

Genome-wide association study uncovers genetic architecture of rubber seed oil composition and seed size traits for edible oil improvement

Fazal Rehman

Chinese Academy of Tropical Agricultural Sciences

Tian Yang

Yunnan Institute of Tropical Crops

Tingkai Wu

Chinese Academy of Tropical Agricultural Sciences

Wenguan Wu

Chinese Academy of Tropical Agricultural Sciences

Dong Xu

Chinese Academy of Tropical Agricultural Sciences

Yan Zhou

Yunnan Institute of Tropical Crops

Yu Wu

Yunnan Institute of Tropical Crops

Xiaobo Wang

Chinese Academy of Tropical Agricultural Sciences

Han Cheng

forcheng@gmail.com

Chinese Academy of Tropical Agricultural Sciences

Article

Keywords: RSO, GWAS, caryopsis, fatty acid composition, vegetable oil, candidate genes

Posted Date: September 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10671720/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Rubber seed oil (RSO), extracted from the discarded seeds of *Hevea brasiliensis* plantations, represents a non-competitive vegetable oil resource with established edible uses and a favorable unsaturated fatty acid (UFA) profile, yet the genetic basis of its composition remains uncharacterized. We conducted the first multi-locus genome-wide association study (GWAS) for seed oil content (SOC), fatty acid (FA) composition, and seed size traits across 186 diverse *H. brasiliensis* accessions using three GWAS models (BLINK, FarmCPU, and MLM). UFAs dominated RSO composition (79.51%), with linoleic acid (ω -6, 44.7%) and α -linolenic acid (ω -3, 21.7%) as the most abundant components, and broad-sense heritability exceeded 90% for most traits. Population structure analysis identified four ancestral subpopulations ($K = 4$), supported by cross-validation, phylogenetic analysis, and principal component analysis, and linkage disequilibrium decayed rapidly with a half-decay distance of approximately 50 kb, enabling precise candidate gene localization. Across all traits and models, 2,667 significant marker-trait associations (MTAs) clustered into 672 independent quantitative trait loci (QTLs), distributed across multiple chromosomes, with Chr10 and Chr17 harboring the highest association densities. A pleiotropic locus (17_28516849) on Chr17 simultaneously associated with three FA traits, suggesting a shared regulatory mechanism for fatty acid metabolism. Fine-mapping and haplotype analysis identified five candidate genes—MOD1, SRK2I, KASII, WOX1, and YABBY2—with roles in FA biosynthesis, metabolic signaling, and organ development. Favorable haplotypes for MOD1 (H002), SRK2I (H004), and WOX1 (H003), provide genomic resources for marker-assisted selection targeting improved RSO nutritional quality and seed yield, advancing the development of this underutilized vegetable oil for food and nutraceutical applications.

Introduction

Rubber tree (*Hevea brasiliensis* Müll. Arg.), a member of the Euphorbiaceae family native to the Amazon basin, is cultivated across tropical regions of Asia, Africa, and Latin America as the primary commercial source of natural rubber¹. Beyond latex, rubber plantations generate substantial quantities of seeds as a largely underutilized by-product—mature plantations can yield approximately 228,400 tons of seeds annually in major producing countries, yet less than 25% are currently utilized for propagation purposes, with the majority discarded as plantation waste^{2,3}. The growing global demand for vegetable oils, coupled with increasing competition between food and fuel applications in conventional oilseed crops, has intensified interest in alternative oil sources that do not compete with existing food production systems. Rubber seed oil (RSO) represents a particularly promising candidate in this context—extracted from seeds otherwise discarded as plantation waste, and therefore providing a genuinely non-competitive vegetable oil supplement³. Rubber seeds are rich in oil, with kernel oil content ranging from 40–50% and whole-seed oil content reaching up to 60% by weight^{3,4}. RSO is characterized by a high proportion of unsaturated fatty acids (UFAs)—predominantly linoleic acid (C18:2, ω -6), oleic acid (C18:1), and α -linolenic acid (C18:3, ω -3)—collectively accounting for approximately 79–80% of total fatty acid (FA) content^{2,5}. This FA profile, rich in essential polyunsaturated fatty acids (PUFAs) including both ω -6

and ω -3 that cannot be synthesized *de novo* by humans, confers significant nutritional value to RSO as a dietary oil.

Importantly, RSO has been safely consumed as an edible oil in regions with established culinary traditions, including Hainan Island and Yunnan, China, and research has confirmed its edible, medicinal, and nutritional potential^{6,7}. Clinical evaluations have further demonstrated that RSO, PUFA-rich in ω -3 with a favorable ω -3/ ω -6 ratio, can reduce blood lipid levels and help prevent hyperlipidemia-associated disorders, underscoring its potential as a health-promoting dietary oil comparable in composition to other commercially accepted edible vegetable oils³. RSO additionally contains bioactive compounds including tocopherols and phenolic antioxidants contributing to oxidative stability and nutraceutical potential, with recent studies demonstrating its applicability in functional food and cosmetic formulations^{8,9}. Beyond food applications, RSO also serves as a viable biodiesel feedstock, positioning it as a dual-purpose renewable resource with both food and non-food value streams within existing rubber plantation systems^{3,10}. The FA composition of RSO spanning SFAs (C14:0, C16:0, C18:0, C20:0), MUFAs (C18:1, C20:1), and PUFAs (ω -6, ω -3), is a primary determinant of oil quality for both food and industrial end uses. However, the genetic factors controlling individual FA accumulation in rubber seeds remain largely unexplored, limiting opportunities for targeted breeding improvement³.

Over the past decade, substantial progress has been made in rubber tree genomics, with high-density SNP platforms enabling QTL identification for agronomic and economic traits^{11–13}. GWAS utilizing NGS-derived SNP arrays have further facilitated candidate gene discovery for latex production traits^{14,15}, and high-quality chromosome-level genome assemblies and pangenome resources now provide a robust foundation for precision association mapping^{16,17}. Despite these advances, the genetic regulation of seed-derived metabolites, including lipid biosynthesis and FA accumulation remains uncharacterized in rubber tree. As highlighted³, integrative genomic approaches represent the most promising route for elucidating RSO biosynthetic regulation and enabling precision breeding for improved oil traits, yet genome-wide association mapping as a foundational step has not previously been applied to RSO composition.

In other oilseed crops, multi-locus GWAS has successfully identified candidate genes controlling FA biosynthesis, including KAS, SAD, and FAD family enzymes enabling marker-assisted improvement of oil quality relevant to both food and industrial applications^{18–20}. Multiple statistical GWAS models have demonstrated particular utility for dissecting polygenic metabolic traits by improving detection power while reducing false positive rates²¹. CRISPR/Cas9-mediated editing of GWAS-identified genes such as FAD2, controlling oleic to linoleic acid conversion, has further validated the utility of association mapping for precision oil quality improvement^{22,23}. Moreover, seed size traits including seed length, width, thickness, fruit and caryopsis weight are also important as primary determinants of seed biomass yield and oil extraction efficiency, with CALAVAT3 gene²⁴, WUSCHEL²⁵ and YABBY transcription factors (TFs)^{26,27} identified as key regulators of organ size and lateral organ elongation in multiple crop species. However, the genomic loci controlling seed size traits in rubber tree have not previously been

characterized. The present study addresses these existing gaps by conducting the comprehensive multi-locus GWAS for rubber seed oil content (SOC), FA composition, and seed size traits across 186 *H. brasiliensis* accessions using three GWAS models (BLINK, FarmCPU, and MLM). Candidate genes at significant loci were identified through LD-based fine-mapping and haplotype analysis of non-synonymous coding variants. The findings establish the first genomic framework for RSO composition in rubber tree, identifying candidate genes relevant to FA biosynthesis and seed development, and providing a foundation for molecular breeding strategies targeting improved RSO nutritional quality and seed yield for food, nutraceutical, and industrial applications.

Results

Phenotypic variation and descriptive statistics of SOC, FA composition, and seed size traits

To comprehensively elucidate the genetic basis of seed quality and fruit development in rubber tree, three categories of traits were evaluated across 186 accessions: (i) SOC; (ii) FA composition traits—myristic acid (C14:0), palmitic acid (C16:0), margaric acid (C17:0), stearic acid (C18:0), arachidic acid (C20:0), palmitoleic acid (C16:1), oleic acid (C18:1), linoleic acid (C18:2), α -linolenic acid (C18:3), and eicosenoic acid (C20:1); and (iii) seed size traits—seed length (SL), seed width (SW), seed thickness (ST), fruit weight (FW), and caryopsis weight (CW). The FA profile of RSO comprises both SFAs and UFAs, including MUFAs and PUFAs. ANOVA revealed significant genotypic differences among all sixteen traits. Broad-sense heritability (H^2) estimates exceeded 90% for SOC and most FA compositions, with the exception of C18:3 ($H^2 = 52\%$), indicating that FA composition in RSO is predominantly under genetic control with limited environmental influence. Similarly, seed size traits exhibited high heritability estimates exceeding 80%, suggesting strong genetic determination (Table S1). The coefficient of variation (CV%) ranged from 6.3% to 78% across all traits, reflecting substantial phenotypic diversity within the panel suitable for GWAS (Table S2). Frequency distribution histograms and boxplots across individual environments demonstrated approximately normal distributions with clear phenotypic variation among accessions for all traits (Fig. 1, S1A). Normality was further assessed using the K–S test across all traits, confirming distributional assumptions required for GWAS analysis (Table S3). The mean SOC was 41.35%, with UFAs collectively accounting for 79.51% of total FA content, substantially exceeding the SFA fraction (20.49%). Among individual FAs, ω -6 was the most abundant (44.7%), followed by ω -3 (21.7%) and oleic acid (12.2%), indicating that RSO is predominantly an ω -6-rich oil (Fig. S1B, Table S4). Descriptive statistics for seed size traits indicated 26.2, 22.06, 18.65, 4.25, and 2.54 for SL, SW, ST, FW, and CW, respectively, with SL showing the greatest phenotypic range among accessions (Table S2).

Pearson's correlation analysis revealed complex and biologically meaningful relationships among SOC, FA composition, and seed size traits (Fig. S1C; Table S5). SOC exhibited significant negative correlations with several FA classes, including C17:0 ($r = -0.63$), C16:1 ($r = -0.64$), C14:0 ($r = -0.56$), C20:1 ($r = -$

0.52), and C20:0 ($r = -0.43$), suggesting metabolic competition between oil accumulation and the biosynthesis of specific FA classes. Correlations of SOC with C16:0 ($r = -0.049$) and C18:3 ($r = -0.052$) were negligible. SOC also showed significant negative correlations with seed size traits, including SL ($r = -0.27$), FW ($r = -0.22$), and CW ($r = -0.23$), indicating an inverse relationship between seed size and oil accumulation. Within FA classes, strong positive co-associations among SFAs and MUFAs were observed, particularly between C14:0 and C17:0 ($r = 0.94$), C14:0 and C16:1 ($r = 0.78$), and C14:0 and C20:0 ($r = 0.49$), suggesting coordinated regulation within the SFA biosynthetic pathway. MUFAs similarly showed strong positive co-associations with SFAs, including C16:1 with C17:0 ($r = 0.91$), C20:1 with C17:0 ($r = 0.88$), C20:1 with C14:0 ($r = 0.83$), and C20:1 with C20:0 ($r = 0.76$). In contrast, PUFAs exhibited negative associations with most SFAs and MUFAs; C18:3 showed particularly strong negative correlations with C18:2 ($r = -0.67$) and C18:1 ($r = -0.49$), reflecting competitive metabolic flux between FA desaturation pathways. Seed size traits were strongly intercorrelated, with SL positively associated with ST ($r = 0.85$), FW ($r = 0.80$), CW ($r = 0.75$), and SW ($r = 0.65$). Similarly, strong positive associations were observed among all pairwise combinations of SW, ST, FW, and CW, consistent with coordinated fruit development. C16:1 showed a significant positive association with FW ($r = 0.19$) and CW ($r = 0.22$), suggesting a possible link between fruit development and MUFA accumulation (Fig. S1C, Table S5). These correlation patterns collectively indicate coordinated metabolic regulation within FA biosynthetic pathways.

PCA was performed to further characterize multivariate relationships among all traits. The first five PCs had eigenvalues greater than 1 and collectively explained 81.80% of total phenotypic variation (Fig. S2A, Table S6). PC1 and PC2 accounted for 32.45% and 22.11% of total variance, respectively, together explaining 54.56% of variation and capturing the primary axes of trait differentiation. PC3, PC4, and PC5 contributed an additional 12.90%, 7.80%, and 6.54% of variance, respectively. The PCA biplot revealed clear trait clustering consistent with FA compositions and seed size grouping (Fig. S2B, C). PC1 separated SFAs, selected MUFAs (C16:1, C20:1), and seed size traits—all positioned on the positive axis—from SOC and C18:2 on the negative axis, consistent with the negative correlations observed between SOC and these FA traits. PC2 further differentiated FA desaturation patterns, with C18:3 oriented in the negative direction, reflecting its distinct metabolic behavior relative to other UFA. The close angular proximity of SFA and MUFA vectors, alongside seed size trait, confirmed their strong positive inter-correlations and suggested possible co-regulation within shared developmental and biosynthetic networks. Accessions 171, 181, 184, 185, and 139 clustered along the positive PC1 axis, indicating relatively higher SFA levels and larger seed size traits in these genotypes, whereas accessions positioned toward the SOC and C18:2 vectors showed favorable oil content with PUFA-rich profiles (Fig. S2B). Overall, PCA effectively captured the major sources of phenotypic variation across all traits, consistent with correlation results, providing a multivariate framework for interpreting the metabolic and developmental relationships underlying seed oil biosynthesis and seed morphology in rubber tree.

Population structure and genetic diversity analysis of Hevea population

To characterize the genetic structure of rubber tree accessions, a comprehensive population structure analysis was conducted using model-based clustering (ADMIXTURE), phylogenetic analysis, and PCA based on 16,629,407 high-quality SNPs retained after quality control filtering. Cross-validation (CV) error analysis across $K = 1$ to 6 ancestral populations revealed a clear minimum at $K = 4$ (CV error = 0.384), providing quantitative support for four ancestral populations as the optimal representation of genetic structure (Fig. 2A). ADMIXTURE analysis across $K = 2$ to $K = 5$ progressively resolved increasing substructure, with $K = 4$ yielding the most interpretable and parsimonious population partitioning (Fig. 2B). At $K = 2$, the panel was broadly partitioned into two ancestral groups with a visible admixture transition zone, while $K = 3$ revealed an additional distinct cluster concentrated among a subset of accessions. At the optimal $K = 4$, individuals were classified into four genetically distinct clusters: Cluster 1 (C1, $n = 26$), Cluster 2 (C2, $n = 56$), Cluster 3 (C3, $n = 40$), and Cluster 4 (C4, $n = 48$) (Fig. 2B). Varying degrees of admixture were observed between adjacent clusters, particularly between C2 and C3, suggesting historical gene flow among ancestral populations. At $K = 5$, existing clusters became increasingly fragmented without resolving clearly distinct new subgroups. The convergence of visual interpretability and minimum CV error at $K = 4$ jointly supported this as the optimal number of ancestral populations. Phylogenetic analysis using neighbor-joining tree independently corroborated the four-cluster structure, with four major clades corresponding closely to the ADMIXTURE-defined clusters at $K = 4$ (Fig. 2C). The concordance between model-based clustering and phylogenetic topology provides mutual validation of the population subdivision and confirms that the identified clusters reflect genuine genealogical relationships rather than statistical artifacts of the ADMIXTURE model. PCA of genome-wide SNP data further supported the population structure findings, with individual accessions distributing along cluster boundaries broadly consistent with ADMIXTURE and phylogenetic results (Fig. 2D). The first three PCs explained 7.11%, 5.66%, and 5.05% of total genetic variance, respectively, accounting for 17.82% of genomic variation. The relatively low variance explained per component is consistent with the high genetic diversity and complex admixture history typical of rubber tree germplasm, where no single axis captures a dominant proportion of variation. Despite the low per component variance, the distribution of accessions along PC1–PC3 axes was broadly concordant with the four ADMIXTURE-defined clusters, supporting the robustness of the identified population structure.

The Q matrix at $K = 4$, and the first three PCs were incorporated as fixed-effect covariates in GWAS models to control for population structure. LD decay was estimated across all chromosomes to characterize the genomic resolution of association mapping and to determine an appropriate physical distance threshold for independent loci. LD decayed rapidly with increasing physical distance in the rubber tree panel, with the maximum mean R^2 of 0.461 observed at the shortest inter-SNP distance of 1 bp, declining to 0.103 at 500 kb (Fig. 2E). The half-decay distance (LD_{50}), defined as the physical distance at which mean R^2 declines to half its maximum value, was estimated at approximately 50 kb. This relatively rapid LD decay reflects the high genetic diversity and complex admixture history of the rubber tree germplasm evaluated, consistent with the four-cluster population structure. The adequacy of population structure correction in GWAS models was subsequently evaluated through quantile–quantile (QQ) plots for each trait, as described in next section.

GWAS identifies significant marker trait associations for SOC, FA compositions, and seed size traits

To dissect the genetic architecture underlying SOC, FA composition, and seed size traits, a multi-locus GWAS was conducted using three models, including BLINK, FarmCPU, and MLM across sixteen traits. Manhattan plots and quantile–quantile (QQ) plots were generated for all traits across the three models to visualize association signals and assess the distribution of observed versus expected P -values. QQ plots indicated that observed P -values followed the expected null distribution across the majority of SNPs, with deviation occurring only at the upper tail corresponding to genuine association signals, suggesting adequate control of population structure (Fig. S3). In total, 1,328 significant marker-trait associations (MTAs) clustered into 629 independent QTLs based on physical proximity (within 50 kb, consistent with estimated LD decay) were identified across all traits and models combined.

Seed oil content (SOC)

GWAS for SOC identified sixteen significant MTAs across eleven genomic regions ($P = 3.29 \times 10^{-8}$ to 1.08×10^{-7}). Of these, thirteen MTAs representing nine QTLs were consistently detected by both BLINK and FarmCPU, while three additional MTAs were uniquely identified by MLM (Table S7). The strongest association was observed at SNP (10_82959143) on Chr10 ($P = 1.87 \times 10^{-8}$), mapping to *qSOC-10.2*. Chr10 harbored the highest density of association signals, with four independent QTLs identified (*qSOC-10.1-qSOC-10.4*). Two SNPs (10_92190334 and 10_92190366) at positions 92,190,334 and 92,190,366 bp on Chr10 were consistently detected by both models with identical P -values ($P = 1.09 \times 10^{-7}$), representing the most reproducible association signal for SOC. Notably, *qSOC-10.4* mapped within the 92.14–92.24 Mb interval on Chr10 (Table S7), in proximity to *SnRK1*, a gene encoding sucrose non-fermenting 1-related protein kinase previously implicated in seed oil biosynthesis regulation.

Myristic acid (C14:0): A total of 159 significant MTAs associated with 115 QTLs were identified across the three models ($P = 9.71 \times 10^{-8}$ to 1.75×10^{-30}). Among these, 37 MTAs representing 28 QTLs were detected by at least two models, of which 20 MTAs were consistently identified across all three models (BLINK, FarmCPU, and MLM), 12 MTAs were shared between BLINK and FarmCPU, and five MTAs between BLINK and MLM (Table S7). The peak MTA was SNP (17_28516849) on Chr17 ($P = 1.75 \times 10^{-30}$), mapping to *qC14:0-17.2* and detected consistently across all three models, representing the strongest single association signal identified (Table S7).

Palmitic acid (C16:0): Twenty-five significant MTAs associated with 22 QTLs were identified ($P = 8.73 \times 10^{-8}$ to 6.32×10^{-11}), with 24 MTAs and 22 QTLs consistently detected across multiple models. The peak MTA was SNP 14_20935205 on Chr14 ($P = 6.32 \times 10^{-11}$), mapping to *qC16:0-14.1* and detected across all three models. Of the 24 multi-model MTAs, three were identified by all three models and 21 were shared between BLINK and FarmCPU (Table S7).

Margaric acid (C17:0): A total of 134 MTAs clustered into 103 QTLs were identified across the three models ($P = 9.77 \times 10^{-9}$ to 1.48×10^{-27}), with 54 MTAs and 44 QTLs detected across multiple models.

Among these, 21 MTAs were consistently identified across all three models, 25 MTAs were shared between BLINK and FarmCPU, and eight between BLINK and MLMM (Table S7). The peak MTA was SNP (17_28516849) on Chr17 ($P = 1.48 \times 10^{-27}$), mapping to *qC17:0-17.2* and detected across all three models. Two QTL clusters on Chr17 (*qC17:0-17.1* and *qC17:0-17.2*) harbored associations detected by all three and two models (BLINK-FarmCPU) simultaneously, and Chr01, Chr10, Chr11, and Chr17 harbored the highest density of association signals (Table S7). Notably, SNP (17_28516849) was also identified as the peak MTA for C14:0 (see above), suggesting a pleiotropic locus on Chr17 with effects on multiple SFAs classes.

Stearic acid (C18:0): Twenty-three MTAs clustered into 19 QTLs were identified across the three models ($P = 1.13 \times 10^{-7}$ to 8.90×10^{-8} , with 21 MTAs and 19 QTLs consistently detected across multiple models. The peak MTA was SNP (7_92539132) on Chr07 ($P = 8.35 \times 10^{-10}$), mapping to *qC18:0-7.1* and detected across all three models, while 20 additional associations were co-detected by BLINK and FarmCPU (Table S7). Chr04 and Chr07 harbored the highest density of association signals. Notably, *qC18:0-4.1* mapped within the 22.31–22.41 Mb interval on Chr04 (Table S7), in proximity to MOD1, a gene encoding a stearyl-ACP desaturase with established roles in FA desaturation and seed oil composition.

Arachidic acid (C20:0): A total of 205 MTAs clustered into 138 QTLs were identified ($P = 9.84 \times 10^{-9}$ to 2.05×10^{-20}), with 58 MTAs and 51 QTLs detected across multiple models. Among these, 18 MTAs were consistently identified by all three models, 35 MTAs were shared between BLINK and FarmCPU, and three MTAs between BLINK and MLMM (Table S7). The peak MTA was SNP (10_3603352) on Chr10 ($P = 2.05 \times 10^{-20}$), mapping to *qC20:0-10.1* and detected across all three models. Chr08 and Chr13 harbored the highest association density. Two SNPs on Chr16 (16_6880730 and 16_6880732) were reproducibly detected by BLINK and FarmCPU with identical P -values ($P = 1.63 \times 10^{-7}$), representing a highly reproducible signal within this region (Table S7).

Palmitoleic acid (C16:1): Seventy-five MTAs clustered into 45 QTLs were identified across the three models ($P = 9.94 \times 10^{-8}$ to 6.20×10^{-14}), with 35 MTAs and 32 QTLs detected across multiple models. Of these, eight MTAs were consistently identified by all three models and 27 MTAs were shared between BLINK and FarmCPU (Table S7). The peak MTA was SNP (13_19904928) on Chr13 ($P = 6.20 \times 10^{-14}$), mapping to *qC16:1-13.2*, with Chr08 harboring the highest overall density of association signals (Table S7). Notably, *qC16:1-8.2* mapped within the 4.22–4.32 Mb interval on Chr08 (Table S7), in proximity to BPM (BTB/POZ domain and MATH domain-containing protein 3), play a role in seed development and metabolic regulation.

Oleic acid (C18:1): Twenty-nine significant MTAs clustered into 23 QTLs were identified ($P = 9.59 \times 10^{-10}$ to 1.10×10^{-12}), with seven MTAs and five QTLs detected across multiple models. The peak MTA was SNP (3_79942628) on Chr03 ($P = 1.10 \times 10^{-12}$), mapping to *qC18:1-3.2* and detected across all three models. Among multi-model MTAs, three associations were co-detected by all three models, and one association each was shared between BLINK–FarmCPU, FarmCPU–MLMM, and BLINK–MLMM, respectively (Table S7). Notably, *qC18:1-6.4* mapped within the 95.95–96.05 Mb interval on Chr06 (Table S7), in proximity to

KASII (β -ketoacyl-ACP synthase II), an enzyme that catalyzes the rate-limiting elongation of palmitic acid to stearic acid within the plastidial fatty acid synthesis cycle.

Eicosenoic acid (C20:1): A total of 619 MTAs clustered into 122 QTLs were identified ($P = 9.97 \times 10^{-10}$ to 1.80×10^{-22}), with 55 MTAs and 24 QTLs detected across multi-models. Among these, 18 MTAs were consistently identified by all three models, 34 MTAs were shared between BLINK and FarmCPU, and three MTAs between BLINK and MLMM (Table S7). The peak MTA was SNP (17_28516849) on Chr17 ($P = 1.80 \times 10^{-22}$), mapping to *qC20:1-17.2* and detected consistently across all three models. Chr07 harbored the highest association density. Importantly, SNP (17_28516849) was also identified as the peak MTA for C14:0 and C17:0, further supporting a pleiotropic role for this Chr17 locus in regulating FA compositions, potentially through a shared regulatory mechanism in the FA elongation pathway (Table S7).

Linoleic acid (C18:2): Six MTAs clustered into five QTLs were identified across the three models ($P = 3.64 \times 10^{-7}$ to 2.15×10^{-8}), with five MTAs/QTLs detected by multi-models. The peak MTA was SNP (10_10389795) on Chr10 ($P = 2.15 \times 10^{-8}$), mapping to *qC18:2-10.2* and detected by BLINK and FarmCPU (Table S7). C18:2 yielded the fewest significant MTAs among all FA traits evaluated, consistent with its complex genetic architecture involving multiple desaturation steps.

Linolenic acid (C18:3): Twenty-five MTAs clustered into 20 QTLs were identified ($P = 1.25 \times 10^{-15}$ to 9.49×10^{-15}), with 16 MTAs and 11 QTLs consistently detected across multi-models, of which 15 MTAs were shared between BLINK and FarmCPU (Table S7). The peak MTA was SNP (3_12763257) on Chr03 ($P = 9.49 \times 10^{-15}$), mapping to *qC18:3-3.1* and detected across all three models. Chr05 and Chr13 harbored the highest density of association signals (Table S7).

Seed size traits

GWAS for the five seed size traits collectively identified 1,339 significant associations across 43 genomic regions, distributed across multiple chromosomes. Manhattan and QQ plots were generated for all seed size traits across the three GWAS models to visualize association signals and evaluate the distribution of observed versus expected P -values. QQ plots indicated that observed P -values conformed to the expected null distribution across the majority of SNPs, with deviation evident only at the upper tail representing genuine association signals, suggesting that the incorporated covariates adequately controlled for population structure in the seed size trait analyses (Fig. S4). For seed length (SL), four significant MTAs across four genomic regions were identified ($P = 7.43 \times 10^{-8}$ to 1.83×10^{-7}), with two MTAs consistently detected across multi-models. The peak MTA was SNP (8_17138757) on Chr08 ($P = 7.43 \times 10^{-8}$), mapping to *qSL-8.1* (Table S8). Notably, *qSL-1.1* mapped within the 22.42–22.52 Mb interval on Chr01 (Table S8), in proximity to YABBY2 TF of the plant-specific YABBY gene family with established roles in fruit size and lateral organ elongation, directly relevant to the seed length. For seed width (SW), 38 significant MTAs across eight genomic regions were identified ($P = 1.09 \times 10^{-7}$ to 4.19×10^{-11}), with two MTAs consistently detected across BLINK and MLMM. The peak MTA was SNP (7_24895895) on Chr07 ($P = 3.37 \times 10^{-8}$), mapping to *qSW-7.1*. Notably, SW was the only trait for which FarmCPU did not identify

significant associations; all 38 MTAs were detected by BLINK and MLM, which may reflect differences in model sensitivity to the genetic architecture of this trait (Table S8). For seed thickness (ST), 1,278 significant MTAs across fifteen genomic regions were identified ($P = 1.05 \times 10^{-7}$ to 7.96×10^{-14}), with four MTAs consistently detected across multiple models. The peak MTA was SNP (13_58500532) on Chr13 ($P = 7.96 \times 10^{-14}$), mapping to *qST-13.2* and detected across all three models. Chr09 and Chr13 harbored the highest density of association signals, with two independent QTLs on Chr09 (*qST-9.1* and *qST-9.2*). Notably, significant *qST-13.1* ($P = 1.85 \times 10^{-9}$) mapped within the 3.64–3.74 Mb interval on Chr13 (Table S8), in proximity to WUSCHEL, a homeodomain TF with established roles in stem cell maintenance and meristematic cell size regulation. For fruit weight (FW), six significant MTAs across six genomic regions were identified ($P = 2.21 \times 10^{-7}$ to 4.97×10^{-10}), with four MTAs consistently detected across multi-models. The peak MTA was SNP (18_45325224) on Chr18 ($P = 4.97 \times 10^{-10}$), mapping to *qFW-18.1*. Chr02 harbored the highest association density, with two independent QTLs identified (*qFW-2.1* and *qFW-2.2*) (Table S8). For caryopsis weight (CW), 13 significant MTAs across ten genomic regions were identified ($P = 1.51 \times 10^{-7}$ to 1.64×10^{-11}), with five MTAs consistently detected across multi-models. The peak MTA was SNP 1_66289757 on Chr01 ($P = 1.64 \times 10^{-11}$), mapping to *qCW-1.1*. Chr11 harbored the highest association density, with two independent QTLs identified (*qCW-11.1* and *qCW-11.2*) (Table S8).

Fine-mapping and candidate gene identification at key GWAS loci for SOC, FA composition, and seed size traits

To identify candidate genes underlying the significant GWAS associations, high-confidence QTLs detected by multiple models were selected for fine-mapping analysis. Candidate genes within each QTL interval were identified based on gene annotation within LD blocks defined for each lead SNP, with priority given to genes harboring non-synonymous coding variants. Haplotype analysis was subsequently performed for the most biologically relevant candidate gene within each LD block to assess the functional significance of identified variants. Here we report the fine-mapping results for five representative loci spanning SOC, FA composition, and seed size traits.

qC18:0-4.1—MOD1 (HbAchr4G0026485), stearic acid: The QTL *qC18:0-4.1*, with lead SNP (4_22355427) on Chr04, was prioritized for candidate gene analysis based on its consistent detection across BLINK and FarmCPU models and the strong biological relevance of the annotated candidate gene to FA biosynthesis (Fig. 3A). Regional fine-mapping of the associated interval on Chr04 revealed a defined LD block spanning approximately 99 kb (22,306,000–22,405,000 bp), with the peak association signal localizing near position 22.35 Mb (Fig. 3B). Gene annotation within this LD block identified five candidate genes: HbAchr4G0026482 (ARF1: auxin response factor 1), HbAchr4G0026483 (NSE4A: non-structural maintenance of chromosomes element 4 homolog A), HbAchr4G0026484 (ARF1: auxin response factor 1), HbAchr4G0026485 (MOD1: enoyl-[acyl-carrier-protein] reductase [NADH], chloroplastic), and HbAchr4G0026486 (At1g51120: AP2/ERF and B3 domain-containing transcription factor). Among these, HbAchr4G0026485 (MOD1) was prioritized as the most likely causal gene based on its established enzymatic role in FA biosynthesis—specifically, the NADH-dependent reduction of enoyl-ACP, a key step in the plastidial fatty acid synthesis cycle directly relevant to C18:0 accumulation. Fine mapping of the

MOD1 CDS revealed seven non-synonymous missense variants clustered within a narrow 164 bp window near 22,389,000 bp on the negative strand of Chr04: T > G (p.Ala159Thr) near 22,388,000 bp, and C > T (p.Glu83Lys), G > A (p.Ser86Phe), G > T (p.Gln54Lys), C > T (p.Arg37Lys), G > A (p.Ala12Val), and A > G (p.Val5Ala) near 22,389,000 bp (Fig. 3C). The tight clustering of seven missense variants within a short coding region suggests a hypervariable functional domain potentially subject to diversifying selection in this rubber tree germplasm panel. Genotype–phenotype analysis of the lead SNP (4_22355427) revealed three observed genotypes (AA, AG, and GG), with a significant difference in stearic acid content between AA and AG genotypes ($P = 0.0003$), with AG heterozygous accessions exhibiting higher stearic acid content than AA homozygous reference accessions (Fig. 3D). The GG homozygous alternative genotype was observed in very few accessions and was therefore insufficient for reliable statistical comparison. The seven missense variants defined seven haplotypes in the panel, with H001 ($n = 127$) and H002 ($n = 29$) occurring at the highest frequencies (Fig. 3E). Pairwise haplotype–phenotype association analysis revealed a significant difference in stearic acid content between H002 and H001 ($P = 0.043$), confirmed by both the association matrix (Fig. 3F) and violin plot comparison (Fig. 3G). H002 accessions exhibited consistently higher stearic acid levels relative to H001 accessions, with the H002 distribution showing a broader range extending to about 15% stearic acid compared to the more concentrated lower-value distribution of H001. These results collectively support MOD1 as a functional candidate gene at *qC18:0-4.1*, with the H002 haplotype—defined by a specific combination of seven non-synonymous coding variants—contributing to elevated stearic acid accumulation in RSO.

qSOC-10.4–SRK2I (HbAchr10G0003664), seed oil content: The QTL *qSOC-10.4*, with lead SNP (10_92190334) on Chr10, was selected for candidate gene analysis based on its reproducible detection across BLINK and FarmCPU models and the biological plausibility of the annotated candidate gene in seed oil regulation (Fig. 4A). Regional fine-mapping of the associated interval on Chr10 revealed a moderately structured LD block spanning approximately 99.94 kb (92,140,000–92,240,000 bp), with the association signal localizing near position 92.19 Mb (Fig. 4B). Gene annotation within the LD block identified eight genes: HbAchr10G0003662 (LBD36: LOB domain-containing protein 36), HbAchr10G0003663 (HHO5: transcription factor HHO5), HbAchr10G0003664 (SRK2I: serine/threonine-protein kinase SRK2I), HbAchr10G0003665 (At5g66900: probable disease resistance protein), HbAchr10G0003666 (protein of unknown function), HbAchr10G0003667 (PCMP-H5: pentatricopeptide repeat-containing protein At4g37170), HbAchr10G0003668 (protein of unknown function), and HbAchr10G0003669 (RTNLB13: reticulon-like protein B13). Among these, HbAchr10G0003664 (SRK2I) was prioritized as the most biologically relevant candidate based on its documented role in abscisic acid signaling and carbon partitioning pathways that regulate seed oil accumulation during seed filling. Fine-mapping of the SRK2I CDS on the positive strand of Chr10 identified two non-synonymous missense variants: G > C (p.Asp2His) at position 92,157,053 bp and A > C (p.Gln272His) at position 92,162,082 bp (Fig. 4C). These two variants—located at opposite ends of the gene spanning approximately 92,156,000–92,163,000 bp—together define the haplotype structure at this locus, with the substitution of aspartic acid with histidine at position 2 and glutamine with histidine at position 272 potentially altering kinase domain activity or substrate recognition. Genotype–phenotype analysis of the lead SNP revealed two

observed genotype classes (TG and TT) with a highly significant association between genotypes ($P < 0.0001$), with TT homozygous accessions exhibiting substantially higher SOC relative to TG heterozygous accessions (Fig. 4D).

The two missense variants defined seven haplotypes in the panel, with H001 ($n = 52$), H002 ($n = 50$), H003 ($n = 35$), and H004 ($n = 18$) meeting the minimum frequency threshold for association testing (Fig. 4E). Pairwise haplotype–phenotype association analysis among the four retained haplotypes revealed a significant difference in SOC specifically between H004 and H001 ($P = 0.019$), while comparisons involving H002 and H003 did not reach significance (Fig. 4F). Violin plot visualization confirmed that H004 accessions exhibited higher SOC values with a broader distribution extending toward about 60%, compared to H001, consistent with the direction of the SNP-level genotypic association (Fig. 4G). These findings suggest that the H004 haplotype—defined by the C allele at position 92,157,053 and the heterozygous A|C combination at position 92,162,082—is associated with elevated SOC in rubber tree, identifying H004 as a potentially favorable haplotype for MAS in SOC breeding programs.

qC16:1-8.1—BPM3 (HbAchr8G0033831), palmitoleic acid: The QTL *qC16:1-8.2*, with lead SNP (8_4266243) on Chr08, was investigated for candidate gene identification based on its significant SNP-level association with palmitoleic acid content and the functional annotation of genes within the defined LD block (Fig. S5A). Regional fine-mapping of the associated interval (4.22–4.32 Mb) on Chr08 revealed a moderately structured LD block beneath the peak association signal, within which gene annotation identified HbAchr8G0033831 as the prioritized candidate gene (Fig. S5B). This gene encodes BPM3 (BTB/POZ domain and MATH domain-containing protein 3), a substrate adaptor component of CUL3-based E3 ubiquitin ligase complexes implicated in targeted protein degradation during seed development and metabolic regulation. Fine-mapping of the BPM3 identified one non-synonymous missense variant—T > G (p.Glu187Ala)—located near position 4,231,000 bp on the third exon of the gene, where substitution of glutamic acid with alanine may alter protein folding or substrate interaction capacity (Fig. S5C). Genotype–phenotype analysis of the lead SNP revealed a significant association with palmitoleic acid content between TC and TT genotype classes ($P = 0.0100$), with TC heterozygous accessions exhibiting higher palmitoleic acid levels than TT homozygous accessions (Fig. S5D).

qC18:1-6.4—KASII (HbAchr6G0031475), oleic acid: The QTL *qC18:1-6.4*, with lead SNP (6_96004035) on Chr06, was selected for candidate gene investigation based on its significant association with oleic acid content and the biological relevance of the annotated gene within the defined LD block (Fig. S6A). Regional fine-mapping of the associated interval (95.95–96.05 Mb) on Chr06 revealed one of the strongest and most extensive LD block structures observed, spanning the KASII gene region and supporting confident localization of the association signal to this interval (Fig. S6B). Gene annotation within the LD block identified HbAchr6G0031475 as the most biologically relevant candidate, encoding KASII (β -ketoacyl-ACP synthase II), an enzyme that catalyzes the rate-limiting elongation of palmitic acid (C16:0) to stearic acid (C18:0) within the plastidial fatty acid synthesis cycle, directly regulating the substrate pool available for subsequent desaturation to oleic acid by stearyl-ACP desaturase. Fine-

mapping of the KASII CDS identified one non-synonymous missense variant—C > G (p.Thr463Arg)—located near position 96,011,000 bp within the multi-exon gene body on the positive strand of Chr06, where substitution of threonine with arginine at position 463 may alter the catalytic efficiency or substrate specificity of the KASII active site (Fig. S6C). Genotype–phenotype analysis of the lead SNP revealed a significant association with oleic acid content between AA and AG genotypes ($P = 0.0041$), with AA homozygous accessions exhibiting higher oleic acid content relative to AG heterozygous accessions, suggesting that the minor G allele at this locus is associated with reduced oleic acid accumulation (Fig. S6D).

qST-13.1–WOX1 (HbAchr13G0007531), seed thickness: The QTL *qST-13.1*, with SNP (13_3692007) on Chr13 ($P = 1.85 \times 10^{-9}$), detected across FarmCPU model, represented the significant association signal corresponding to ST and was selected for detailed candidate gene analysis (Fig. 5A). Regional fine-mapping of the associated interval on Chr13 revealed a moderately defined LD block spanning approximately 99.86 kb (3,642,000–3,742,000 bp), with the peak association signal localizing near position 3.69 Mb (Fig. 5B). Gene annotation within the LD block identified five candidate genes: HbAchr13G0007530 (TUBB: tubulin beta chain), HbAchr13G0007531 (WOX1: WUSCHEL-related homeobox 1), HbAchr13G0007532 (At1g18250: thaumatin-like protein), HbAchr13G0007533 (At3g18020: pentatricopeptide repeat-containing protein At3g18020), and HbAchr13G0007534 (ZEU1: thymidylate kinase). Among them, HbAchr13G0007531 was prioritized as the most biologically relevant candidate based on its established roles in lateral organ size determination and cell proliferation regulation through interaction with CLAVATA signaling components directly relevant to seed thickness development.

Fine-mapping of the WOX1 on the positive strand of Chr13 identified six non-synonymous missense variants distributed across two genomic clusters spanning approximately 3,705,000–3,708,000 bp (Fig. 5C). The first cluster contained one variant near 3,705,500 bp: C > A (p.His51Asn). The second cluster near 3,707,000–3,707,500 bp contained five additional variants including C > G (p.Thr257Ser), G > A (p.Ala253Thr), A > G (p.Lys295Glu), and additional substitutions, where the combination of amino acid changes across these positions may cumulatively alter WOX1 transcriptional regulatory activity governing seed development programs. Genotype–phenotype analysis of the lead SNP revealed three genotypes (AA, GA, and GG) with significant associations between AA and GA genotypes ($P = 0.0029$) and between AA and GG genotypes ($P = 0.0035$), while the GA–GG comparison was non-significant ($P = 0.4175$) (Fig. 5D). AA homozygous accessions exhibited the highest ST values, followed by GA heterozygous and GG homozygous accessions, suggesting that the reference A allele at this locus is associated with greater seed thickness in an additive manner. The six missense variants defined ten haplotypes in the panel, of which H001 ($n = 90$), H002 ($n = 29$), H003 ($n = 22$), and H004 ($n = 20$) allocated for association analysis (Fig. 5E).

Pairwise haplotype–phenotype association analysis among the four retained haplotypes revealed a highly significant difference in seed thickness between H003 and H001 ($P = 0.003$), while all other pairwise comparisons did not reach statistical significance (Fig. 5F). Violin plot visualization confirmed

that H003 accessions exhibited higher seed thickness relative to H001 accessions, with the H003 distribution extending toward higher values compared to the more concentrated lower distribution of H001 (Fig. 5G). These findings identify the H003 haplotype of WOX1 as significantly associated with elevated seed thickness in rubber tree, providing a genomic resource for marker-assisted selection targeting improved seed size characteristics.

qSL-1.1–YABBY2 (HbAchr1G0000732), seed length: The QTL *qSL-1.1*, with lead SNP (1_22473802) on Chr01 was investigated for candidate gene identification based on its significant SNP-level association with SL and the positional and functional relevance of the annotated candidate gene within the LD block (Fig. S7A). Regional fine-mapping of the associated interval (22.42–22.52 Mb) on Chr01 revealed a moderately structured LD block beneath the peak association signal, within which gene annotation identified HbAchr1G0000732 as the prioritized candidate gene (Fig. S7B). This gene encodes YABBY2, a TF of the plant-specific YABBY gene family with established roles in abaxial cell fate determination and lateral organ elongation, may directly relevant to the developmental processes controlling seed length along the proximal-distal axis. Fine-mapping of the YABBY2 identified one non-synonymous variant: G > C (p.Ala164Gly), located within the conserved HMG-box DNA-binding domain of the encoded protein, where amino acid substitution may alter transcriptional regulatory activity governing organ elongation (Fig. S7C). Genotype–phenotype analysis of the lead SNP revealed a significant difference in SL between AA and AT genotypes ($P = 0.0000$), with the AT heterozygous genotype associated with altered SL relative to the AA homozygous genotype (Fig. S7D).

Discussion

In this study, we employed a multi-model GWAS strategy integrating BLINK, FarmCPU, and MLM to dissect the genetic architecture underlying seed oil content, fatty acid composition, and seed size traits across a diverse rubber tree germplasm panel. The broad phenotypic variation observed for SOC, FA composition, and seed size traits across the accession panel indicates the presence of substantial genetic diversity within this germplasm and supports its suitability for association mapping. High broad-sense heritability ($H^2 > 90\%$) for most traits shows that genetic factors predominantly drive phenotypic variation under the evaluated conditions, consistent with high heritability reported for FA composition in other oilseed crops including soybean and oil palm^{28,29}. The moderate heritability observed for ω -3 ($H^2 = 52\%$) suggests environmental sensitivity of PUFA desaturation processes, consistent with temperature-dependent regulation of FA desaturation reported in other plant species³⁰. From a food quality perspective, the dominance of PUFAs, with ω -6 as the most abundant individual FA (44.7%) followed by ω -3 (21.7%) characterizes RSO as an essential FA-rich vegetable oil with nutritional relevance comparable to established edible oils. This FA profile—providing both ω -6 and ω -3 essential FAs in a single non-competitive oil source—aligns with previous biochemical characterizations of RSO and highlights the nutritional potential that genetic improvement could further optimize^{3,31}. Compared with commonly edible vegetable oils, including sunflower oil (C18:2, approximately 65%), soybean oil (ω -6, approximately 53%; ω -3 approximately 8%), and palm oil (C16:0, approximately 44%)—RSO provides

a notably higher ω -3 content (21.7%) and more balanced ω -6/ ω -3 ratio^{3,32–34}, suggesting nutritional advantages that genetic improvement could further optimize. The negative correlations between SOC and several individual FA components may reflect competitive allocation of carbon between triacylglycerol accumulation and FA compositional regulation during seed development, a pattern documented in other oilseed species where oil content and FA composition share biosynthetic substrates³⁵. Strong positive co-associations among SFAs and among MUFA components suggest shared metabolic regulation within biosynthetic sub-pathways, while negative associations between PUFAs and SFA/MUFA components are consistent with competitive FA desaturation processes at the *FAD2/FAD3* enzymatic steps²². The inverse association between SOC and seed size traits aligns with carbon allocation trade-offs reported in other oilseed species and suggests that breeding for simultaneously improved oil content and seed size may require allele pyramiding strategies rather than directional selection³⁶.

The identification of four genetically distinct ancestral subpopulations, provides an important characterization of the genetic diversity landscape in rubber tree accession's panel. Concordance among three independent analytical frameworks strengthens confidence in the population structure inference, consistent with methodological approaches increasingly recommended in population genomics studies³⁷. The admixture observed between adjacent clusters reflects the documented breeding history of rubber tree germplasm—shaped by controlled crosses, vegetative propagation, and germplasm exchange across major producing regions^{15,38}. The low variance explained by the first three PCs is characteristic of highly diverse outcrossing perennial crops including cacao³⁹, grapevine⁴⁰, and Eucalyptus⁴¹, where genomic variation is broadly distributed across axes due to extensive heterozygosity and historical recombination. Rapid LD decay ($LD_{50}=50$ kb) reflects the high genetic diversity of the mixed cultivated and wild germplasm evaluated and contrasts with slower decay reported for cultivated-only rubber tree panels ($R^2=0.49$)⁴², where reduced diversity extends haplotype blocks and limits fine-mapping resolution. The rapid LD decay in this study provides an advantage over self-pollinating crops such as soybean and rice where domestication bottlenecks extend LD blocks, enabling localization of association signals to narrower genomic intervals and facilitating candidate gene prioritization⁴³. Conventional single-locus GWAS models can be limited in detecting small-effect loci controlling complex quantitative traits¹⁹; therefore, the three complementary models employed here—utilize distinct statistical approaches for handling polygenic backgrounds²¹, collectively improved detection power and robustness of identified associations.

The substantial variation in MTAs across SOC and FA traits reflects heterogeneous genetic architecture among components of seed lipid metabolism—minor FA classes (C14:0, C20:0, C20:1) showing more dispersed polygenic control than major regulatory traits like total oil content, consistent with patterns in other oilseed crops where minor FAs tend to show more complex genetic regulation^{20,43}. The relatively few MTAs for C18:2 (6 MTAs) and SOC (16 MTAs) may indicate fewer large-effect loci controlling these traits—which are arguably the most nutritionally important components of RSO—suggesting that

targeted manipulation of these traits may be achievable through a limited number of key genomic loci. The convergence of association signals at SNP (17_28516849) on Chr17 for three FA traits differing in chain length and saturation—C14:0, C17:0, and C20:1—raises the possibility of a regulatory locus with coordinated effects on multiple branches of FA metabolism. From a biosynthetic perspective, these FAs are connected through shared elongation and modification pathways, and a TF or pathway regulator at this locus could plausibly influence multiple enzymatic steps simultaneously—analogueous to the coordinated effects of master lipid regulators such as WRI1 on multiple lipogenic genes^{43,44}. A comparable genomic region on Chr17 in soybean has been associated with multiple SFA and MUFA traits and attributed to a FA elongase gene cluster⁴⁵, and whether the rubber tree Chr17 locus represents an analogous regulatory hub warrants investigation through conditional association testing, colocalization analysis, and transcriptome profiling of accessions carrying contrasting Chr17 haplotypes. The detection of multiple association signals for both SOC and arachidic acid on Chr10, mapping to genomic intervals separated by distances substantially exceeding the estimated LD₅₀ of 50 kb, is consistent with the presence of multiple independently acting loci on this chromosome rather than a single broad association driven by extended LD. While GWAS alone cannot establish whether this pattern reflects physical clustering of functionally related lipid metabolism genes or independent regulatory variation at separate loci; thus, the Chr10 region represents a priority target for future fine-mapping and candidate gene characterization. Similar concentrations of seed lipid-associated loci on individual chromosomes have been reported in soybean and *Brassica napus*, where they have been attributed to clusters of biosynthetically related genes^{18,43}, though whether an analogous mechanism underlies the Chr10 pattern in rubber tree remains to be determined through functional investigation. The distribution of seed size trait associations across multiple chromosomes is consistent with the polygenic inheritance commonly observed for quantitative morphological traits in perennial crops^{46,47}. The absence of shared GWAS loci between SOC and seed size traits, despite their negative phenotypic correlation, suggests that the observed trade-off may arise from developmental resource competition or independent genetic pathways rather than from detectable pleiotropic loci. If confirmed in larger independent populations, the apparent genetic separability of these trait groups would suggest that pyramiding favorable SOC and seed size alleles within a single genetic background may be achievable⁴⁸.

The five candidate genes identified through integrated fine-mapping and haplotype analyses span FA biosynthesis, metabolic signaling, and developmental regulation, collectively providing a biologically coherent framework for RSO quality improvement. MOD1 encoding an enoyl-ACP reductase catalyzing the final reductive step in each plastidial FA elongation cycle, represents the direct link between identified genetic variation and stearic acid accumulation relevant to RSO oxidative stability as an edible oil. Loss of function studies in *Arabidopsis* confirm MOD1's essential role in maintaining saturated to unsaturated FA balance and normal plastid lipid metabolism⁴⁹, and the significant H002 haplotype association ($P = 0.043$) alongside the strong SNP-level effect ($P = 0.0003$) may provide complementary genetic evidence, consistent with recommended approaches for prioritizing candidates⁵⁰. Notably, stearic acid is nutritionally distinct from other SFAs in that it does not raise low-density lipoprotein cholesterol levels in human and is considered cholesterol-neutral by the WHO⁵¹. Furthermore, oils

enriched in stearic acid exhibit physical properties resembling cocoa butter, introducing potential applications in specialty food fats and confectionery lipid systems where structured fat functionality is required⁵². Genetic improvement of stearic acid content through MOD1 targeted breeding could therefore expand RSO utility in premium food ingredient applications beyond conventional vegetable oil uses. SRK2I encoding an SnRK2-family serine/threonine kinase⁵³, represents a plausible regulator of seed oil through ABA signaling-mediated effects on seed filling—ABA-SnRK2 pathways have documented downstream effects on lipogenic gene expression including WRI1 targets⁴⁴—with the H004 haplotype showing significant SOC elevation ($P = 0.019$). As SOC directly determines the volume of edible oil extractable per seed, genetic improvement through SRK2I-targeted marker-assisted selection would have direct economic relevance for RSO production as a food-grade vegetable oil in smallholder rubber plantation systems.

KASII directly catalyzing palmitic acid to stearic acid elongation that determines SAD substrate availability for oleic acid production, represents biologically coherent FA biosynthesis candidates—directly relevant to oleic acid content, a key determinant of RSO nutritional quality and oxidative stability for edible use⁵⁴. KASII gene has been silenced to increase palmitic acid content in cotton seed oil using RNAi and validate its FA biosynthesis involvement⁵⁵. Increasing oleic acid content in edible oils is a key breeding objective across oilseed crops, as oleic acid imparts superior oxidative stability, longer shelf life, boost memory, hinder the diseases related to brain and adrenal glands, and reduced requirements for partial hydrogenation—thereby avoiding trans-fatty acid formation—relative to PUFA-rich oil as is RSO^{3,56,57}. High oleic variants of soybean⁵⁸, olive⁵⁷, *Brassica* oil⁵⁹, and sunflower⁶⁰ securing premium food market prices demonstrate the commercial food value of oleic acid enrichment, and genetic improvement of RSO toward higher oleic acid content through KASII-targeted breeding could similarly enhance its competitiveness as a food-grade edible oil⁶¹. BPM3 proteins interact with WRI, functioning as a CUL3-E3 ubiquitin ligase substrate adaptor potentially influencing lipid-regulatory protein stability, with FA level significantly increased in BPM overexpressed lines similar to overexpression of WRI1^{62,63}. These suggests BPM proteins in seeds functions is strongly coordinated with WRI1 being the primary regulator of fatty acid biosynthesis in seeds³.

WUSCHEL and YABBY2 encoding TFs with established roles in meristematic cell proliferation and lateral organ elongation respectively^{26,64}, considering potential candidate genes with seed developmental regulators directly relevant to oil extraction efficiency through seed size improvement. From food production perspective, greater seed thickness is associated with higher seed mass and improved mechanical oil extraction efficiency, linking WOX1-targeted genetic improvement directly to the processing and economic viability of RSO as an edible oil resource. For KASII, BPM3, and YABBY2, the significant SNP-level associations suggest that causal variants at these loci may reside in regulatory rather than coding regions—a pattern increasingly recognized for metabolic and developmental genes where promoter-level variation drives natural phenotypic diversity, as demonstrated for YABBY family genes in tomato²⁷, pepper⁶⁵, and cucumber⁶⁶. Future integration of promoter resequencing and eQTL analysis at these loci would therefore complement the coding-variant evidence reported here and

provide a more complete characterization of functional variation underlying FA composition and seed size traits in rubber tree⁶⁷.

The genomic resources generated in this study have promising implications for improving RSO as an edible vegetable oil through marker-assisted breeding. Favorable haplotypes for MOD1 (H002), SRK2I (H004), and WOX1 (H003) are present at considerable frequencies within the evaluated panel of rubber tree accessions, providing identifiable donor materials for allele introgression into elite rubber tree genetic backgrounds. The genetic independence of SOC and seed size loci suggests that pyramiding favorable SOC alleles with favorable seed size alleles—which would improve both oil content and extraction efficiency—should be achievable, though validation in independent populations⁴⁸. However, several limitations of the current study should be considered including single environment evaluation limits characterization of genotype by environment interactions particularly for ω -3 ($H^2=52\%$), and the accession panel (186) restricts power for small effect loci, which may partly explain the low MTAs for nutritionally important ω -6. Expanding the panel of accessions with greater geographic diversity and conducting multi-environment trials across rubber-growing regions would improve detection power, allele representation, and breeding deployment confidence³⁸. Beyond these genetic and phenotyping limitations, two compositional features have historically been flagged as practical constraints on unrestricted edible use of RSO: its semi-drying character, arising from relatively high ω -6 content, and the presence of trace polyisoprene (about 1%) co-extracted within the unsaponifiable fraction^{68,69}. The semi-drying property could potentially be mitigated through blending with other vegetable oils or partial hydrogenation, as practiced for other high- ω -6 oils such as linseed oil, which is commercially marketed for edible use despite an even higher ω -6 content⁷⁰. The long-term physiological consequences, if any, of trace polyisoprene ingestion do not appear to have been systematically investigated and warrant dedicated safety evaluation as part of future RSO food-grade development efforts.

Future research priorities include functional validation of these potential candidate genes through expression profiling across seed developmental stages, genome editing, and biochemical characterization. In addition, investigation of the pleiotropic locus (Chr17) through transcriptomic and metabolomic profiling of contrasting haplotype accessions. Integration of the association mapping results with emerging *H. brasiliensis* pangenomic resources including structural variant catalogues^{16,17}. And evaluation of rubber seed oil FA composition improvement in the context of established food safety evidence, building on clinical data demonstrating the ability of RSO to reduce blood lipid levels and support cardiovascular health when consumed as a dietary oil^{3,6,7}. RSO intended for food applications should undergoes conventional extraction and refining, bleaching, and deodorization processing to achieve food-grade quality⁷¹, consistent with standard edible oil production practices applied globally and already employed in regions where RSO has an established history of edible use^{3,7}. Collectively, these findings provides genomic foundation for targeted improvement of RSO as a sustainable, non-competitive vegetable oil resource, positioning rubber tree seed quality improvement as a tractable molecular breeding objective with direct relevance to food security, agricultural diversification, and the valorization of underutilized plantation by products. Beyond nutritional attributes, the genetic

improvement of RSO aligns with broader food system sustainability objectives such as providing a nutritious edible oil from plantation waste, thereby addressing both food security and environmental sustainability concerns associated with expanding conventional oilseed production^{72,73}. Earlier compositional studies of rubber seed kernel and oil similarly reported that potentially toxic kernel constituents—including cyanogenic compounds, aflatoxin, gossypol, and saponins—do not carry over into the extracted oil itself, providing additional historical support for RSO's suitability as an edible-oil base^{68,69}. As an underutilized oil source with established edible safety^{6,7} and a favorable essential FA profile, genetically improved RSO has the potential to contribute to vegetable oil supply diversification in tropical rubber-producing regions where both food security and agricultural income are priority concerns^{74,75}. This need is underscored by the diverse vegetable oil consumption patterns observed in tropical oil-producing regions, including natural rubber-producing countries where edible oil supplies often depend on a combination of locally produced and imported oils. In Sri Lanka, coconut oil has historically been the dominant edible vegetable oil used by households, while other nutritionally important oils have had a smaller role in the domestic market^{76,77}. This highlights the potential value of locally available, underutilized oil resources including RSO, which contains substantial amounts of PUFAs, including ω -6 and ω -3, and may contribute to improving dietary fatty acid diversity^{3,31}.

Conclusion

This study establishes the first genomic framework for the genetic architecture of rubber seed oil content, fatty acid composition, and seed size traits in *Hevea brasiliensis*, providing a foundation for molecular breeding of this underutilized non-competitive vegetable oil resource. Multi-locus GWAS integrating three models with LD-based fine-mapping identified promising candidate genes—MOD1, SRK2I, KASII, WOX1, and YABBY2—spanning fatty acid biosynthesis, metabolic signaling, and seed developmental regulation, with favorable haplotypes for MOD1, SRK2I, and WOX1 representing actionable genomic resources for marker-assisted improvement of RSO nutritional quality and seed yield. The pleiotropic Chr17 locus simultaneously influencing multiple fatty acid traits suggests coordinated genetic regulation warranting priority functional investigation. Given RSO's established edible safety, favorable essential PUFA-rich profile comprising ω -6 and ω -3, and its unique advantage as a plantation by product requiring no additional land cultivation, the genomic resources generated here contribute directly to developing rubber tree seed quality improvement as a tractable breeding objective—advancing RSO as a sustainable, nutritious edible oil for food, nutraceutical, and food security applications in tropical rubber producing regions.

Material and methods

Plant materials and experimental design

A diverse panel consisting of 186 *H. brasiliensis* accessions maintained at the Jinghong Rubber Tree Germplasm Resource Nursery of Yunnan Institute of Tropical Crops was evaluated in this study. Mature

seeds were collected from the nursery at the Xishuangbanna region of Yunnan Province, China (21°93'N, 101°27'E), which represents one of the major rubber-growing areas in China. Fully mature rubber seeds were collected during the fruit maturation period (Aug-Sep 2024) from at least five independent biological replicates per accession to minimize environmental effects. After harvesting, seeds were immediately cleaned, kernels were manually separated from seed shells, flash-frozen in liquid nitrogen, and stored at -80°C until further analysis. seed size traits, including seed length (SL), seed width (SW), seed thickness (ST), fruit weight (FW), and caryopsis weight (CW) were recorded for each accession at both environments using digital calipers and an analytical balance, with a minimum of 20 seeds measured per accession per environment.

Seed oil extraction and quantification

Rubber seed kernels were separated manually from the seed coat and homogenized to a fine powder using a high-speed grinder. Approximately 200 g of kernel powder per accession was freeze-dried using a vacuum freeze dryer (Labconco FreeZone, USA) to remove residual moisture prior to oil extraction. Seed oil extraction was performed using n-hexane solvent extraction^{3,78}. Ground samples were extracted in n-hexane under constant agitation at room temperature for 24 h. The solvent extract was subsequently filtered through Whatman No. 1 filter paper and concentrated using a rotary evaporator under reduced pressure at 40°C . The recovered oil was weighed using an analytical balance to calculate SOC as a percentage of dry kernel weight, and stored at 4°C in sealed amber vials under nitrogen gas prior to FA profiling.

Fatty acid profiling using GC-MS

The FA composition was determined by converting extracted lipids to fatty acid methyl esters (FAMES) prior to gas chromatography–mass spectrometry (GC-MS) analysis. Approximately 1 g of freeze-dried kernel powder was subjected to ultrasonic extraction using a methanol: chloroform mixture (2:1, v/v) containing 0.9% NaCl solution. Following centrifugation at $12000 \times g$ for 5 min, the lower organic phase was collected and evaporated to dryness under a stream of nitrogen gas. The extracted lipids were saponified using 0.5 M potassium hydroxide in methanol at 60°C for 30 min, followed by methylation using 14% boron trifluoride–methanol reagent at 60°C for 30 min. The resulting FAMES were extracted with n-hexane, filtered through a $0.22 \mu\text{m}$ polytetrafluoroethylene (PTFE) membrane filter, and analyzed within 24 h of preparation. GC-MS analysis was performed using a TRACE GC Ultra gas chromatograph coupled with an ISQ mass spectrometer (Thermo Fisher Scientific, USA). Chromatographic separation was achieved on a DB-5MS capillary column ($30 \text{ m} \times 0.25 \text{ mm} \times 0.25 \mu\text{m}$) with helium as the carrier gas at a flow rate of 1 mL/min. The oven temperature program was as follows: initial temperature of 80°C for 2.5 min, ramp rate of $15^{\circ}\text{C}/\text{min}$ – 210°C to final temperature of 230°C . Mass spectra were acquired under electron ionization (EI) mode at 70 eV with a scan range of 40–460 m/z. Individual FAs were

identified by comparison with the NIST mass spectral library and quantified by peak area normalization relative to an internal standard⁷⁹.

Phenotypic data analysis

Analysis of variance (ANOVA) was performed to assess the significance of genotypic differences among accessions for all evaluated traits. Broad-sense heritability (H^2) was estimated using the formula $H^2 = VG / (VG + VGE/e + VE/re)$, where VG is genetic variance, VGE is genotype by environment interaction variance, VE is residual variance, e is the number of environments, and r is the number of replicates, implemented in R software. Descriptive statistics were calculated for all traits across. Normality of trait distributions was assessed using the Kolmogorov–Smirnov (K–S) test. Pearson correlation coefficients among all trait were calculated in R v4.5.1, and visualized as a correlation heatmap using the ggcorrplot package. Principal component analysis (PCA) of phenotypic trait data was performed using the prcomp function in R v4.5.1, and visualized using ggbiplot, with eigenvalue decomposition of the correlation matrix used to identify PCs explaining the major axes of phenotypic variation.

Whole-genome resequencing and SNP genotyping

Genomic DNA was extracted from young leaves of each accession using the CTAB method with slight modifications, and DNA quality was assessed by NanoDrop spectrophotometry and 1% agarose gel electrophoresis. High-quality DNA samples were used for library construction. Paired-end sequencing libraries were prepared using the NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA, USA) library preparation kit and sequenced on the Illumina NovaSeq 6000 platform (Illumina, USA) to generate 150 bp paired-end reads with a target average genome coverage of $\geq 10\times$ per accession. Raw reads were filtered to remove adaptor sequences and low-quality reads using Trimmomatic v0.39⁸⁰, and clean reads were aligned to the *H. brasiliensis* reference T2T genome (haplotype A)¹⁷ using BWA-MEM v0.7.17⁸¹. Duplicate reads were marked and removed using Picard (MarkDuplicates). Variant calling was performed using the GATK HaplotypeCaller pipeline v4.2 in GVCF mode, followed by joint genotyping and variant quality score recalibration⁸². Multi-allelic SNPs (> 2 alleles) and SNPs positioned within or 5 bp flanking any InDels were discarded. Genotypes with read depth (DP) < 5 and genotype quality (GQ) < 10 were defined as missing genotypes. Subsequently, SNPs were retained after quality filtering using the following criteria: minor allele frequency (MAF) > 0.05 , and missing rate $> 20\%$. The remaining high-quality variants were further phased and imputed using beagle 5.5 (v27Feb25.75f). Functional annotation of retained SNPs was performed using SnpEff v5.0 with the *H. brasiliensis* genome annotation as reference⁸³.

Population structure and genetic diversity analysis

Population genetic structure was assessed using model-based clustering implemented in ADMIXTURE v1.3.0⁸⁴. Prior to analysis, the quality-filtered SNP dataset was subjected to linkage disequilibrium (LD) pruning using PLINK v1.90⁸⁵, with the following parameters: `--indep-pairwise 50 5 0.1`. After LD pruning, 1,210,48 SNPs were retained for population structure inference. ADMIXTURE was run for $K = 1$ to 10 with

five-fold cross-validation (–cv). The optimal number of ancestral populations (K) was determined by selecting the K value with the lowest cross-validation error. Ancestry proportions (Q-matrix) for each individual were extracted at the optimal K. PCA was performed based on 16,629,407 high-quality SNPs using PLINK v1.90⁸⁵. An unrooted neighbor-joining phylogenetic tree was constructed using high-quality SNP dataset in MEGA v11⁸⁶ with 1000 bootstrap replicates to assess phylogenetic relationships among accessions. Linkage disequilibrium (LD) decay was evaluated from the quality-filtered SNP dataset using PopLDdecay v3.43⁸⁷. Pairwise squared correlation coefficients (r^2) between SNPs were calculated for each chromosome independently, and the mean r^2 was computed for physical distances up to 500 kb. The half-decay distance (LD_{50}), defined as the distance at which r^2 declines to half of its maximum value, was determined from the binned decay curve to characterize the extent of LD in the population.

Genome-wide association analysis

Multi-locus GWAS was conducted for SOC, FA composition, and seed size traits using three GWAS models implemented in the GAPIT3 R package⁸⁸: Bayesian information and linkage disequilibrium iteratively nested keyway (BLINK)⁸⁹, fixed and random model circulating probability unification (FarmCPU)²¹, and multiple locus mixed model (MLMM)⁹⁰. All three models incorporated the Q matrix (K = 4) to control for population structure. The significance threshold for GWAS was determined using a Bonferroni correction based on the number of independent markers tested. Specifically, the significance threshold was calculated as $P = 0.05/N$, where N represents the number of LD-pruned SNPs retained after filtering (1,210,48), yielding a conservative genome-wide significance threshold of $P < 4.13e-07$ (equivalent to $-\log_{10}(P) > 6.38$). SNPs were considered high-confidence associations if detected by all three models, moderately supported if detected by two models, and a single model. Significant SNPs identified across models were deduplicated and clustered into independent quantitative trait loci (QTLs) based on a physical distance window of 50 kb, consistent with the estimated LD decay distance in the rubber accession panel. QTLs were named following the convention q (trait, chromosome, locus number) (for example, *qSOC-10.1*). Manhattan plots and quantile–quantile (QQ) plots were generated for all traits using the qqman R package to visualize association signals.

Linkage disequilibrium and candidate gene identification

For each significant QTL detected by multiple GWAS models, an LD block was defined around the lead SNP using the confidence interval method implemented in Haploview v4.2⁹¹. Gene annotations within each LD block were retrieved from the *H. brasiliensis* reference genome annotation¹⁷, and candidate genes were prioritized based on: (i) the presence of non-synonymous (missense) coding variants in strong LD with the lead SNP; (ii) functional annotation consistent with known roles in FA biosynthesis, seed oil accumulation, or seed development; and (iii) literature evidence from homologous genes in model species or related crops. Non-synonymous variants within prioritized candidate genes were annotated using SnpEff³³ with amino acid changes reported in standard HGVS notation.

Haplotype analysis and phenotypic association testing

Haplotype analysis was performed for the most biologically relevant candidate gene at each prioritized QTL. Haplotypes were defined based on the combination of non-synonymous SNP alleles identified within each candidate gene CDS. Haplotypes with a sample size below ten accessions were excluded from statistical testing to ensure adequate power. Pairwise haplotype–phenotype associations were evaluated using Student’s t-test or one-way ANOVA with Tukey’s HSD post-hoc test, with significance determined at $P < 0.05$. Genotype–phenotype associations for the lead SNP at each QTL were assessed using the same statistical framework. Association matrices and violin plots with embedded boxplots were generated in R using the ggplot2 package to visualize haplotype–phenotype relationships. All statistical analyses were performed in R v4.5.1.

Data availability

The authors declare that all pertinent data that support this study have been included within the paper. Raw data will be made available by corresponding authors upon reasonable request.

Declarations

Acknowledgement

This study was funded by the Central Public-interest Scientific Institution Basal Research Fund (Grant No. 1630022026006), Innovational Fund for Scientific and Technological Personnel of Hainan Province (Grant No. KJRC2023A04 and Grant No. KJRC2025A04), and the Project of State Key Laboratory for Tropical Crop Breeding (Grant No. SKLTCBBSH202503).

Author contributions

Conceptualization F.R., X.W., and H.C.; investigation T.W., and D.X.; methodology X.W., and F.R.; formal analysis and data curation X.W., and F.R.; writing original draft F.R.; writing-review and editing F.R., W.W., Y.Z., D.X., and Y.W.; sample collection, and preparation F.R., T.Y., Y.Z., Y.W., and W.W.; metabolic profiling F.R., and T.W.; phenotypic data collection F.R.; funding acquisition F.R., X.W., and H.C; project administration H.C. All authors read and approved the submission of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information

Correspondence and requests for materials should be addressed to Han Cheng and Xiaobo Wang.

References

1. Salimon, J., Abdullah, B.M. & Salih, N. Rubber (*Hevea brasiliensis*) seed oil toxicity effect and Linamarin compound analysis. *Lip. Health Dis.* 11, (2012).
2. Ikhuoria, E. U., Okieimen, F. E., Obazee, E. O. & Erhabor, T. Synthesis and characterization of styrenated rubber seed oil alkyd. *African J. Biotech.* 10, 1913–1918 (2011).
3. Rehman, F. et al. Unlocking rubber seed oil potential: Integrative omics to panomics perspectives and future breeding strategies. *Indus. Crops Prod.* 238, 122382 (2025).
4. Pinzi, S. et al. The ideal vegetable oil-based biodiesel composition: a review of social, economical and technical implications. *Ener. Fuels* 23, 2325–2341 (2009).
5. Demirbas, A. Political, economic and environmental impacts of biofuels: a review. *Appl. Energ.* 86, 108–117 (2009).
6. Chen, J., He, M. & Bai, X. Research and development progress of rubber seed oil in Yunnan. *Trop. Agric. Sci. Tech.* 27, 4 (2004).
7. He, M., Li, G. & Li, H. Current situation and thinking of edible-oriented development of rubber seed oil in Yunnan. *Trop. Agric. Sci. Tech.* 04, (2010).
8. Lu, Q., Chen, P., Chai, Y., Li, Q. & Mao, H. Effects of dietary rubber seed oil on production performance, egg quality and yolk fatty acid composition of hy-line brown layers. *Animal Biosci.* 34, 119–126 (2020).
9. Chaikul, P., Lourith, N. & Kanlayavattanakul, M. Decelerated skin aging effect of rubber (*Hevea brasiliensis*) seed oil in cell culture assays. *Sci. Rep.* 14, (2024).
10. Atabani, A.E. et al. A comprehensive review on biodiesel as an alternative energy resource and its characteristics. *Renew. Sus. Ener. Rev.* 16, 2070–2093 (2013).
11. Mantello, C.C. et al. Microsatellite marker development for the rubber tree (*Hevea brasiliensis*): characterization and cross-amplification in wild *Hevea* species. *BMC Res. Not.* 5, (2012).
12. Pootakham, W. et al. Construction of a high-density integrated genetic linkage map of rubber tree (*Hevea brasiliensis*) using genotyping-by-sequencing (GBS). *Front. Plant Sci.* 6, (2015).
13. Wu, W. et al. Ultra high-density genetic map construction and identification of quantitative trait loci for growth in rubber tree (*Hevea brasiliensis*). *Ind. Crops Prod.* 178, 114560 (2022).
14. Francisco, F.R. et al. Unravelling rubber tree growth by integrating GWAS and biological Network-Based approaches. *Front. Plant Sci.* 12, (2021).
15. Cheng, H. et al. Chromosome-level wild *Hevea brasiliensis* genome provides new tools for genomic-assisted breeding and valuable loci to elevate rubber yield. *Plant Biotech. J.* 21, 1058–1072 (2023).
16. Fang, Y. et al. Pan-genome and phylogenomic analyses highlight *Hevea* species delineation and rubber trait evolution. *Nat. Commun.* 15, (2024).
17. Li, C. et al. The haplotype-resolved telomere-to-telomere genome and OMICS analyses reveal genetic responses to tapping in rubber tree. *Nat. Commun.* 16, 6255 (2025).
18. Yuan, X. et al. Integrative omics analysis elucidates the genetic basis underlying seed weight and oil content in soybean. *Plant Cell* 36, 2160–2175 (2024).

19. Elsafy, M. et al. Genome-wide association study of oil content and fatty acid composition in sesame (*Sesamum indicum* L.) under diverse environmental conditions. *Crop Sci.* 65, (2025).
20. Zhang, Y. et al. Integrative omics analysis reveals the genetic basis of fatty acid composition in *Brassica napus* seeds. *Genom Biol.* 26, 83 (2025).
21. Liu, X., Huang, M., Fan, B., Buckler, E. S. & Zhang, Z. Iterative usage of fixed and random effect models for powerful and efficient Genome-Wide association studies. *PLoS Genet.* 12, e1005767 (2016).
22. Okuzaki, A. et al. CRISPR/Cas9-mediated genome editing of the fatty acid desaturase 2 gene in *Brassica napus*. *Plant Physiol. Biochem.* 131, 63–69 (2018).
23. Neelakandan, A. K. et al. Application of CRISPR/CAS9 system for efficient gene editing in peanut. *Plants* 11, 1361 (2022).
24. Rehman, F. et al. CRISPR gene editing of major domestication traits accelerating breeding for Solanaceae crops improvement. *Plant Mol. Biol.* 108, 157–173 (2022).
25. Vandenbussche, M. et al. Differential recruitment of WOX transcription factors for lateral development and organ fusion in *Petunia* and *Arabidopsis*. *Plant Cell* 21, 2269–2283 (2009).
26. Sarojam, R. et al. Differentiating *Arabidopsis* Shoots from Leaves by Combined YABBY Activities. *Plant Cell* 22, 2113–2130 (2010).
27. Han, H., Liu, Y., Jiang, M., Ge, H. & Chen, H. Identification and expression analysis of YABBY family genes associated with fruit shape in tomato (*Solanum lycopersicum* L.). *Genet. Mol. Res.* 14, 7079–7091 (2015).
28. Pham, A., Lee, J., Shannon, J. G. & Bilyeu, K. D. Mutant alleles of FAD2-1A and FAD2-1B combine to produce soybeans with the high oleic acid seed oil trait. *BMC Plant Biol.* 10, 195 (2010).
29. Osorio-Guarín, J. A. et al. Genome-wide association study (GWAS) for morphological and yield-related traits in an oil palm hybrid (*Elaeis oleifera* x *Elaeis guineensis*) population. *BMC Plant Biol.* 19, 533 (2019).
30. Menard, G. N. et al. Genome Wide Analysis of Fatty Acid Desaturation and Its Response to Temperature. *Plant Physiol.* 173, 1594–1605 (2017).
31. Oyekunle, D. T. et al. Review on utilization of rubber seed oil for biodiesel production: Oil extraction, biodiesel conversion, merits, and challenges. *Clean. Eng. Tech.* 21, 100773 (2024).
32. Calder, P. C. Functional Roles of Fatty Acids and Their Effects on Human Health. *J. Parent. Ent. Nutr.* 39, 18S-32S (2015).
33. Simopoulos, A. An increase in the Omega-6/Omega-3 fatty acid ratio increases the risk for obesity. *Nutr.* 8, 128 (2016).
34. Innes, J. K. & Calder, P. C. Omega-6 fatty acids and inflammation. *Prostaglandins Leukotrienes Essen. Fatty Acids* 132, 41–48 (2018).
35. Chapman, K. D. & Ohlrogge, J. B. Compartmentation of triacylglycerol accumulation in plants. *J. Biolog. Chem.* 287, 2288–2294 (2011).

36. Weselake, R. J. et al. Increasing oil content in Brassica oilseed species. *Prog. Lipid Res.* 96, 101306 (2024).
37. Lawson, D. J., Van Dorp, L. & Falush, D. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat. Commun.* 9, 3258 (2018).
38. Priyadarshan, P. & Clément-Demange, A. Breeding Hevea rubber: Formal and Molecular Genetics. *In Adv. Genet.* 52, pp. 51–115 (2004).
39. Motamayor, J. C. et al. Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree (*Theobroma cacao* L). *PLoS One* 3, e3311 (2008).
40. Myles, S. et al. Genetic structure and domestication history of the grape. *Proc. Nat. Acad. Sci.* 108, 3530–3535 (2011).
41. Silva-Junior, O. B., Faria, D. A. & Grattapaglia, D. A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 Eucalyptus tree genomes across 12 species. *New Phyt.* 206, 1527–1540 (2015).
42. de Souza, L. M. et al. Linkage Disequilibrium and Population Structure in Wild and Cultivated Populations of Rubber Tree (*Hevea brasiliensis*). *Front. Plant Sci.* 9, 815 (2018).
43. Fang, C. et al. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biol.* 18, 161 (2017).
44. Wei, W. et al. The comprehensive regulatory network in seed oil biosynthesis. *J. Integ. Plant Biol.* 67, 649–668 (2025).
45. Li, X. et al. Genetic loci and causal genes for seed fatty acids accumulation across multiple environments and genetic backgrounds in soybean. *Mol. Breed.* 41, 31 (2021).
46. Ma, X., Su, Z. & Ma, H. Molecular genetic analyses of abiotic stress responses during plant reproductive development. *J. Experim. Bot.* 71, 2870–2885 (2020).
47. Zhang, P., Yang, Z., Jia, S., Chen, G., Li, N., Karikari, B. & Cao, Y. Genome-Wide Association study and candidate gene mining of seed size traits in soybean. *Agron.* 14, 1183 (2024).
48. Soto-Cerda, B. J. et al. Genomic regions underlying agronomic traits in linseed (*Linum usitatissimum* L.) as revealed by association mapping. *J. Integ. Plant Biol.* 56, 75–87 (2013).
49. Mou, Z., He, Y., Dai, Y., Liu, X. & Li, J. Deficiency in fatty acid synthase leads to premature cell death and dramatic alterations in plant morphology. *Plant Cell* 12, 405–417 (2000).
50. Gao, J., Espin-Garcia, O., Paterson, A. D. & Sun, L. Integrating variant functional annotation scores have varied abilities to improve power of genome-wide association studies. *Sci. Rep.* 12, 10720 (2022).
51. Mensink, R. P. Effects of saturated fatty acids on serum lipids and lipoproteins: A systematic review and regression analysis. World Health Organization, Geneva (2016).
52. Salas, J. J. et al. Production of stearate-rich butters by solvent fractionation of high stearic–high oleic sunflower oil. *Food Chem.* 124, 450–458 (2011).

53. Kulik, A., Wawer, I., Krzywińska, E., Bucholc, M. & Dobrowolska, G. SNRK2 Protein Kinases—Key regulators of plant response to abiotic stresses. *OMICS a Journal of Integrative Biology*, 15, 859–872 (2011).
54. Kim, K., Park, J., Norah, N., Kim, S., Jo, H., Song, J. T. & Lee, J. Structural and functional characterization of a novel GmKASII-A allele associated with saturated fatty acid composition in EMS-induced mutant PE1544. *BMC Plant Biol.* 25, 818 (2025).
55. Liu, Q. et al. Genetic enhancement of palmitic acid accumulation in cotton seed oil through RNAi down-regulation of ghKAS2 encoding β -ketoacyl-ACP synthase II (KASII). *Plant Biotech. J.* 15, 132–143 (2017).
56. Choe, E. & Min, D. B. Mechanisms and factors for edible oil oxidation. *Compreh. Rev. Food Sci. Food Safety* 5, 169–186 (2006).
57. Lu, Y. et al. Protective effects of oleic acid and polyphenols in extra virgin olive oil on cardiovascular diseases. *Food Sci. Human Well.* 13, 529–540 (2024).
58. Pham, A., Shannon, J. G. & Bilyeu, K. D. Combinations of mutant FAD2 and FAD3 genes to produce high oleic acid and low linolenic acid soybean oil. *Theor. Appl. Genet.* 125, 503–515 (2012).
59. Scarth, R. & Tang, J. Modification of brassica oil using conventional and transgenic approaches. *Crop Sci.* 46, 1225–1236 (2006).
60. Fernández-Martínez, J. M., Pérez-Vich, B. & Velasco, L. Sunflower. In: Vollmann, J., Rajcan, I. (eds) *Oil Crops. Handbook of Plant Breeding*, vol 4. Springer, New York, NY (2009).
61. Zhong, Y. et al. Mapping and identification of QTLs for seed fatty acids in soybean (*Glycine max* L.). *J. Integ. Agri.* 23, 3966–3982 (2024).
62. Cernac, A. & Benning, C. WRINKLED1 encodes an AP2/EREB domain protein involved in the control of storage compound biosynthesis in *Arabidopsis*. *Plant J.* 40, 575–585 (2004).
63. Chen, L. et al. *Arabidopsis* BPM Proteins Function as Substrate Adaptors to a CULLIN3-Based E3 Ligase to Affect Fatty Acid Metabolism in Plants. *Plant Cell* 25, 2253–2264 (2013).
64. Mayer, K. F. et al. Role of WUSCHEL in regulating stem cell fate in the *Arabidopsis* Shoot meristem. *Cell* 95, 805–815 (1998).
65. Fang, K. et al. Genome-wide analysis of the CaYABBY family in pepper and functional identification of CaYABBY5 in the regulation of floral determinacy and fruit morphogenesis. *J. Integ. Agric.* 24, 3024–3039 (2025).
66. Zhao, J., Song, W. & Zhang, X. Genetic and molecular regulation of fruit development in cucumber. *New Phyt.* 244, 1742–1749 (2024).
67. Wang, B. et al. Genome-wide selection and genetic improvement during modern maize breeding. *Nat. Genet.* 52, 565–571 (2020).
68. Achinewhu, S. C. Unconventional sources of food: Chemical composition of rubber seed (*Hevea brasiliensis*). *Food Chem.* 21, 17–25 (1986).

69. Nwokolo, E., Kitts, D. D. & Kanhai, J. Serum and liver lipids of rats fed rubber seed oil. *Plant Food Hum. Nutr.* 38, 145–153 (1988).
70. Nykter, M., Kymäläinen, H. & Gates, F. Quality characteristics of edible linseed oil. *Agric. Food Sci.* 15, 402 (2006).
71. Gunstone, F. D. *Vegetable Oils in Food Technology: Composition, Properties and Uses* (2nd ed.). (Wiley-Blackwell, 2011).
72. Searchinger, T. D. T. et al. Assessing the efficiency of changes in land use for mitigating climate change. *Nat.* 564, 249–253 (2018).
73. Willett, W. et al. Food in the Anthropocene: the EAT–Lancet Commission on healthy diets from sustainable food systems. *Lancet* 393, 447–492 (2019).
74. Knez, M., Ranić, M. & Gurinović, M. Underutilized plants increase biodiversity, improve food and nutrition security, reduce malnutrition, and enhance human health and well-being. Let's put them back on the plate! *Nut. Rev.* 82, 1111–1124 (2023).
75. Akmeemana, C., Somendrika, M. D. & De A. Abeysundara, P. Underutilized plant oil sources: Fatty acid composition, physicochemical properties, applications, and health benefits. *In Elsevier eBooks* pp. 25–56 (2025).
76. Peiris, T. S. G., Fernando, M. T. N. & Samarajeewa, S. Factors influencing the use of coconut oil by the householders in Sri Lanka and their policy relevance to popularize the consumption of coconut oil. *CORD*, 20, 34 (2004).
77. Karunarathna, N. B., Fernando, C. J., Munasinghe, D. & Fernando, R. Occurrence of aflatoxins in edible vegetable oils in Sri Lanka. *Food Cont.* 101, 97–103 (2019).
78. Kittigowittana, K. et al. Fatty acid composition and biological activities of seed oil from rubber (*Hevea brasiliensis*) cultivar RRIM 600. *Inter. J. Appl. Res. Nat. Prod.* 6, 1–7 (2013).
79. Zhao, C. et al. Apple TIME FOR COFFEE contributes to freezing tolerance by promoting unsaturation of fatty acids. *Plant Sci.* 302, 110695 (2021).
80. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinform.* 30, 2114–2120 (2014).
81. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. DROPS (Schloss Dagstuhl – Leibniz Center for Informatics) (2013).
82. McKenna, A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303 (2010).
83. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* 6, 80–92 (2012).
84. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664 (2009).
85. Purcell, S. et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Amer. J. Human Genet.* 81, 559–575 (2007).

86. Tamura, K., Stecher, G. & Kumar, S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.* 38, 3022–3027 (2021).
87. Zhang, C., Dong, S., Xu, J., He, W. & Yang, T. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinform.* 35, 1786–1788 (2018).
88. Wang, J. & Zhang, Z. GAPIT Version 3: Boosting power and accuracy for genomic association and prediction. *Genom. Proteom. Bioinform.* 19, 629–640 (2021).
89. Huang, M., Liu, X., Zhou, Y., Summers, R. M. & Zhang, Z. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *Giga Sci.* 8, (2019).
90. Segura, V. et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830 (2012).
91. Gabriel, S. B. et al. The structure of haplotype blocks in the human genome. *Sci.* 296, 2225–2229 (2002).

Figures

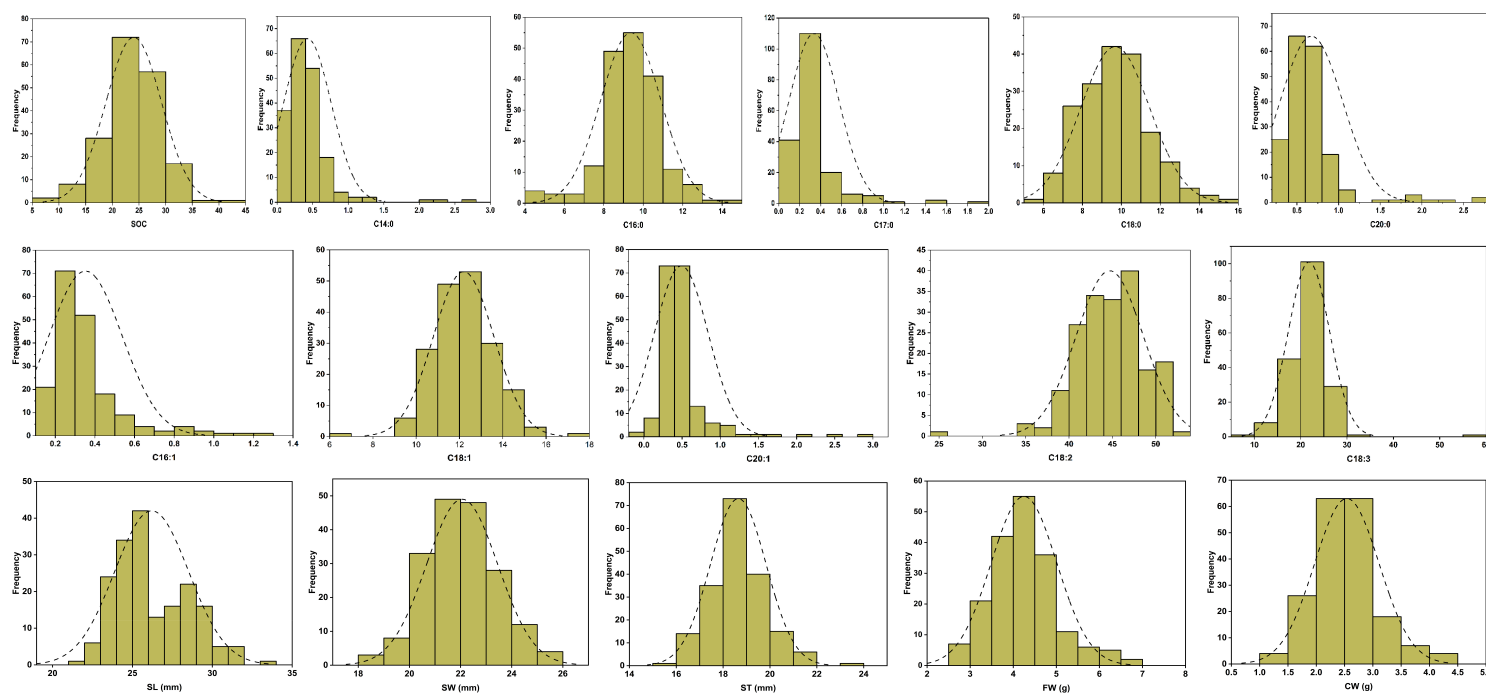


Figure 1

Frequency distribution of SOC, FA compositions, and seed size traits across 186 rubber tree accessions. Frequency distribution histograms illustrating phenotypic variation in SOC, FA composition traits, and seed size traits. Each histogram displays the frequency distribution of trait values, with the x-axis representing trait values and the y-axis representing the frequency within each class interval. The overlaid curve represents the fitted normal distribution for each trait. FA composition traits include myristic acid (C14:0), palmitic acid (C16:0), palmitoleic acid (C16:1), margaric acid (C17:0), stearic acid

(C18:0), oleic acid (C18:1), linoleic acid (C18:2), linolenic acid (C18:3), arachidic acid (C20:0), and eicosenoic acid (C20:1). Seed size traits include SL, SW, ST, FW, and CW. SOC, seed oil content; FA, fatty acid; SL, seed length; SW, seed width; ST, seed thickness; FW, fruit weight; CW, caryopsis weight.

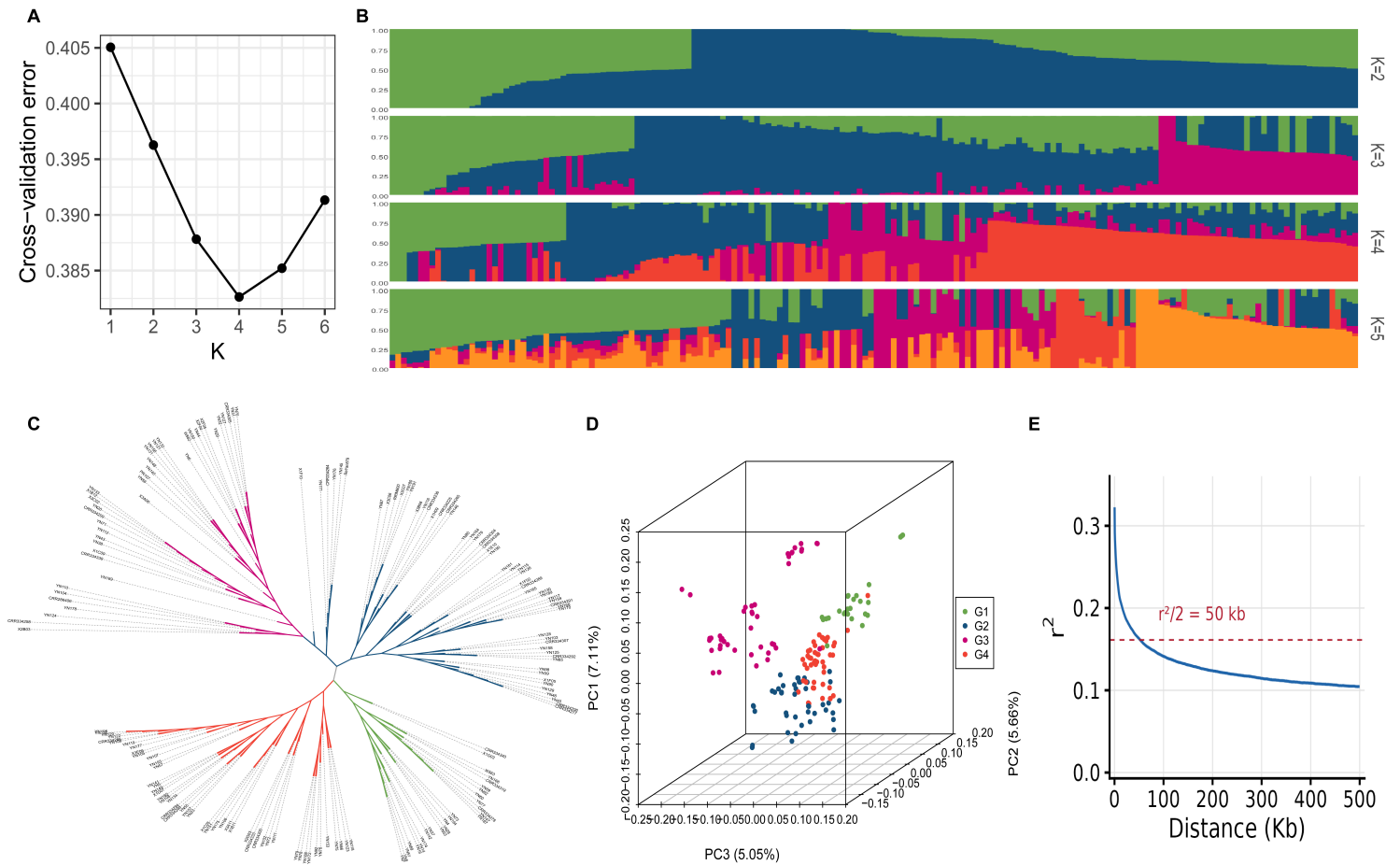


Figure 2

Population structure analysis of the 186 rubber tree accessions used for GWAS. **A** Cross-validation (CV) error plot for ADMIXTURE analysis across K=1 to 6 ancestral populations. The minimum CV error at K=4 indicates the optimal number of ancestral clusters. **B** ADMIXTURE bar plots for K=2 to K=5 showing the proportion of ancestry from each inferred ancestral population for each individual. Each vertical bar represents one individual, and colors indicate the proportion of ancestry assigned to each cluster. **C** Phylogenetic tree constructed from genome-wide SNP data, with branches color-coded according to ADMIXTURE cluster membership at K=4, providing independent validation of population subdivision. **D** Three-dimensional PCA scatter plot showing the distribution of individuals along the first three PCs. Colors correspond to the four clusters identified at optimal K=4. **E** LD decay plot showing the relationship between average pairwise R^2 values and physical genomic distance (kb) across all chromosomes. Each grey point represents an individual SNP pair R^2 value plotted against physical separation distance. The blue curve indicates the locally smoothed mean R^2 across all pairwise comparisons at each distance interval. The horizontal red dashed line indicates the R^2 threshold of 0.1. The maximum mean R^2 was 0.461, with LD_{50} of approximately 50 kb. The vertical blue line at 50 kb indicates the LD_{50} distance, which

was used as the basis for defining LD block boundaries and physical distance windows for independent QTL clustering in GWAS analyses. PCA, principal component analysis; LD, linkage disequilibrium; GWAS, genome-wide association study.

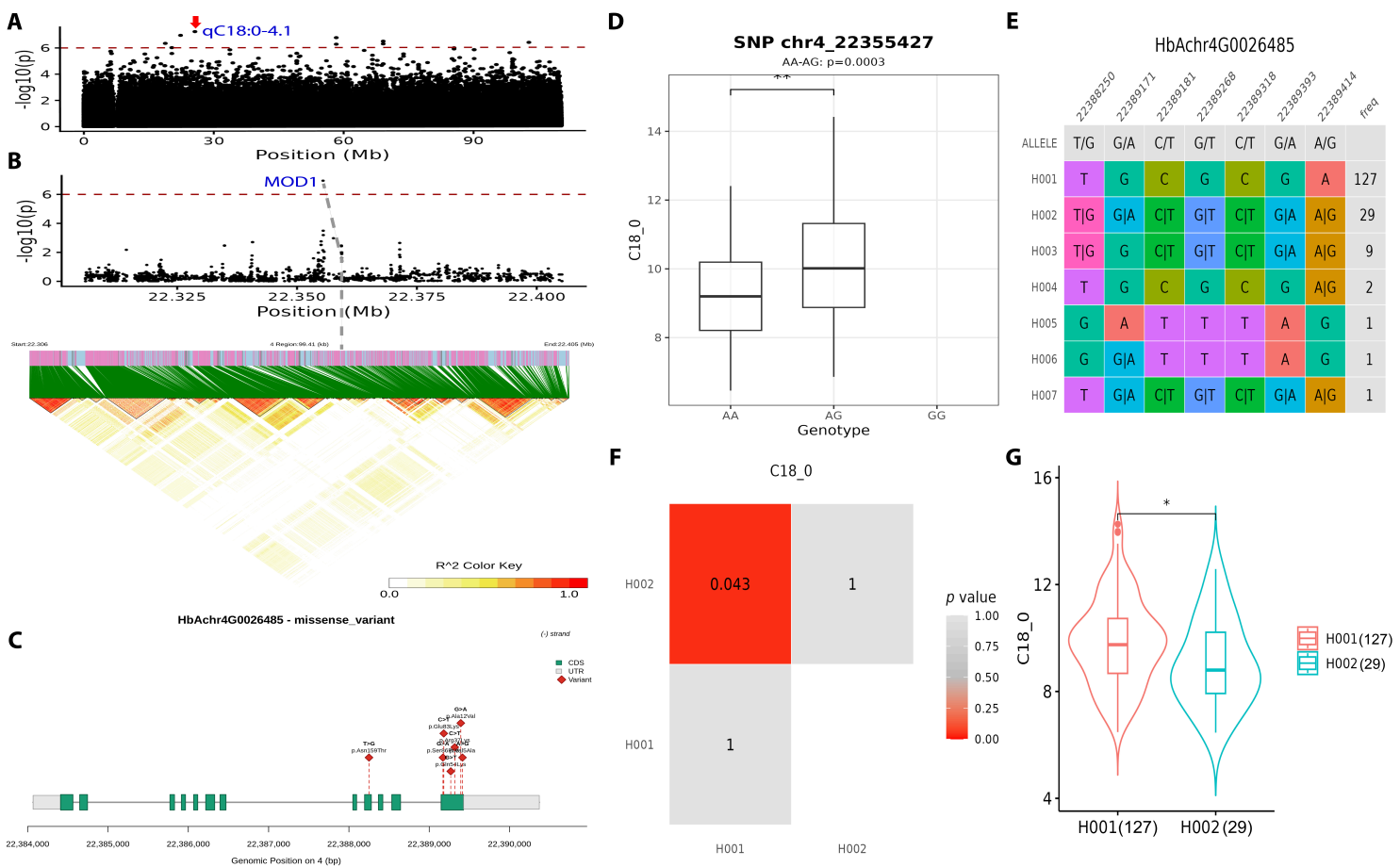


Figure 3

Fine mapping, candidate gene identification, and haplotype analysis of HbAchr4G0026485 at locus (qC18:0-4.1). **A** Genome-wide Manhattan plot showing association signals for C18:0 across all chromosomes, with $-\log_{10}(P\text{-value})$ on the y-axis and chromosomal position (Mb) on the x-axis. The red horizontal dashed line indicates the genome-wide significance threshold ($-\log_{10}(P) > 6$). The peak association signal at qC18:0-4.1 on Chr04 is indicated by the red arrow, with the QTL name labeled in blue. **B** Zoomed regional Manhattan plot of the qC18:0-4.1 interval (22.30–22.40 Mb) on Chr04, with the candidate gene MOD1 labeled in blue. The grey dashed lines connect the peak association signal in panel B to the LD heatmap. The LD heatmap below the regional Manhattan plot displays pairwise R^2 values among all SNPs within the displayed interval, color-coded from white/yellow to red according to the R^2 color key. **C** Gene structure of HbAchr4G0026485 on Chr04, spanning (22,384,000–22,390,000) bp. CDS are shown as teal blocks, UTR as grey blocks, and introns as connecting lines. Seven non-synonymous missense variants are annotated with red diamond symbols at their respective positions within the CDS. Genomic coordinates (bp) are indicated on the x-axis (22,384,000–22,390,000 bp). **D** Boxplot showing the genotype–phenotype association of SNP (4_22355427) with C18:0, comparing AA, AG, and GG genotypes.

AG, and GG genotypes. The y-axis represents C18:0 content. The horizontal line within each box indicates the median, box boundaries represent the 25th and 75th percentiles. AG heterozygous accessions exhibited significantly higher C18:0 compared to AA homozygous accessions (** $P=0.0003$). **E** Haplotype allele table for HbAchr4G0026485 showing allele combinations at seven non-synonymous SNP positions for each of the seven defined haplotypes and their corresponding frequencies. Reference alleles are shown in green and alternative/heterozygous alleles in pink/teal. H001 ($n = 127$) and H002 ($n = 29$) represent the two major haplotypes. **F** Pairwise haplotype–phenotype association matrix showing the P -value from the comparison of C18:0 between H001 and H002. The deep red color indicates a statistically significant difference ($P = 0.043$), as shown by the P -value color scale. **G** Violin plot with embedded boxplot comparing C18:0 between H001 (pink) and H002 (teal) haplotype groups. Violin width represents the kernel density distribution of C18:0 values, the embedded box represents the interquartile range, the horizontal line within the box indicates the median, and individual data points are shown as dots. MOD1, enoyl-[acyl-carrier-protein] reductase [NADH] chloroplastic; C18:0, stearic acid; CDS, coding sequence; UTR, untranslated region; LD, linkage disequilibrium; FA, fatty acid.

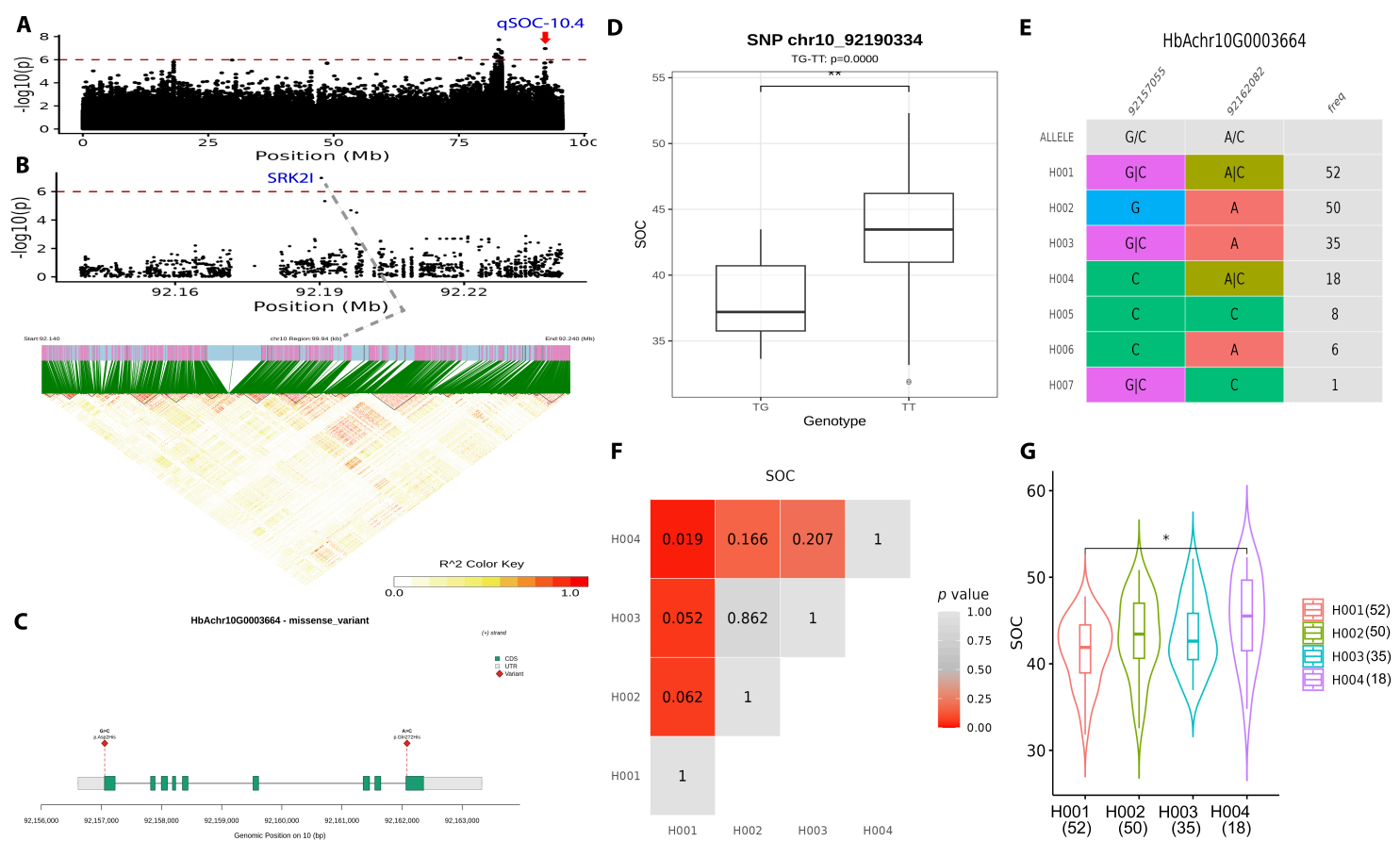


Figure 4

Fine mapping, candidate gene identification, and haplotype analysis of HbAchr10G0003664 at locus (qS0C-10.4). **A** Genome-wide Manhattan plot showing association signals for SOC across all chromosomes. The red horizontal dashed line indicates the genome-wide significance threshold ($-\log_{10}(P) > 6$). The peak association signal at qS0C-10.4 on Chr10 is indicated by the red arrow, with the

QTL name. **B** Zoomed regional Manhattan plot of the qSOC-10.4 interval (92.14–92.24 Mb) on Chr10, with the candidate gene SRK2I. The grey dashed lines connect the peak association signal in panel B to the LD heatmap. The LD heatmap below the regional Manhattan plot displays pairwise R^2 values among all SNPs within the displayed interval. **C** Gene structure of HbAchr10G0003664 on Chr10, spanning (92,156,000–92,163,000 bp). CDS are shown as teal blocks, UTR as grey blocks, and introns as connecting lines between exons. Two non-synonymous missense variants are annotated with red diamond symbols at their respective positions within the CDS. Genomic coordinates are indicated on the x-axis (92,156,000–92,163,000 bp). **D** Boxplot showing the genotype–phenotype association of the lead SNP (10_92190334) with SOC, comparing TG and TT genotypes. The y-axis represents SOC. The horizontal line within each box indicates the median, box boundaries represent the 25th and 75th percentiles, and open circles represent outliers. TT homozygous accessions exhibited significantly higher SOC compared to TG heterozygous accessions ($**P<0.0001$). **E** Haplotype allele table for HbAchr10G0003664 showing allele combinations at two non-synonymous SNP positions (92,157,053 bp and 92,162,082 bp) for each of the seven defined haplotypes and their corresponding frequencies. **F** Pairwise haplotype–phenotype association matrix showing P -values from comparisons of SOC among H001, H002, H003, and H004 haplotype groups. A statistically significant difference was observed between H004 and H001 ($P=0.019$), indicated by the deep red, while all other pairwise comparisons were non-significant. **G** Violin plot with embedded boxplot comparing SOC (y-axis) among H001 (pink), H002 (green), H003 (teal), and H004 (purple) haplotype groups. Violin width represents the kernel density distribution of SOC values, the embedded box represents the interquartile range, the horizontal line within the box indicates the median, and individual data points are shown as dots. SRK2I, serine/threonine-protein kinase SRK2I; SOC, seed oil content; CDS, coding sequence; UTR, untranslated region; LD, linkage disequilibrium.

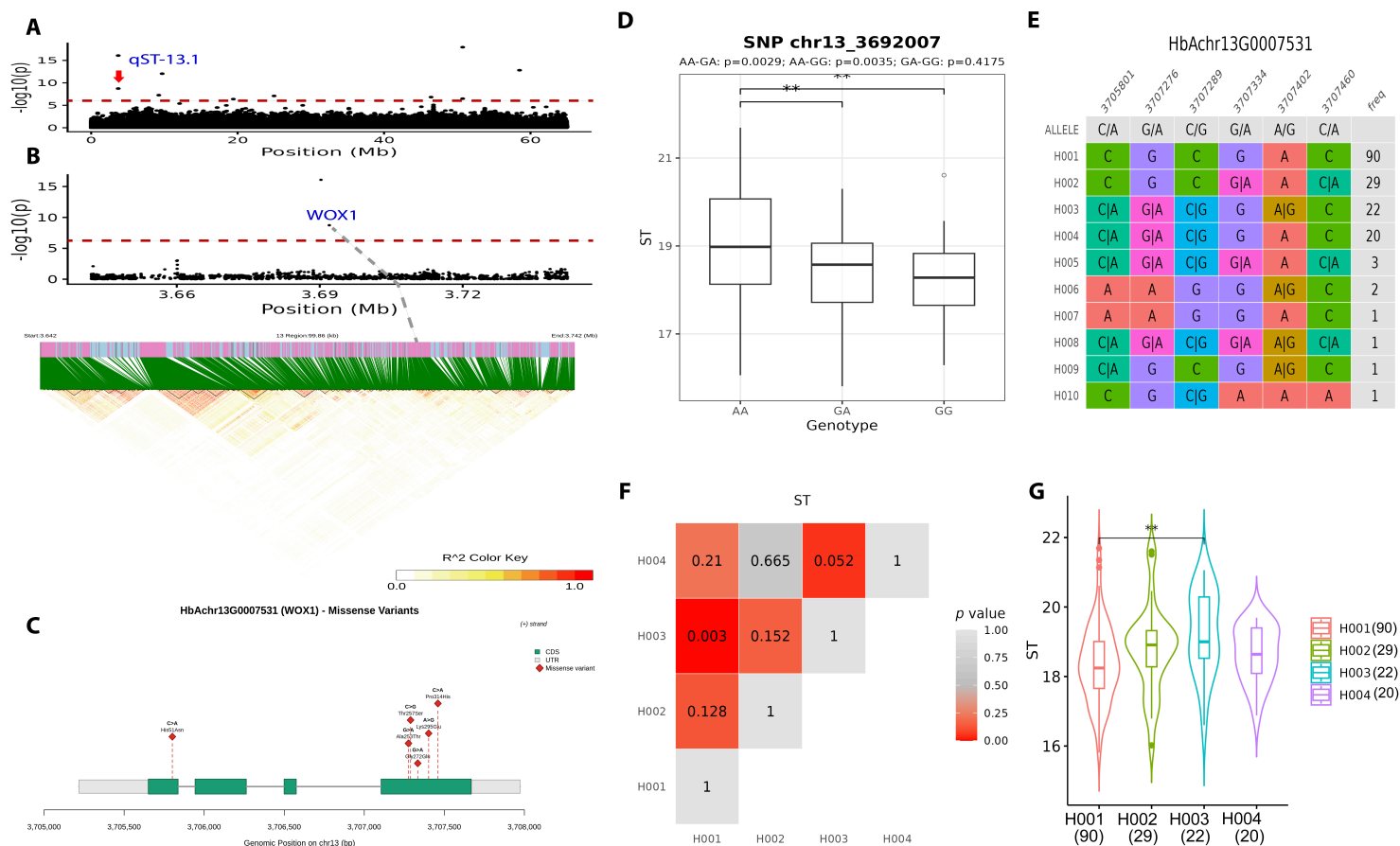


Figure 5

Fine mapping, candidate gene identification, and haplotype analysis of HbAchr13G0007531 at locus (qST-13.1). **A** Genome-wide Manhattan plot showing association signals for ST across all chromosomes, with $-\log_{10}(P\text{-value})$ on the y-axis and chromosomal position (Mb) on the x-axis. The red horizontal dashed line indicates the genome-wide significance threshold ($-\log_{10}(P) > 6$). The peak association signal at qST-13.1 on Chr13 is indicated by the red arrow, with the QTL name. **B** Zoomed regional Manhattan plot of the qST-13.1 interval (3.64–3.74 Mb) on Chr13, with the candidate gene WOX1. The grey dashed line in panel B connects the physical position of the WOX1 gene within the LD heatmap displayed below. The LD heatmap below the regional Manhattan plot displays pairwise R^2 values among all SNPs within the displayed interval (3.642–3.742 Mb), color-coded from white/yellow to red according to the R^2 color key. **C** Gene structure of HbAchr13G0007531 on Chr13, spanning (3,705,000–3,708,000 bp). CDS are shown as teal blocks, UTR as grey blocks, and introns as connecting lines between exons. Six non-synonymous missense variants are annotated with red diamond symbols at their respective positions within the CDS. Genomic coordinates are indicated on the x-axis (3,705,000–3,708,000 bp). **D** Boxplot showing the genotype–phenotype association of significant SNP (13_3692007) with ST, comparing AA, GA, and GG genotypes. The y-axis represents ST. The horizontal line within each box indicates the median, box boundaries represent the 25th and 75th percentiles, and open circles represent outliers. AA accessions exhibited significantly higher ST than both GA ($P=0.0029$) and GG ($P=0.0035$) accessions, while the GA–GG comparison was non-significant ($P=0.4175$), indicating an additive effect of the reference A allele on ST (** $P<0.01$). **E** Haplotype allele table for

HbAchr13G0007531 showing allele combinations at six non-synonymous SNP positions for each of the ten defined haplotypes and their corresponding frequencies. Reference alleles are shown in green and alternative/heterozygous alleles in pink/teal/yellow. **F** Pairwise haplotype–phenotype association matrix showing *P*-values from comparisons of ST among H001, H002, H003, and H004 haplotype groups. A highly significant difference was observed between H003 and H001 ($P=0.003$), indicated by the deep red cell, while all other pairwise comparisons were non-significant. The *P*-value color scale ranges from red to grey. **G** Violin plot with embedded boxplot comparing ST (y-axis) among H001 (pink), H002 (green), H003 (teal), and H004 (purple) haplotype groups. Violin width represents the kernel density distribution of ST values, the embedded box represents the interquartile range, the horizontal line within the box indicates the median, and individual data points are shown as dots. H003 accessions exhibited significantly higher ST compared to H001 accessions ($**P = 0.003$). WOX1, WUSCHEL-related homeobox 1; ST, seed thickness; CDS, coding sequence; UTR, untranslated region; LD, linkage disequilibrium.

Supplementary Files

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

- [Supplementarytables.xlsx](#)
- [Supplementaryfigures.pdf](#)