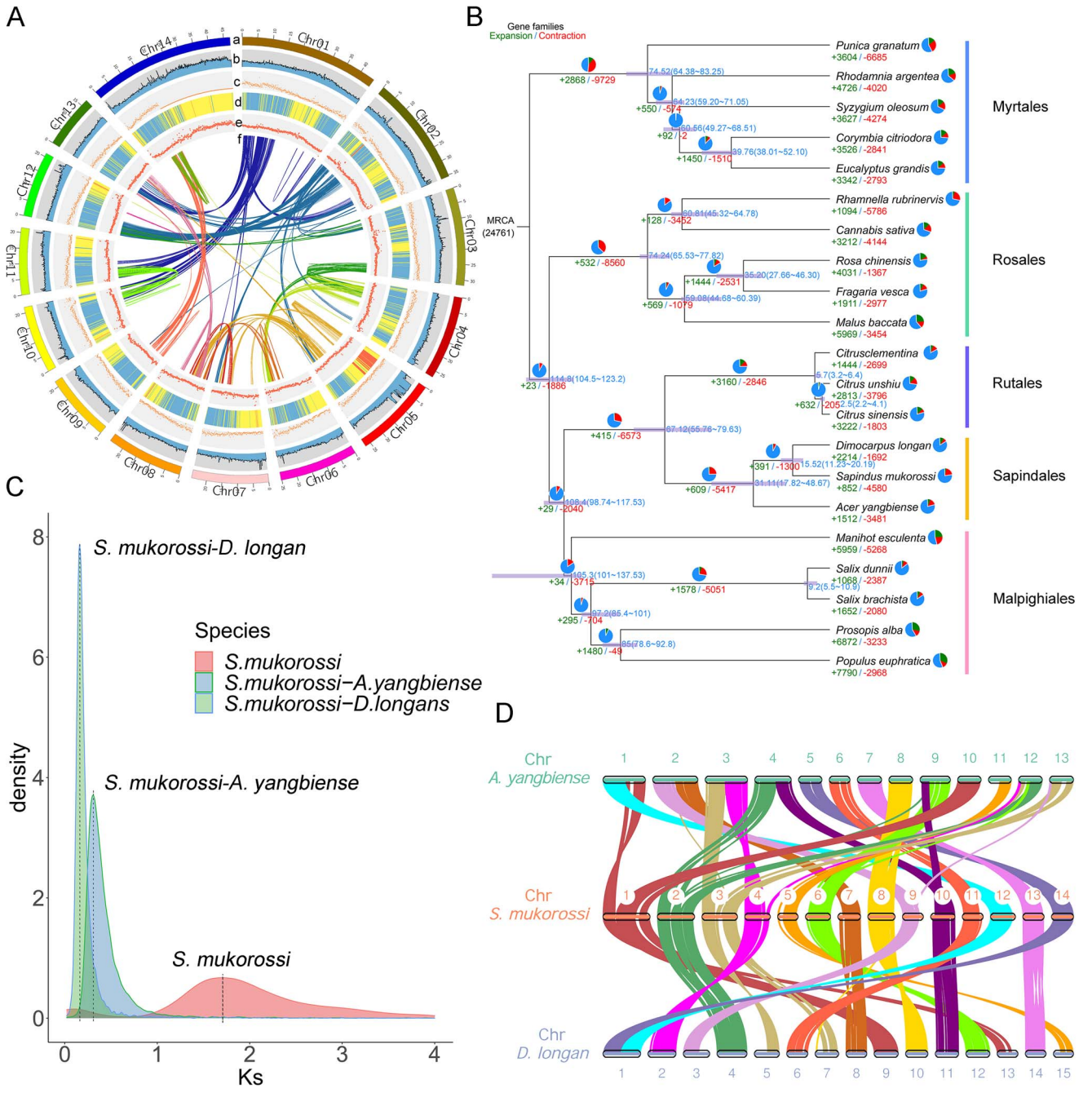

Figure 2. Characterization and evolution of the *S. mukorossi* genome. **a** Genome features across 14 chromosomes. (a) The 14 assembled chromosomes; (b) GC content; (c) gene density; (d) summary of gene expression in roots, stems, and leaves; (e) Transposable element density over the genome in non-overlapping windows; (f) summary of major interchromosomal syntenic relationships. Central colored lines represent syntenic blocks; block size = 3 kb. **b** Phylogenetic tree and evolution analysis of *S. mukorossi*. Red and green fonts indicate the numbers of gene families that have expanded and contracted among 21 species, respectively. Gene family expansions are indicated in green in the pie chart, while gene family contractions are indicated in red. Bars represent the 95% highest probability densities of the estimated divergence time. **c** Ks distribution of syntenic blocks between *S. mukorossi*, *D. longan*, and *A. yangbiense*. **d** Macrosynteny analysis between *S. mukorossi*, *D. longan*, and *A. yangbiense*. Each line connects a pair of orthologous genes.

852 expanded gene families), respectively. Enrichment analysis of contracted and expanded gene families was carried out with the following thresholds: $P < .05$ and false discovery rate < 0.05. GO and KEGG enrichment results showed that the expanded gene families of *S. mukorossi* were mostly enriched in cell, cell part, cellular process, energy metabolism, carbohydrate metabolism, cell growth, and lipid metabolism (Supplementary Tables S12 and –). These families included ATP-dependent Clp protease (*SmCLPP*, whz_003864), which plays an important role in the cell cycle and development; acetyl-CoA carboxylase carboxyl transferase (*SmAccD*, whz_015069), and alcohol dehydrogenase I (*SmADH1*, whz_020025), which are associated with cell envelope lipid biosynthesis and α-linolenic acid metabolism; and steroid 17α-monooxygenase (*SmCYP17A*, whz_000520),

Downloaded from https://academic.oup.com/hr/article/doi/10.1093/hr/uhac012/6529164 by guest on 30 July 2026

which participates in the synthesis of steroids, fatty acids, and eicosanoids. KEGG enrichment results showed that these contracted gene families were mostly enriched in starch and sucrose metabolism, hormone signal transduction, and sesquiterpenoid and triterpenoid biosynthesis (Supplementary Table). These families included trehalose 6-phosphate synthase (*SmTPS1*, whz_009200), β-glucosidase (whz_002676), fructokinase (*SmFK*, whz_000626), sucrose-phosphate synthase (*SmSPS*, whz_024199), glucan endo-1,3-β-glucosidase 5 (whz_002416), and sucrose synthase (*SmSUS1*, whz_023880), some of which are involved in the biosynthesis of polysaccharides. Furthermore, we also identified 9521 gene families specific to the *S. mukorossi* genome among the nine plants. KEGG enrichment analysis showed that these gene families mainly participated in energy metabolism, environmental adaptation, carbohydrate metabolism, cell growth, signal transduction, and lipid metabolism (Supplementary Table). These families included *SmRPM1* (whz_000421), *SmRPS2* (whz_000391) and *SmCML* (whz_000614), which are associated with disease resistance via plant–pathogen interactions; 3-oxoacyl-[acyl-carrier protein] reductase (whz_003695), enoyl-[acyl-carrier protein] reductase I (whz_001582), acetyl-CoA acyltransferase 1 (whz_005588), lipoxygenase (whz_001602), and linoleate 9S-lipoxygenase (whz_003125), which participate in linoleic acid metabolism and α-linolenic acid metabolism; and mevalonate kinase (*SmMVK*, whz_024682), mevalonate pyrophosphate decarboxylase (*SmMVD*, whz_005809), and squalene synthase (*SmSS*, whz_013402), which are involved in the hederagenin biosynthetic pathway. A total of 126 candidate genes in *S. mukorossi* appeared to be under positive selection according to the following parameters: $Ka/Ks > 1$, $P < .05$. Most of these genes were enriched in GO terms and KEGG pathways related to energy metabolism, cell growth and death, and stimulus response, including *Sm*ND5 (whz_000010), *Sm*ND2 (whz_006058), *Sm*CDC4 (whz_004052), and *Sm*NPR1 (whz_008089). We also found that among these positively selected TF families, *Sm*ARR-B (cytokinin signaling, whz_000208) and *Sm*B3 [the auxin response factor (ARF) family and the LEAFY COTYLEDON 2-ABSCISIC ACID INSENSITIVE 3-VP1/ABI3-LIKE (LAV) family, whz_004052] were involved in hormone signaling pathways during seed development (Supplementary Table S16).

To estimate the nature of whole-genome duplication (WGD) events in *S. mukorossi*, the *Ks* distribution of homologous genes from *S. mukorossi*, *D. longan*, and *A. yangbiense* was characterized. Two *Ks* values peaked at 0.18 and 0.36 for orthologs between *S. mukorossi* and *D. longan* and between *S. mukorossi* and *A. yangbiense*, suggesting speciation events at ~15 and ~31 Mya, which is consistent with the findings of the phylogenetic tree. The major peak of duplication in *S. mukorossi* was 1.90, which indicates that an ancient WGD event had occurred at ~161 Mya but did not reveal a recent WGD event, which is consistent with the *Ks* peaks reported for *D. longan* and *A. yangbiense* (Fig. 2c). These results indicated that the ancient WGD event of *S. mukorossi* occurred before the divergence of *S. mukorossi* and *D. longan*. Although synteny block analysis revealed strong collinearity in *S. mukorossi*, *D. longan*, and *A. yangbiense*, large interchromosomal rearrangements were detected for the homologous chromosome pairs AyChr03/SmChr04, SmChr04/DlChr02, AyChr08/SmChr08, AyChr01/SmChr01, SmChr02/DlChr04, SmChr03/DlChr05, SmChr07/DlChr08, and SmChr13/DlChr14 (Fig. 2d). In addition, partial chromosomal fusions were also observed in the homologous chromosome pairs among these species (AyChr01/Chr10 → SmChr01, AyChr03/Chr13 → SmChr03, AyChr04/Chr12 → SmChr02, DlChr09/Chr13 → SmChr01, SmChr12/Chr14 → DlChr01, DlChr05/Chr07 → SmChr03) despite high collinearity (Fig. 2d). These results suggested that major chromosomal fusions and rearrangements may have led to the divergence of the three species in Sapindaceae, which might be one of the reasons for the difference in chromosome formation.

Population genetic analysis

A total of 104 *S. mukorossi* accessions were collected from nine provinces of China covering nearly all *S. mukorossi* planting areas (Supplementary Table S1). Resequencing of the 104 *S. mukorossi* accessions by Illumina HiSeq PE150 generated a total of 545.83 Gb of clean data with an average depth of 12.27× per or 4.91 Gb of clean data per sample (Supplementary Table S3). After mapping against the assembled reference genome, we identified a total of 19 443 452 high-quality single-nucleotide polymorphisms (SNPs) and 7 960 863 InDels after filtering. Phylogenetic tree and principal component analysis (PCA) revealed that 104 samples were clustered into three independent groups corresponding to the Group I, Group II, and Group III populations (Fig. 3a and b). Group I mainly consisted of accessions from the northeast distribution area, with enrichment of accessions from FJ and ZJ. Group II mainly consisted of accessions from the southern distribution area, with enrichment of accessions from GX and JX. The accessions of the southwestern distribution area (YN) were placed in an independent branch as Group III, showing stronger geographical distribution patterns than the other two groups. To further analyze the genetic relationship between these accessions, we performed a structure analysis using ADMIXTURE (*K*, 2–10) (Fig. 3c). At a *K* value of 2 or 3, Group III was one separate cluster, demonstrating that Group III has had a linear evolutionary route due to geographical isolation. Additionally, the genes of Group III gradually penetrated into Group I and Group II accessions. At a *K* value of 3, the accessions were clearly divided into three separate clusters: Group I, Group II, and Group III, which is consistent with the above PCA and phylogenetic tree. Group I and Group II showed a relatively high level of admixture, which may be attributable to a shared ancestor or recent introgression events. To further understand the evolution of the *S.*

Downloaded from https://academic.oup.com/hr/article/doi/10.1093/hr/uhac012/6529164 by guest on 30 July 2026

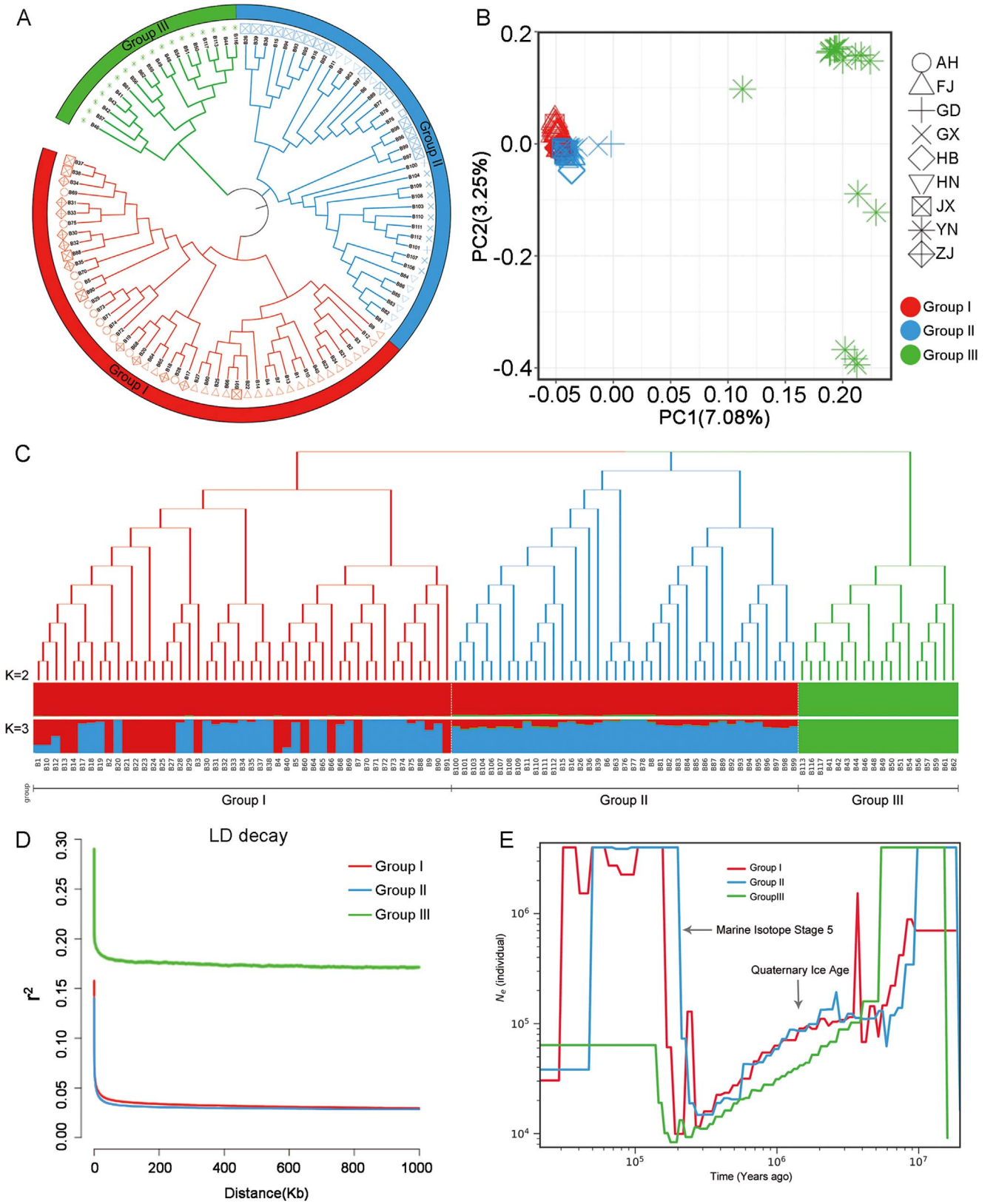

Figure 3. Phylogeny and population genetics of 104 *S. mukorossi* accessions. **a** Phylogenetic tree of 10 *S. mukorossi* accessions constructed based on whole-genome SNPs. **b** PCA plots of 104 *S. mukorossi* accessions. Different colors and shapes represent different subgroups and regions, respectively. **c** Population structure analysis of 104 *S. mukorossi* accessions (K = 2 or 3). The background of each putative ancestor is shown in different colors. **d** Genome-wide decay of LD. **e** Demographic history of N_e with a generation time of 5 years. Neutral mutation rate per generation (μ) = 1.25×10^{-8}.

Downloaded from https://academic.oup.com/hr/article/doi/10.1093/hr/uhac012/6529164 by guest on 30 July 2026

mukorossi population from the perspective of genomics, we analyzed the decay rate of linkage disequilibrium (LD) of the three subpopulations (indicated by r^2). Based on the one-half r^2 maximum of pairwise SNPs among these three groups, the LD decay rates in Group I, Group II, and Group III were 459 bp ($r^2 = .075$), 643 bp ($r^2 = .0703$), and 403 bp ($r^2 = .1446$), respectively (Fig. 3d). The rapid decline in LD in each population may be due to the high density of SNPs and the high degree of recombination in the *S. mukorossi* genome. Similarly, a study of the quail population structure also revealed a genomic background of low-level LD bird populations, which promote adaptive variation. Genetic diversity (π) increased from 5.42×10^{-3} in Group I to 5.59×10^{-3} in Group II and 6.71×10^{-3} in Group III. We also identified the direction of gene flow from Group III (YN) to Group II (HB) to Group I (AH, FJ, JX, GD), indicating that the species in the southwest may be the donor population for *S. mukorossi* planting areas in China (Supplementary Fig. S6). By calculating the fixation index value (F_{ST}) between these three groups, we found that the F_{ST} values were ~0.42 between Group I and Group III populations and ~0.41 between Group II and Group III populations, and the F_{ST} between the Group I and Group II populations was only 0.027. These results further indicated that Group I and Group II populations have relatively low genetic differentiation. Moreover, Group I ($\pi = 5.42 \times 10^{-3}$) and Group II ($\pi = 5.59 \times 10^{-3}$) harbored nearly the same level of genetic diversity with little genetic differentiation ($F_{ST} = 0.027$). Negative mean Tajima's *D* values between Group I (−0.24) and Group II (−0.41) showed that the *S. mukorossi* population has recently accumulated more low-frequency gene mutations, and regional amplification may have occurred in the population history (Supplementary Fig. 7). Overall, Group I and Group II populations, which existed in close geographical areas, formed one robust clade or a single cluster in structure analyses, with *K* values ranging from 2 to 3, confirming their close genetic relationship.

Demographic history

To explore the demographic history of *S. mukorossi*, we estimated N_e using the pairwise sequentially Markovian coalescent (PSMC) method and assumed a mutation rate of 1.25×10^{-8} per generation for a generation time of 5 years. The N_e values of the three groups reflected one dramatic decline period and a recovery period (Fig. 3e). Group I, Group II, and Group III reached the maximum N_e at ~10 Mya (million years ago) and then began to decline dramatically. One sharp reduction may have resulted from climatic fluctuations corresponding to the start of the Quaternary Ice Age (~2 Mya) [24]. The recovery period occurred ~180 Kya (thousand years ago) during the interglacial period of Marine Isotope Stage 5 (~130 Kya), which may have resulted from a warming climate [39]. According to the results of demographic trajectories, we found that the N_e values of Group I and Group II were basically identical and higher than that of Group III. The results may be due to relatively low genetic diversity and geographical distributions during the evolution of *S. mukorossi*.

Genome-wide selective sweep signals

It is generally believed that the difference between these three populations lies in their phenotypic characteristics and habitat mainly. In view of the above results, we found that Group I and Group II are closely related and have had gene flow events. To determine the potential divergent selection of these differences, we used three approaches [including the π ratio ($\pi_{GroupIII}/\pi_{GroupI}$), F_{ST}, and the cross-population composite likelihood ratio (XP-CLR)] to scan selected regions between the Group I and Group III populations with the top 5% of values. In total, we identified 212 and 149 selective sweeps based on the π ratio (>1.25) and F_{ST} (>0.42) results, and these selective sweeps harbored 3857 and 4166 genes, respectively (Fig. 4a). GO and KEGG enrichment results showed that these sweep regions were mostly enriched in secondary metabolite biosynthesis, oxidative phosphorylation, plant hormone signal transduction, α-linolenic acid metabolism, and response stimulus (Supplementary Figs S8 and S9; Supplementary Tables S17–S20). To further narrow our list of candidate genes, we found that 2585 candidate genes were shared according to the π ratio and F_{ST} results. Group I accessions were characterized by the selection of a large number of *RPS* (ribosomal protein S) genes compared with the Group III accessions, including *SmRPS2* (whz_007360), *SmRPS7* (whz_023082), *SmRPS10* (whz_015443), *SmRPS12* (whz_011857), and *SmRPS23* (whz_018422). *Arabidopsis* homologs containing *RPS3* and *RPS5* have been shown to participate in DNA damage and resistance to *Pseudomonas syringae* [40, 41]. We also used the KEGG database to functionally annotate these 2585 genes, which revealed that the most highly enriched categories were involved in lipid metabolism, signal transduction, environmental adaptation, and terpenoid and polyketide metabolism (Supplementary Fig. S10; Supplementary Table S21). In lipid metabolism, 12 genes were involved in α-linolenic acid metabolism and fatty acid biosynthesis, including *SmDAD1* (whz_020348), *SmACOX4* (whz_006370), *SmOPR3* (whz_020852), *SmACOT1* (whz_026593), *SmFAB2* (whz_000745), *SmgalA* (whz_002077), *SmPPT* (whz_020366), *SmSGPL1* (whz_009676), and *SmACAT* (whz_014940) (Fig. 4a). For terpenoid metabolism, 10 genes were associated with monoterpenoid, diterpenoid, and triterpenoid biosynthesis, including *SmSQLE* (whz_022458), *SmKAO* (whz_020277), *SmGA3ox* (whz_020438), *SmGA2ox* (whz_020996), *SmFOLK* (whz_001194), *SmHMGCR* (whz_016467), and *SmGGPS* (whz_003902) (Fig. 4a). Additionally, 160 TFs were also identified in these sweep regions ($P < .05$), including 30 MYBs, 6 NACs, 13 WRKYs, 5 bZIPs, 11 bHLHs, and 1 AP2 (Supplementary Table S21). NACs and WRKYs have been reported to be involved in defense responses, growth, and development, including *SmNAC2* (whz_006513), *SmNAC75* (whz_020862), *SmNAC83* (whz_006262),

Downloaded from https://academic.oup.com/hr/article/doi/10.1093/hr/uhac012/6529164 by guest on 30 July 2026

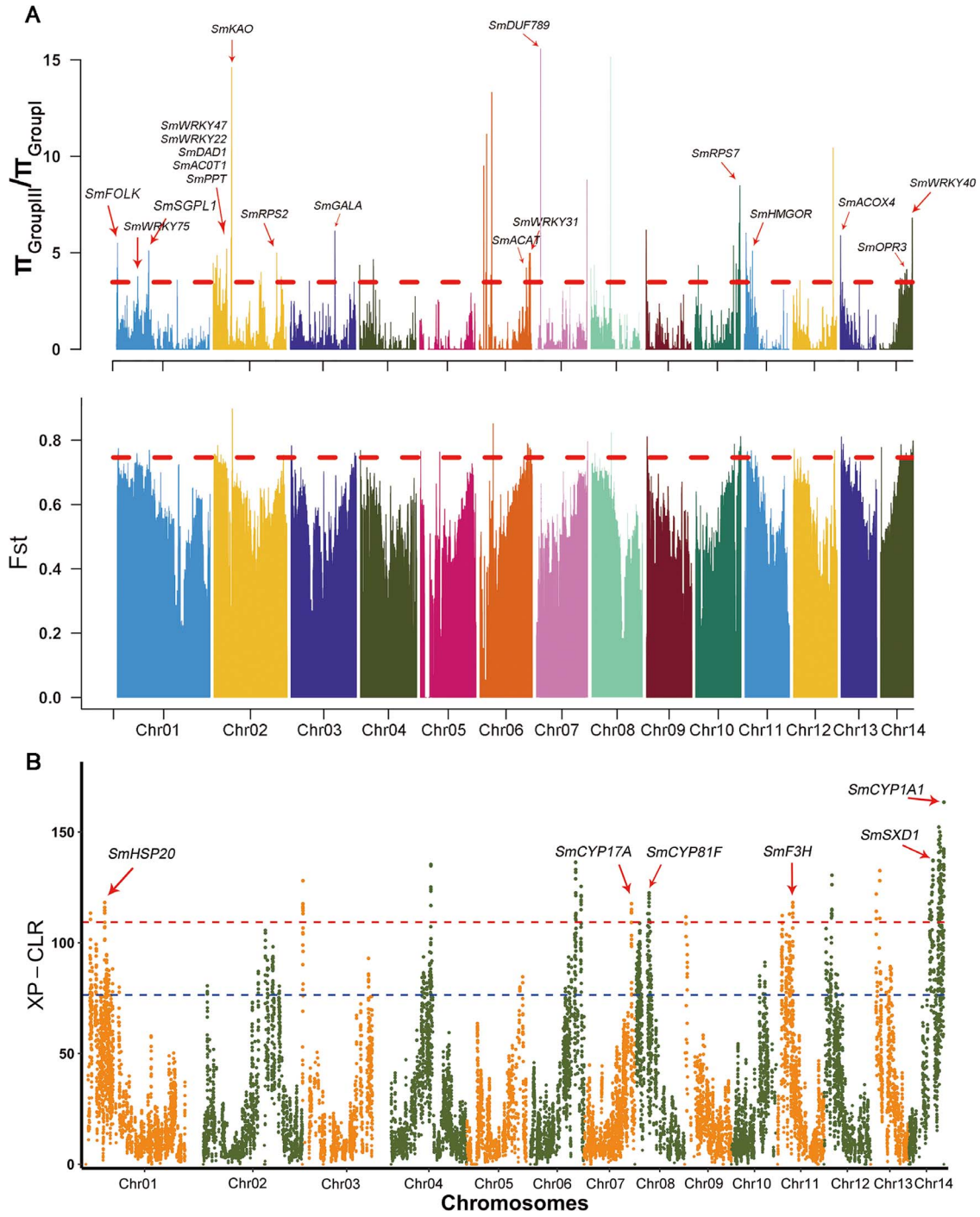

Figure 4. Genome-wide detection and annotation of selective sweeps in 104 *S. mukorossi* accessions. **a** Selective sweeps were identified by $\pi_{GroupIII}/\pi_{GroupI}$ and F_{ST}. The red dashed line corresponds to the thresholds for very high significance in the top 5% of the highest values. **b** Manhattan plots of the CLRs among the Group I and Group III *S. mukorossi* populations. Red and blue dashed lines indicate the threshold for the top 1% and top 5% of CLR values, respectively. Genes located within the significant CLR peaks and corresponding annotations are denoted.

*Sm*WRKY40 (whz_021259), *Sm*WRKY22 (whz_005902), *Sm*WRKY47 (whz_006030), *Sm*WRKY31 (whz_014884), and *Sm*WRKY75 (whz_009675) [42–45].

Using the data generated from XP-CLR with the top 5% of CLR scores, we identified 163 selective sweep regions that harbored 126 genes and 929 selective sweep regions that harbored 2487 genes that were significantly selected in Group I and Group III (Fig. 4b; Supplementary Fig. S11). Using the data generated from functional annotation analysis, we found that oxidative phosphorylation signaling pathways, including *SmND2* (whz_009672), *SmCOX1* (whz_015024), *SmATP6* (whz_015028), *SmND5* (whz_015033), and *SmPPTC7* (whz_026327), were significantly enriched in Group I but not in Group III. Group III accessions were characterized by the selection of a large number of genes involved in flavonoid and triterpene synthesis, including *SmF3H* (whz_016322), *SmSXD1* (whz_020513), *SmCYP17A*

Downloaded from https://academic.oup.com/hr/article/doi/10.1093/hr/uhac012/6529164 by guest on 30 July 2026

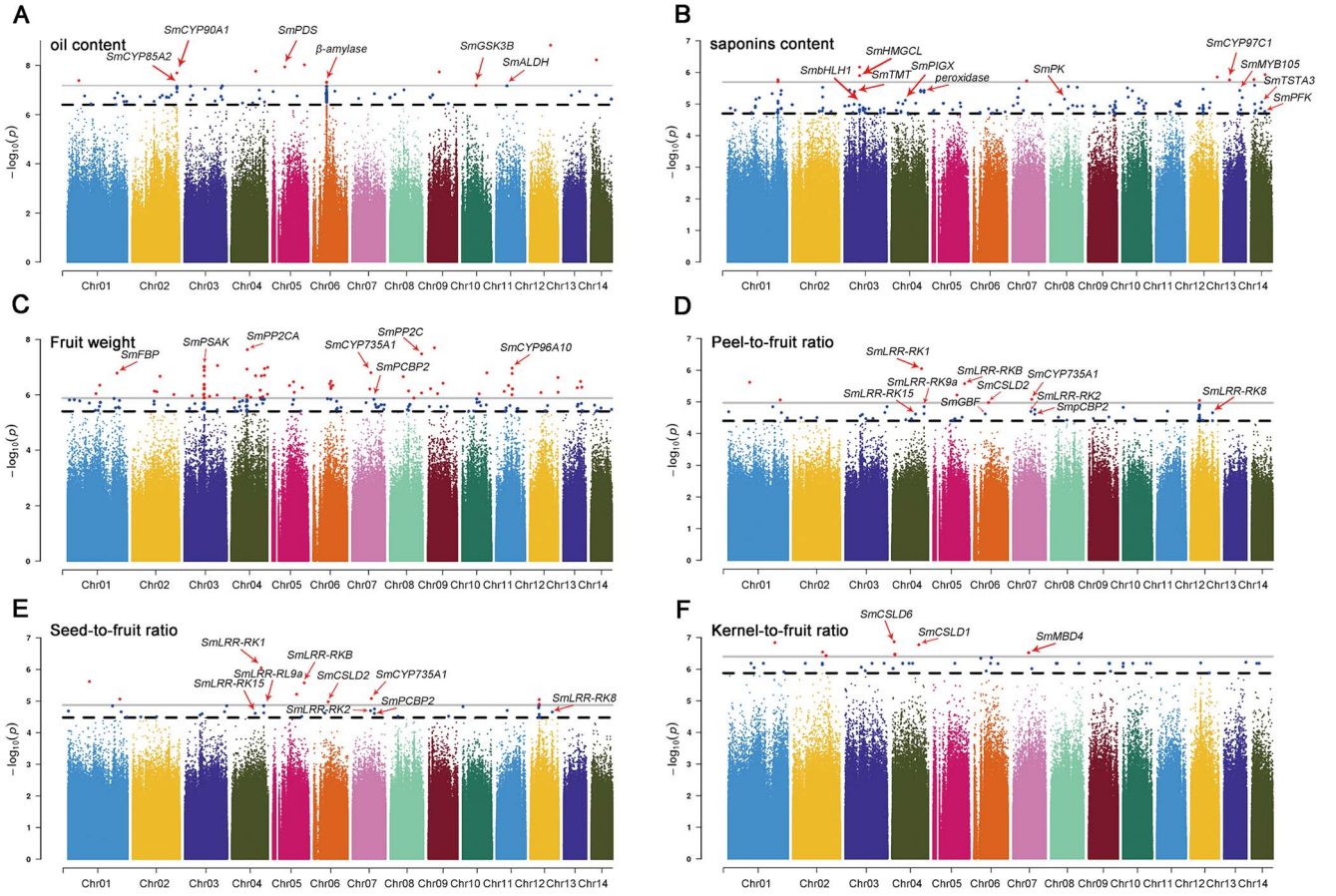

Figure 5. Manhattan plots from GWAS for important agronomic traits in 57 *S. mukorossi* accessions. **a** Oil content. **b** Saponin content. **c** Fruit weight. **d** Peel-to-fruit ratio. **e** Seed-to-fruit ratio. **f** Kernel-to-fruit ratio. Gene names are highlighted in black in candidate regions potentially related to six traits in *S. mukorossi*. Black and gray lines indicate the threshold for the top 5% and top 1% of the highest values, respectively.

(whz_019898), *SmCYP81F* (whz_018510), *SmCYP1A1* (whz_020835), *SmKAO* (whz_018519), and *SmMYBP* (whz_019890). In addition, we also identified related genes involved in resistance in Group III, including *SmHSP20* (whz_001399), *SmWRKY15* (whz_001652), *SmHSP70* (whz_024119), *SmRPS2* (025316), and *SmRSP19* (whz_019904).

Genome-wide association studies for six agronomic traits

We used the phenotyping data for six agronomic traits of fruits of 57 *S. mukorossi* samples collected from seven provinces of China, including FJ (17), JX (15), ZJ (6), YN (3), AH (6), HN (5), and GX (5) (Supplementary Table S2). Saponin and oil contents are two important economic traits for superior individuals of *S. mukorossi*. We detected six strong GWAS signals harboring 47 candidate genes, including *SmCYP90A1* (whz_004291), *SmPDS* (whz_013074), *SmGSK3B* (whz_022075), *SmRPS3* (whz_021468), *β*-amylase (whz_014801), and *SmALDH* (whz_006230) (Fig. 5a; Supplementary Fig. S12). We found that two of the six GWAS signals, located on Chr02 and Chr05, were significantly associated with oil contents, which is consistent with previous reports. For example, two oil-related genes were located in the strong association peaks, including *SmCYP90A1* (whz_004291), which participates in the metabolism of various internal and external substances, such as steroid hormones and fatty acids [46], and *SmGGTS* (whz_017512), which is related to arachidonic acid and cyanoamino acid metabolism [47]. We further detected 73 genes corresponding to six chromosomes (Chr01, Chr03, Chr04, Chr08, Chr13, and Chr14), and a series of genes that have been reported to be significantly associated with the saponin phenotype, including *SmTMT* (whz_003247), *SmCYP97C1* (whz_006615), *SmPIGX* (whz_009575), *SmMYB105* (whz_007070), and *SmbHLH1* transcription factor (whz_003247) (Fig. 5b; Supplementary Fig. S13). In addition, we also identified related genes involved in the metabolism of polysaccharides and flavonoids, including *SmTSTA5* (whz_021561), *SmDFR* (whz_021579) and *SmPK* (whz_019106). *SmAHP* (whz_017177) and *SmPP2C* (whz_017178), which are associated with the fruit weight phenotype, spanning from 5.24 to 5.39 Mb on Chr07 (Fig. 5c; Supplementary Fig. S14). *PP2C* is closely related to abscisic acid signaling, and negatively regulates abscisic acid signaling in the early stages [48]. The *AHP* factor is involved in the signal transduction of cytokinins [49]. GWAS identified a poly(rC)-binding protein 2 gene (*SmPCBP2*, whz_007852) and six leucine-rich repeat receptor kinases (*SmLRR-RKs*),

Downloaded from https://academic.oup.com/hr/article/doi/10.1093/hr/uhac012/6529164 by guest on 30 July 2026

which are associated with the peel-to-fruit ratio and seed-to-fruit ratio (Fig. 5d and e; Supplementary Figs S15 and S16). *SmPCBP2*, as a negative regulator in MAVS-mediated signaling, has been reported to regulate cell growth and death [50]. We performed GWAS for the kernel-to-fruit ratio and found that one SNP located upstream of cellulose synthases (*SmCSLD1*, whz_008313) was strongly associated with the kernel-to-fruit ratio (Fig. 5f; Supplementary Fig. S17). These results provide valuable information for molecular breeding considering genes responsible for fruit weight, the peel-to-fruit ratio, saponin contents, the seed-to-fruit ratio, the kernel-to-fruit ratio, and oil contents in *S. mukorossi*.

Discussion

To obtain a high-confidence SNP set from resequencing data, a reference genome resource is very important. Several genome assemblies in Sapindaceae have been reported, including *A. yangbiense* (~665 Mb) and *D. longan* Lour. (~471 Mb) [30, 31]. However, no report has involved the genome assembly and annotation of *S. mukorossi*. Diverse chromosome numbers (the chromosome numbers are 11, 13, 14, 15, etc.) and high heterozygosity (1.25%) in Sapindaceae have hindered genome assembly. *S. mukorossi* is an autodiploid with $2n = 2x = 28$ chromosomes based on the results obtained for the molecular karyotype and *K*-mer analysis, differing significantly from the chromosome numbers of previously reported Sapindaceae species (*A. yangbiense*, $2n = 2x = 26$; *D. longan*, $2n = 2x = 30$) [30, 31]. Using the latest sequencing and assembly technology, we obtained the genome of *S. mukorossi* at the chromosome level, revealing a genome size of ~391 Mb with a contig N50 length of 24.66 Mb. By thoroughly mapping resequencing data, 104 accessions generated ~545.83 Gb of clean data and identified 4 351 219 high-quality SNPs and 7 960 863 InDel datasets, yielding the richest genomic resource for further molecular breeding and biological studies of *S. mukorossi* to date.

The resequencing data of the 104 accessions collected from nine provinces of China covered nearly all *S. mukorossi* planting areas. The nucleotide diversity of *S. mukorossi* (7.53×10^{-3}) was significantly higher than that of other woody plants, such as peach (1.10×10^{-3}), tea (6.12×10^{-3}), apple (2.20×10^{-3}), jujubes (1.73×10^{-3}), pear (5.50×10^{-3}), and *Prunus mume* (2.42×10^{-3}), indicating that the evolutionary process of *S. mukorossi* is characterized by relatively high genetic diversity, which is conducive to adaption to various adverse environments. We further calculated the F_{ST} between these three groups and found that the FST of the Group I and Group II accessions, at 0.002, was lower than that of the Group III accessions. These results also suggested that the close relation between accessions from the northeast and south may be attributable to a shared ancestor or recent introgression events due to close geographical areas.

KEGG and GO analyses showed that some of the specific or expanded genes in *S. mukorossi* were significantly enriched in disease resistance, hormone signaling pathways, and lipid metabolism, including *SmRPM1*, *SmRPS2*, acetyl-CoA acyltransferase 1, lipoxygenase and linoleate 9S-lipoxygenase, *Sm*ARR-B, and *Sm*B3. Linoleate 9S-lipoxygenase and lipoxygenase, as key enzymes in linoleic acid and α-linolenic acid metabolism, could enhance key oxylipin synthesis pathways in tillers when upregulated [51]. Expansion of these genes is speculated to affect the accumulation of lipid metabolites. Kernel oil content plays an important role in *S. mukorossi* breeding. Moreover, Group I accessions were also characterized by the selection of a large number of *RPS*s compared with Group III accessions, including *SmRPS2*, *SmRPS4*, *SmRPS7*, *SmRPS10*, *SmRPS12*, and *SmRPS23*. The resistance genes *RPS3/5* in *Arabidopsis* mediate resistance to *P. syringae* and DNA damage, and *P. syringae* has been reported to infect many economically important crops [52]. It is speculated that the strong selectivity of these genes enhances environmental adaptability, resulting in better cold tolerance in the northeast than in the southwest. Genome-wide selective sweep analysis revealed that most candidate genes were highly involved in lipid metabolism and terpenoid and polyketide metabolism, including *SmDAD1*, *SmACOX4*, *SmOPR3*, *SmACOT1*, *SmFAB2*, *SmgalA*, *SmPPT*, *SmSGPL1*, *SmACAT*, *SmSQLE*, *SmKAO*, *SmGA3ox*, *SmGA2ox*, *SmFOLK*, *SmHMGCR*, *SmGGPS*, *SmF3H*, *SmCYP17A*, *SmCYP81F*, and *SmCYP1A1*. *SmGA3ox* and *SmGA2ox* are diterpenoid plant hormones with a key role in regulating various processes throughout the life cycles of plants [53]. *SmGGPS* encodes an intermediate in terpenoid biosynthesis that is synthesized by geranylgeranyl diphosphate synthetases [54]. Many studies have focused on the identification of CYP450s for terpenoid biosynthesis, such as CYP716A47, which is related to ginsenoside biosynthesis [55]; CYP76AH1, which catalyzes the turnover of miltiradiene in tanshinone biosynthesis [56]; and CYP71AV14 and CYP71AX30, which are related to sesquiterpene lactone biosynthesis [57]. We speculate that differential selection in these regions may result in richer saponin and oil contents in northeastern accessions than in southwestern accessions, which is consistent with the phenotypic data.

Fruit weight, the peel-to-fruit ratio, saponin contents, the seed-to-fruit ratio, the kernel-to-fruit ratio, and oil contents are important agronomic traits for selecting excellent germplasm resources. To identify potential candidate genes associated with these traits, we performed GWAS using resequencing data from 57 accessions. We found that four peel-to-fruit ratio- and seed-to-fruit ratio-related *LRR-RK* genes have been reported to regulate cell proliferation, stem cell maintenance, hormone perception and other developmental processes in *Arabidopsis* [58]. Similarly, we detected one SNP located upstream of cellulose synthases (*SmCSLD1*, whz_008313) that is responsible for the kernel-to-fruit ratio. *SmCSLD1* is a component of the cell wall biosynthetic machinery and participates in the production of energy [59].

Downloaded from https://academic.oup.com/hr/article/doi/10.1093/hr/uhac012/6529164 by guest on 30 July 2026

Saponins and oils are also important active compounds in *S. mukorossi*, and different saponin and oil contents may contribute to variation in economic value with respect to breeding. Interestingly, we found that two saponin-related genes were located in strong association peaks, including *SmTMT* (whz_003247), which is involved in ubiquinone and other terpenoid-quinone biosynthesis [60], and *SmbHLH1* transcription factor (whz_003247), which regulates triterpene saponin biosynthesis in *Medicago truncatula* [61]. Our analysis identified candidate genes associated with major agronomic traits, and this information can be used for further reference in the selection of superior germplasm resources. Chromosome-level reference genome assembly and population resequencing data have also substantially expanded basic and applied research for woody trees of Sapindaceae.

Materials and methods

Molecular karyotype analysis and genome size estimation of *S. mukorossi*

S. mukorossi samples were obtained from Jianou, Nanping, Fujian, China (118.09′20″ E, 27.03′01″ N). Fluorescence *in situ* hybridization and karyotype analysis were carried out according to a previously reported method [62]. The size of the *S. mukorossi* genome was estimated by an optimized DNA flow cytometry method according to Huang *et al.* [63]. The genome size (1C) of *S. mukorossi* (~380 Mb) was estimated to be ~0.43-fold that of tomato (~880 Mb) as an internal reference. Jellyfish software was used to estimate genome size and heterozygosity by 21-*K*-mers with ~21.0 Gb (47×) of clean short reads on the Illumina HiSeq X Ten platform [64].

Sequencing and assembly of the *S. mukorossi* genome

Genomic DNA was isolated from young leaves using the CTAB method according to a previously reported protocol [65]. One SMRTbell library was constructed and sequenced on a PacBio Sequel II for HiFi CCS. A total of 409.15 Gb of CCS subreads were generated with a read N50 length of 13.45 kb and an average length of 13.16 kb. After trimming and quality control of all reads, ~22.96 Gb of HiFi CCS data were used for genome assembly with hifiasm software (https://github.com/chhylp123/hifiasm), followed by removal of heterozygous sequences with Khaper (https://github.com/lardo/khaper) based on the 21-*K*-mers result. Fresh leaves of *S. mukorossi* were harvested and placed in formaldehyde for 20 min for cross-linking, followed by endonuclease digestion (*Dpn* II), end repair, cyclization, and purification. The captured DNA fragments were used to construct a Hi-C sequencing library. The scaffolds/contigs of the assembled genome sequence were divided, sorted, and oriented into chromosomes using LACHESIS with parameters: CLUSTER_NONINFORMATIVE_RATIO = 1.4, CLUSTER_MAX_LINK_DENSITY = 2.5, CLUSTER_MIN_RE_SITES = 100, ORDER_MIN_N_RES_IN_TRUNK = 60, and ORDER_MIN_N_RES_IN_SHREDS = 60 [66]. Placement and orientation errors exhibiting obvious discrete chromatin interaction patterns were manually adjusted.

Genome quality assessment

Genome completeness was assessed using the BUSCO databases with the Embryophyta model. High-quality ESTs were mapped to the *S. mukorossi* genome using BLAST (version 0.35) [67]. After trimming and quality control, second-generation data from the Illumina platform were mapped to the genome assembly using Bowtie2 to assess coverage [68].

Genome annotation

Repeat sequences were identified by combining *de novo* annotation and homology-based methods. For *de novo* annotation, RepeatModeler (version 1.0.4) was used to search for repetitive sequences in the *S. mukorossi* genome, and then the results were used to build a repetitive sequence library [69]. We further used RepeatMasker (version 2.1) to identify repetitive sequences in a previously built repetitive sequence library [70]. For homology-based prediction, a homology-based detection procedure was performed using a conserved BLASTN search in Repbase of RepeatMasker [71]. The repetitive sequence results by *de novo* annotation and homology-based annotation were combined by removing redundant transposable elements. LTR_FINDER (version 1.05) was used to search for intact LTR retrotransposons in the *S. mukorossi* genome [72]. rRNAs and tRNAs were predicted using RNAmmer (version 1.2) and tRNAsan-SE (version 1.23), respectively [73]. In addition, miRNAs and snRNAs were identified using INFERNAL (version 1.1.1) to search the Pfam database [74].

Protein-coding genes were annotated by the MAKER pipeline (version 2.31.9) based on an integrated strategy, including homology-based prediction, RNA-seq data and *ab initio* annotation [75]. For *de novo* gene prediction, we first used BRAKER software to automatically generate full gene structure models with genomic and RNA-seq data [76]. Augustus (version 3.0.3) was applied with the default parameters, and the protein sequences of the training sets used were from the related species *A. yangbiense* and *D. longan* [77]. All of the predicted genes were functionally annotated according to homologous alignments with BLASTP (e-value ≤1e−5, max_target_seqs 1) against the NR, Swiss-Prot, GO, KOG, and KEGG databases [78–82]. TFs were identified by performing BLASTP (e-value ≤1e−5) searches against the plant transcription factor database (PlantTFDB version 5.0).

Construction of a phylogenetic tree and estimation of gain and loss of gene families

To identify the expanded and contracted *S. mukorossi* gene families, a phylogenetic tree was constructed using orthologous genes for *Punica granatum*, *R. argentea*, *S. oleosum*, *Corymbia citriodora*, *E. grandis*, *R. rubrinervis*, *C. sativa*, *R. chinensis*, *F. vesca*, *M. baccata*, *C. clementina*, *C. unshiu*, *C.*

Downloaded from https://academic.oup.com/hr/article/doi/10.1093/hr/uhac012/6529164 by guest on 30 July 2026

sinensis, *D. longan*, *S. mukorossi*, *A. yangbiense*, *Manihot esculenta*, *Salix dunnii*, *S. brachista*, *P. alba*, and *P. euphratica*. A total of 718 single-copy gene families were shared by the 21 analyzed genomes and used to construct a phylogenetic tree using FastTree (version 2.1.9) and RAxML (version 8.2.12) based on the maximum likelihood method [83]. MCMCtree in PAML (version 4.9) and the TimeTree website (http://www.timetree.org/) were used to estimate the time of species divergence with estimated confidence intervals at the 95% level [84]. CAFE software (version 1.6) was used to calculate the expansion and contraction of gene families [85]. Significantly expanded or contracted genes ($P \leq .01$) were annotated and enriched by GO and KEGG analyses, respectively.

Synteny and whole-genome doubling analysis

We analyzed WGD events and determined the source of the homologous genes in *S. mukorossi*. The protein sequences from *S. mukorossi* and other species were searched to identify the best matching pairs using BLASTP (e-value ≤1e−5). Syntenic block analysis was performed using MCScanX with parameters –s 5 –m 5, and *Ks* values (synonymous substitution rate) for syntenic gene pairs were calculated using KaKs_Calculator software [86, 87].

Positively selected genes in *S. mukorossi*

Orthologous gene clusters in *S. mukorossi* and closely related plants were identified using OrthoFinder (version 2.3.12) [88]. The single-copy genes of *S. mukorossi* and closely related species were aligned using MUSCLE [89]. Positive selection sites were detected using Codeml software of the PAML package. The *Ka*/*Ks* ratios of the single-copy genes were evaluated with a threshold of $Ka/Ks \geq 1$ and *P* value ≤.05.

DNA preparation for population sequencing, variant calling and annotation

A total of 104 *S. mukorossi* strains collected from various parts of the world were sequenced with the Illumina HiSeq X Ten platform with an insert size of ∼400 bp. DNA of each sample was isolated using the CTAB method according to a previously reported protocol [65]. The complete raw data of the 104 samples were processed for quality control using Trimmomatic with the default parameters [90], and the clean reads were then aligned to *S. mukorossi* using Bwa [91]. The alignments were sorted using SAMtools (version 1.9) [92], and PCR-generated duplicates were removed using Picard (version 2.17). Variant detection was performed for each sample using the GATK (version 3.8.0) HaplotypeCaller tool and subsequently joint-called across all samples using the GATK GenotypeGVCFs tool [92]. To obtain high-quality SNPs, a GATK hard filter was used to filter and merge VCF files with the parameters QD < 2.0 || FS ≤ 60 || MQRankSum< −12.5 || ReadPosRankSum< −8.0 || SOR > 3.0 || MQ < 40.0. The marker set was annotated using SnpEff (version 4.3) with our gene annotation for *S. mukorossi* [93].

Population genetics analysis

A subset of SNPs from 104 resequencing samples were filtered using PLINK (version 1.9) with the default parameters —geno 0.1, —maf 0.01 [94]. The final SNP set was used to construct a phylogenetic tree using PHYLIP (version 3.2) software based on the neighbor-joining method and implemented within FigTree (version 1.4.2) (http://tree.bio.ed.ac.uk/software/figtree/). PCA of the final SNP set was performed with PLINK (version 1.9), and patterns of genetic variation were visualized using the R package (version 3.4) [95]. To further illustrate the evolutionary history of the *S. mukorossi* genome, a model-based clustering algorithm implemented in ADMIXTURE was used to estimate the relative genome composition for each sample. Population structure analysis was carried out using ADMIXTURE (version 1.3.0), which was run for *K* values from 1 to 10 with 20 bootstrapping replicates [96]. To perform TreeMix analysis, the per-site allele count of each group was calculated using PLINK with the —within command, and the results were then converted into TreeMix input format using the plink2treemix.py script. We then ran TreeMix to construct a maximum-likelihood population tree [97]. LD between pairs of SNPs at up to 100-kb intervals was calculated using PLINK (version 1.9), and PopLDdecay (version 3.26) was used to plot the LD decay graphs by setting the parameter MaxDist to 1000 [98]. To identify candidate regions potentially affected by selection, nucleotide diversity (π), and genetic differentiation (F_{ST}) were calculated using VCFtools (version 0.1.13) in a non-overlapping 100-kb sliding window with a step size of 10 kb [99]. Tajima's *D* was calculated using VariScan (version 2.0.3) in a non-overlapping 100-kb window with a step size of 10 kb [100].

Genome-wide selective sweep analysis

Whole-genome detection of selective sweeps was conducted using F_{ST} between Group I (47) and Group III (18) populations. To detect the regions under selective sweeps, we used the XP-CLR score to confirm the selective sweeps, with the highest being 5% via the XP-CLR method. The CLR test was implemented in SweepFinder2 for multiple markers across a contiguous grid space [101]. GO term enrichment analysis was performed for candidate genes located within these sweep regions.

Genome-wide association analysis

We used 57 *S. mukorossi* accessions to perform GWAS on fruit weight, the peel-to-fruit ratio, saponin contents, the seed-to-fruit ratio, the kernel-to-fruit ratio, and oil contents. A subset of SNPs from 57 resequencing samples were filtered using PLINK (version 1.9) with the default parameters —geno 0.1, —maf 0.05 [94]. The final SNP set

Downloaded from https://academic.oup.com/hr/article/doi/10.1093/hr/uhac012/6529164 by guest on 30 July 2026

was used for GWAS by EMMAX and a mixed linear model, and the calculation formula was as follows:

$$Y = \mu + Xb + Z\gamma + Mu + P\lambda + e.$$

where Y is the observed value vector of the population quantitative trait phenotype; μ is the overall mean; b is the fixed effect vector; γ is the SNP effect vector; u is the polygenic effect vector; λ is the primary weight vector; and X and Z are the correlation matrices corresponding to the fixed effect vectors b and γ, respectively. M is the correlation matrix corresponding to the random effect vector u; P is the correlation matrix corresponding to the primary weight vector λ; and e is the random residual error.

To minimize false-positives, a kinship matrix was estimated with the EMMAX-kin-intel64 program. The Bonferroni threshold for genome-wide significant truncation was set at 0.05/total SNPs. *P* values were used to screen out potential candidate SNPs. Manhattan diagrams and QQ plots were used to represent and evaluate the effect of association analysis, respectively.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (31971643) (Q.Z.), the Industry-University Cooperation Project of Fujian Science and Technology Department (2020 N5008, 2019 N5009) (B.L.) and the Special Funding Project of Fujian Provincial Department of Finance (SC-299) (B.L.).

Author contributions

B.L. and T.X. designed and coordinated the entire project. B.L., T.X., Y.C., and Q.Z. together led and performed the entire project. T.Z., H.F., and Y.H. collected and processed the samples. T.X. and D.C. performed the genome evolution analyses and GWAS. B.L., T.X., and Y.C. participated in manuscript writing and revision. All authors read and approved the final manuscript.

Data availability

Whole-genome sequencing data and resequencing data have been deposited at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, under BioProject accession numbers CRA004811, CRA004795 and CRA004820, respectively (https://ngdc.cncb.ac.cn). Whole-genome assembled data have been deposited at the Genome Warehouse of the National Genomics Data Center under accession number GWHBECP00000000 (https://ngdc.cncb.ac.cn/gwh/).

Conflict of interest

The authors declare that they have no conflicts of interest.

Supplementary data

Supplementary data is available at *Horticulture Research* online.

References

1. Goyal S, Kumar D, Menaria G *et al.* Medicinal plants of the genus *Sapindus* (Sapindaceae) – a review of their botany, phytochemistry, biological activity and traditional uses. *J Drug Deliv Ther*. 2014;**4**:7–20.
2. Li SZ. *The Compendium of Materia Medica*. People's Medical Publishing House; 1982.
3. Zhu S, Fang PJ. Beijing: People's Medical Publishing House, 1959.
4. Shao WH. Geographic variation of saponins contents in *Sapindus mukorossi* peels from different habitats. *Bull Bot Res*. 2012;**32**:627–31.
5. Rupeshkumar G, Surekha SK, Balu CA *et al.* Study of functional properties of *Sapindus mukorossi* as a potential bio-surfactant. *Indian J Sci Technol*. 2011;**4**:530–3.
6. Huang SM, Wang JW, Du MH *et al.* Fatty acid composition analysis of *Sapindus mukorossi* Gaerth. seed oil. *China Oils Fats*. 2009;**34**:74–6.
7. Sun S, Ke X, Cui L *et al.* Enzymatic epoxidation of *Sapindus mukorossi* seed oil by perstearic acid optimized using response surface methodology. *Ind Crops Prod*. 2011;**33**:676–82.
8. Kanchanapoom T, Kasai R, Yamasaki K. Acetylated triterpenes saponins from the Thai medicinal plant. *Biol Pharm Bull*. 2001;**49**:1195–7.
9. Guo Y, Jian-ping XIE, Ai-hua LIU *et al.* Overview of the pharmacological action of *Sapindus mukorossi*. *J Path Biol*. 2011;**6**:873–4.
10. Shao WH, Yue HF, Jiang JM *et al.* Study on genetic variation of seedling growth from different provenances of *Sapindus mukorossi*. *J ZheJiang For Sci Technol*. 2012;**32**:21–5.
11. Fan HH, Yao XM, Tang XH *et al.* A study on superior individual comprehensive appraisal of *Sapindus mukurossi*. *J Fujian For Sci Technol* 2014;**41**:1–6.
12. Huang S, Pan WM. Good landscape trees: *Sapindus mukorossi* Gaertn. *Garden*. 2004;**9**:42.
13. Zhang FL. Study on tissue culture for soapberry. *J Mountain Agr Biol*. 2005;**24**:119–23.
14. Wei FY, Fang C. Enzyme-assisted aqueous extraction of *Sapindus saponins*. *Appl Chem Ind*. 2010;**39**:1149–51.
15. Kanchanapoom T, Kasai R, Yamasaki K. Acetylated triterpene saponins from the Thai medicinal plant. *Chem Pharm Bull (Tokyo)*. 2001;**49**:1195–7.
16. Diao SF, Shao WH, Chen T *et al.* Study on variation in phenotypic traits of fruit and seed of *Sapindus mukorossi* in seedling plantation. *J Northwest A&F Univ*. 2016;**29**:176–82.
17. Liu J, Liu S, Xu Y *et al.* Screening of *Sapindus* germplasm resources in China based on agro-morphological traits. *J For Res*. 2021;**21**:1350–8.
18. Mahar KS, Palni LMS, Ranade SA *et al.* Molecular analyses of genetic variation and phylogenetic relationship in Indian soap nut (*Sapindus* L.) and closely related taxa of the family Sapindaceae. *Meta Gene*. 2017;**13**:50–6.
19. Hong L, Bai ME, Zhang JZ *et al.* Genetic diversity and relationship analysis of *Sapindus mukorossi* germplasm by ISSR and SRAP molecular markers. *J Zhejiang Agr Sci*. 2013;**5**:71–3.
20. Diao SF, Shao WH, Chen T *et al.* Genetic diversity of *Sapindus mukorossi* natural populations in China based on ISSR. *Eur For Inst Proc*. 2016;**29**:176–82.

Downloaded from https://academic.oup.com/hr/article/doi/10.1093/hr/uhac012/6529164 by guest on 30 July 2026

21. Sun CW, Wang L, Liu J *et al.* Genetic structure and biogeographic divergence among *Sapindus* species: an inter-simple sequence repeat-based study of germplasms in China. *Ind Crops Prod*. 2018;**118**:1–10.
22. Jiang CC, Lu XK, Ye XF. Molecular genetic diversity analysis of energy plant *Sapindus mukorossi* germplasm resources. *Mol Plant Breed*. 2016;**14**:2888–95.
23. Zeng LX, Chen S, Yang YY *et al.* The resistance evaluation and their utilization of the germplasm resources from international rice bacterial blight nursery (IRBBN) in South China. *J Plant Genet Res*. 2013;**14**:930–5.
24. Guo M, Zhang Z, Li S *et al.* Genomic analyses of diverse wild and cultivated accessions provide insights into the evolutionary history of jujube. *Plant Biotechnol J*. 2021;**19**:517–31.
25. Zhou Z, Jiang Y, Wang Z *et al.* Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat Biotechnol*. 2015;**33**: 408–14.
26. Varshney RK, Thudi M, Roorkiwal M *et al.* Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. *Nat Genet*. 2019;**51**:857–64.
27. Li Y, Cao K, Zhu G *et al.* Genomic analyses of an extensive collection of wild and cultivated accessions provide new insights into peach breeding history. *Genome Biol*. 2019;**20**: 36–54.
28. Chen JH, Huang Y, Brachi B *et al.* Genome-wide analysis of cushion willow provides insights into alpine plant divergence in a biodiversity hotspot. *Nat Commun*. 2019;**10**: 5230–42.
29. Wang XC, Feng H, Chang Y *et al.* Population sequencing enhances understanding of tea plant evolution. *Nat Commun*. 2020;**11**:4447–57.
30. Lin YL, Min J, Lai R *et al.* Genome-wide sequencing of longan (*Dimocarpus longan* Lour.) provides insights into molecular basis of its polyphenol-rich characteristics. *GigaScience*. 2017;**6**:1–14. https://doi.org/10.1093/gigascience/gix023.
31. Yang J, Wariss HM, Tao L *et al.* De novo genome assembly of the endangered *Acer yangbiense*, a plant species with extremely small populations endemic to Yunnan province, China. *GigaScience*. 2019;**8**. https://doi.org/10.1093/gigascience/giz085.
32. Huang SX, Diang J, Deng D *et al.* Draft genome of the kiwifruit *Actinidia chinensis*. *Nat Commun*. 2013;**4**:1055–73.
33. Ming R, VanBuren R, Wai CM *et al.* The pineapple genome and the evolution of CAM photosynthesis. *Nat Genet*. 2015;**47**: 1435–42.
34. Jaillon O, Aury JM, Noel B *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*. 2007;**449**:463–7.
35. Ma Q, Feng K, Yang W *et al.* Identification and characterization of nucleotide variations in the genome of *Ziziphus jujuba* (Rhamnaceae) by next generation sequencing. *Mol Biol Rep*. 2014;**41**:3219–23.
36. Xu Q, Chen LL, Ruan X *et al.* The draft genome of sweet orange (*Citrus sinensis*). *Nat Genet*. 2013;**45**:59–66.
37. Wu J, Wang Z, Shi Z *et al.* The genome of the pear (*Pyrus bretschneideri* Rehd.). *Genome Res*. 2013;**23**:396–408.
38. Riccardo V, Chagne D, Merriman TR *et al.* The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nat Genet*. 2010;**42**:833–9.
39. Petit JR, Jouzel J, Raynaud D *et al.* Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature*. 1999;**399**:429–36.
40. Mackey D, Holt BF, Wiig A *et al.* RIN4 interacts with *Pseudomonas syringae* type III effector molecules and is required for RPM1-mediated resistance in *Arabidopsis*. *Cell*. 2002;**108**:743–54.
41. Zhang ZB, Wu Y, Gao M *et al.* Disruption of PAMP-induced MAP kinase cascade by a *Pseudomonas syringae* effector activates plant immunity mediated by the NBLRR protein SUMM2. *Cell Host Microbe*. 2012;**11**:253–63.
42. Bu QY, Jiang H, Li CB *et al.* Role of the *Arabidopsis thaliana* NAC transcription factors ANAC019 and ANAC055 in regulating jasmonic acid-signaled defense responses. *Cell Res*. 2008;**18**: 756–67.
43. Sun L, Zhang H, Li D *et al.* Functions of rice NAC transcriptional factors, ONAC122 and ONAC131, in defense responses against *Magnaporthe grisea*. *Plant Mol Biol*. 2013;**81**:41–56.
44. Xue T, Zheng XH, Chen D *et al.* A high-quality genome provides insights into the new taxonomic status and genomic characteristics of *Cladopus chinensis* (Podostemaceae). *Hortic Res*. 2020;**7**:46.
45. Li JB, Luan YS, Jin H. The tomato SlWRKY gene plays an important role in the regulation of defense responses in tobacco. *Biochem Biophys Res Commun*. 2012;**427**:671–6.
46. Cheng J, Frishman WH, Aronow WS. Updates on cytochrome P450-mediated cardiovascular drug interactions. *Dis Mon*. 2010;**56**:163–79.
47. Wickham S, West MB, Cook PF *et al.* Gamma-glutamyl compounds: substrate specificity of gamma-glutamyl transpeptidase enzymes. *Anal Biochem*. 2011;**414**:208–14.
48. Yue M, Szostkiewicz I, Korte A *et al.* Regulators of PP2C phosphatase activity function as abscisic acid sensors. *Science*. 2009;**324**:1064–8.
49. Shiota H, Ko S, Wada S *et al.* A carrot G-box binding factor-type basic region/leucine zipper factor DcBZ1 is involved in abscisic acid signal transduction in somatic embryogenesis. *Plant Physiol Biochem*. 2008;**46**:550–8.
50. Qin Y, Xue B, Liu C *et al.* NLRX1 mediates MAVS degradation to attenuate hepatitis C virus-induced innate immune response through PCBP2. *J Virol*. 2017;**91**:1264–81.
51. Yan HF, Zhou H, Luo H *et al.* Characterization of full-length transcriptome in *Saccharum officinarum* and molecular insights into tiller development. *BMC Plant Biol*. 2021;**21**:228.
52. Xin XF, Kvitko B, He SY. *Pseudomonas syringae*: what it takes to be a pathogen. *Nat Rev Microbiol*. 2018;**16**:316–28.
53. Pan C, Tian K, Ban Q *et al.* Genome-wide analysis of the biosynthesis and deactivation of gibberellin-dioxygenases gene family in *Camellia sinensis* (L.) O. Kuntze. *Genes*. 2017;**8**:235.
54. Li F, Wei CY, Qiao C *et al.* Three duplication events and variable molecular evolution characteristics involved in multiple GGPS genes of six Solanaceae species. *J Genet*. 2016;**95**:453–7.
55. Li C, Zhu Y, Guo X *et al.* Transcriptome analysis reveals ginsenosides biosynthetic genes, microRNAs and simple sequence repeats in *Panax ginseng* C. A. Meyer. *BMC Genomics* 2013;**14**:245.
56. Guo J, Zhou YJ, Hillwig ML *et al.* CYP76AH1 catalyzes turnover of miltiradiene in tanshinones biosynthesis and enables heterologous production of ferruginol in yeasts. *Proc Natl Acad Sci USA*. 2013;**110**:12108–13.
57. Li Y, Gou J, Chen F *et al.* Comparative transcriptome analysis identifies putative genes involved in the biosynthesis of xanthanolides in *Xanthium strumarium* L. *Front Plant Sci*. 2016;**7**:1317.
58. Torii KU. Leucine-rich repeat receptor kinases in plants: structure, function, and signal transduction pathways. *Int Rev Cytol*. 2004;**234**:1–46.

Downloaded from https://academic.oup.com/hr/article/doi/10.1093/hr/uhac012/6529164 by guest on 30 July 2026

59. Saxena IM, Brown MR. Biochemistry and molecular biology of cellulose biosynthesis in plants: prospects for genetic engineering. *Bioeng Mol Biol Plant Pathw*. 2008;**1**:135–60.
60. David S, Dean D. Elevating the vitamin E content of plants through metabolic engineering. *Science*. 1998;**282**:2098–100.
61. Mertens J, Pollier J, Bossche RV *et al.* The bHLH transcription factors TSAR1 and TSAR2 regulate triterpene saponin biosynthesis in *Medicago truncatula*. *Plant Physiol*. 2016;**170**:194–210.
62. Yu F, Zhao X, Chai J *et al.* Chromosome-specific painting unveils chromosomal fusions and distinct allopolyploid species in the saccharum complex. *New Phytol.*. 2022;**233**:1953–65.
63. Huang Y, Ding W, Zhang M *et al.* The formation and evolution of centromeric satellite repeats in *Saccharum* species. *Plant J*. 2021;**106**:616–29.
64. Guillaume C, Rugen M, Bockmayr A *et al.* A model-based method for investigating bioenergetic processes in autotrophically growing eukaryotic microalgae: application to the green algae *Chlamydomonas reinhardtii*. *Biotechnol Prog*. 2011;**27**:631–40.
65. Chen DH, Ronald PC. A rapid DNA minipreparation method suitable for AFLP and other PCR applications. *Plant Mol Biol Report*. 1999;**17**:53–7.
66. Korbel JO, Lee C. Genome assembly and haplotyping with Hi-C. *Nat Biotechnol*. 2013;**31**:1099–101.
67. Camacho C, Coulouris G, Avagyan V *et al.* BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;**10**:421.
68. Langdon WB. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Min*. 2015;**8**:1–7.
69. Flynn JM, Hubley R, Goubert C *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA*. 2020;**117**:9451–7.
70. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*. 2009;**25**:1–14.
71. Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protoc Bioinformatics*. 2004;**4**: 10.1–10.14.
72. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res*. 2007;**35**:W265–8.
73. Lowe TM, Chan PP. tRNAscan-SE on-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res*. 2016;**44**:W54–7.
74. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. *Bioinformatics*. 2009;**25**:1335–7.
75. Kondrashov N, Pusic A, Stumpf CR *et al.* Ribosome-mediated specificity in Hox mRNA translation and vertebrate tissue patterning. *Cell*. 2011;**145**:383–97.
76. Hoff KJ, Lange S, Lomsadze A *et al.* BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*. 2016;**32**:767–9.
77. Stanke M, Keller O, Gunduz I *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res*. 2006;**34**: W435–9.
78. Ashburner M, Ball CA, Blake JA *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet*. 2000;**25**:25–9.
79. Kanehisa M, Goto S, Sato Y *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*. 2014;**42**:D199–205.
80. Tatusov RL, Galperin MY, Natale DA *et al.* The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*. 2000;**28**:33–6.
81. Boeckmann B, Bairoch A, Apweiler R *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*. 2003;**31**:365–70.
82. Jensen L, Julien P, Kuhn M *et al.* eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res*. 2007;**36**:D250–4.
83. Liu K, Randal LC, Tandy W. RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS One*. 2011;**6**:e27731.
84. Puttick MN. MCMCtreeR: functions to prepare MCMCtree analyses and visualize posterior ages on trees. *Bioinformatics*. 2019;**35**:5321–2.
85. Xu Y, Hou H, Liu Q *et al.* Removal behavior research of orthophosphate by CaFe-layered double hydroxides. *Desalin Water Treat*. 2016;**57**:7918–25.
86. Wang Y, Tang H, Debarry JD *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*. 2012;**40**:e49.
87. Zhang Z, Li J, Zhao XQ *et al.* KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genom. Proteom. Bioinf.*. 2006;**4**:259–63.
88. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 2015;**16**:157.
89. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;**32**: 1792–7.
90. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;**30**:2114–20.
91. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;**25**: 1754–60.
92. Li H, Handsaker B, Wysoker A *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;**25**:2078–9.
93. Cingolani P, Platts A, Wang LL *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118. *Fly*. 2012;**6**:80–92.
94. Purcell S, Neale B, Todd-Brown K *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;**81**:559–75.
95. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;**26**:139–40.
96. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;**19**: 1655–64.
97. Beerenwinkel N, Rahnenfuhrer J, Kaiser R. Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics*. 2005;**21**:2106–7.
98. Zhang C, Dong SS, Xu JY *et al.* PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*. 2019;**35**:1786–8.
99. Danecek P, Auton A, Abecassis G *et al.* The variant call format and VCFtools. *Bioinformatics*. 2011;**27**:2156–8.
100. Vilella AJ, Blanco-Garcia A, Hutter S *et al.* VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics*. 2005;**21**:2791–3.
101. Degiorgio M, Huber CD, Hubisz MJ *et al.* SWEEPFINDER2: increased sensitivity, robustness, and flexibility. *Bioinformatics*. 2016;**32**:1895–7.

Downloaded from https://academic.oup.com/hr/article/doi/10.1093/hr/uhac012/6529164 by guest on 30 July 2026