Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Genetic diversity and core SNP marker development for dragon fruit (Selenicereus spp.) germplasm identification and breeding

Juncheng Li
Guangdong Academy of Agricultural Sciences
Mengfei Lin
Jiangxi Academy of Sciences
Youjie Liu
Fujian Academy of Agricultural Sciences
Qingming Sun
254074711@qq.com

Guangdong Academy of Agricultural Sciences

Research Article

Keywords: Dragon fruit, Selenicereus spp., SNP markers, Genetic diversity, Germplasm identification, Betalain biosynthesis

Posted Date: May 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9587066/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

Dragon fruit (*Selenicereus spp.*) is a nutritious and economically important tropical and subtropical fruit crop, mainly cultivated in Central America, Southeast Asia and East Asia. However, the molecular basis of flesh color variation remains poorly understood. Traditional germplasm classification mainly relies on phenotypic traits such as fruit skin and flesh color and stem morphology, which often leads to misclassification due to phenotypic similarity and frequent interspecific hybridization, thus hindering efficient breeding and germplasm utilization. In this study, we used SLAF-seq to develop SNP markers and perform genetic characterization on 45 dragon fruit accessions, and obtained 1,206,498 high-quality SNPs after strict filtering. Genetic diversity analysis showed that the diversity levels of the three main taxa were significantly higher than the overall level. Population structure and phylogenetic analyses clearly distinguished three genetic groups and corrected several misclassified accessions. Furthermore, we constructed a hierarchical core SNP marker set, among which the 10-SNP ultra-core panel showed optimal discrimination ability with 100% accuracy in independent germplasm validation. This panel effectively overcame the misleading effect of flesh color phenotype caused by hybrid germplasm.

This study clarified the genetic diversity and population structure of cultivated dragon fruit, resolved germplasm classification confusion, and developed a set of efficient and diagnostic SNP markers. Furthermore, through phenotype-driven SNP screening and transcriptomic analysis, we identified four candidate genes (bHLH, CYP450, POD12, and MATE transporter) that show strong co-expression with key betalain biosynthesis genes (*CYP76AD1-1* and *DODA1*), providing molecular insights into flesh color regulation. These results provide a solid foundation for germplasm management, parental selection and molecular-assisted breeding of dragon fruit, and will promote the genetic improvement and sustainable development of this important horticultural crop.

1. Introduction

Dragon fruit (*Selenicereus* spp.), a commercially significant horticultural crop, is extensively cultivated in tropical and subtropical regions worldwide, with China and Vietnam emerging as leading producers. This genus, classified within the Cactaceae family, exhibits complex convergent evolutionary patterns related to ecological adaptations and phenotypic traits (Korotkova and Borsch et al., 2017). Frequent interspecific hybridization further complicates t genetic relationships among cultivated *Selenicereus* genotypes, presenting challenges for germplasm utilization and breeding. Until 2017, dragon fruit was commonly classified under the genus *Hylocereus*; however, Korotkova et al. provided compelling molecular evidence that prompted the reclassification of all *Hylocereus* species into *Selenicereus*, a change subsequently recognized by the International Plant Names Index (Croft and Cross et al., 1999). This genetic classification revision not only resolved longstanding debates but also underscored the limitations of phenotype-based classification, as high morphological similarity among closely related genotypes often leads to misidentification.

Traditionally, dragon fruit is divided into three main cultivated types based on key phenotypic traits: red-skinned white-fleshed (*S. undatus*), red-skinned red-fleshed (*S. polyrhizus*), and yellow-skinned white-fleshed (*S. megalanthus*) (Hernández and Salazar, 2012). However, the lack of reproductive isolation among these genotypes has facilitated the emergence of numerous natural and artificial interspecific hybrids, which exhibit diverse phenotypic variations and substantial hybrid vigor (Ha and Oanh et al., 2018). Consequently, hybrid breeding has become the predominant approach for the genetic improvement of dragon fruit (Tel-Zur, 2022). Despite these advancements, the entanglement of genetic relationships and phenotypic plasticity remains a significant barrier: morphological traits such as skin and flesh color, as well as floral characteristics, are often inadequate for accurate variety identification and quality assessment (Mercado-Silva, 2018). This limitation hampers effective parent selection in breeding programs and obstructs the rational utilization of germplasm resources, highlighting the urgent need to elucidate the genetic diversity, evolutionary relationships, and germplasm classification status of dragon fruit utilizing molecular tools.

Molecular markers have revolutionized research on genetic diversity and crop breeding, evolving through three generations of technological advancements (Hasan and Choudhary et al., 2021). First-generation markers, such as Restriction Fragment Length Polymorphism (RFLP), depend on molecular hybridization (Berry and Leon et al., 1995). In contrast, second-generation

markers—including Randomly Amplified Polymorphic DNA (RAPD), Amplified Fragment Length Polymorphism (AFLP), and Simple Sequence Repeats (SSR)—are PCR-based and have been extensively employed in assisted breeding (Powell and Morgante et al., 1996; Pan and Fu et al., 2017). Third-generation markers, particularly Single Nucleotide Polymorphisms (SNPs), offer unparalleled advantages, including high genome coverage, uniform distribution, reproducibility, accuracy, and cost-effectiveness (Jiang and Wen et al., 2022). These attributes render SNPs ideal for variety identification, genetic mapping, and population structure analysis (Allemailem and Almatroudi et al., 2021; Chen and He et al., 2022; Jiang and Wen et al., 2022), catalyzing their widespread adoption in horticultural research.

Specific-Locus Amplified Fragment Sequencing (SLAF-seq) is a high-throughput, cost-effective genome sequencing technology that facilitates efficient SNP discovery and genotyping (Xu and Lu et al., 2015). This approach has been successfully applied in SNP development and genetic analysis for various crops, including cucumber (Zhou and Zhou et al., 2017), Miscanthus (Chen and He et al., 2022), rapeseed (Zhao and Xiao et al., 2024), and sweetpotato (Tamby Chik and Bachok et al., 2011), demonstrating its versatility and reliability. In this study, we employed SLAF-seq to genotype 45 representative dragon fruit accessions from the Germplasm Nursery at the Guangdong Academy of Agricultural Sciences. Utilizing the *S. undatus* genome as a reference(Chen and Xie et al., 2021), we identified a set of high-consistency SNPs to analyze population structure, phylogenetic relationships, and genetic diversity. Our objectives were threefold: (1) to rectify misclassifications of dragon fruit accessions and resolve germplasm classification ambiguities; (2) to develop a set of core SNP markers for rapid and accurate germplasm and variety identification; (3) to provide a theoretical foundation for parent selection in hybrid breeding and molecular marker-assisted breeding. Additionally, we addressed the critical challenge of distilling high-density SNP data into minimal diagnostic marker sets—essential for practical applications in germplasm management and routine breeding diagnostics. By achieving these goals, this study aims to advance the scientific understanding of dragon fruit genetics and accelerate the development of improved varieties for the global market. Furthermore, to dissect the molecular basis of flesh color, we conducted a phenotype-driven SNP association analysis and integrated transcriptomic profiling to identify candidate genes co-expressed with betalain biosynthesis pathways. This study provides reliable molecular tools and a theoretical basis for the accurate identification, rational utilization and genetic improvement of dragon fruit germplasm, which is of great significance for the breeding and industrial development of this important horticultural fruit crop.

2. Materials and Methods

2.1. Plant Materials

Plant materials were obtained from the Germplasm Nursery at the Institute of Fruit Tree Research, Guangdong Academy of Agricultural Sciences, Guangzhou, China. Sample selection was guided by two primary criteria: (1) inclusion of the three principal phenotypic varieties of dragon fruit: red-skinned white-fleshed (*S. undatus*), red-skinned red-fleshed (*S. polyrhizus*), and yellow-skinned white-fleshed (*S. megalanthus*); (2) verification of well-documented origins for all varieties, confirmed by a variety committee or clear records of introduction sources. A total of 45 dragon fruit varieties were selected for this study. Notable accessions include "YWG" (2n = 4x = 44), a commercially cultivated tetraploid variety; "WCHL", characterized by yellow skin and white flesh (similar to "YWG"); "QP", which has green skin and white flesh; and "HJ", recognized by entirely yellow stems. The principal characteristics of the 45 samples are summarized in Table 1, with the "Currently classified" column representing genetic grouping results rather than formal taxonomic conclusions. Additionally, an extra 15 samples were selected from the germplasm nursery for subsequent validation of SNP markers, the selection of these 15 samples was based solely on the criterion of pulp color, including 5 white-fleshed, 5 red-fleshed, and 5 pink or bicolor-fleshed dragon fruit accessions. The representative phenotypic traits of the selected samples are depicted in Fig. 1.

Table 1
Phenotypic traits and taxonomic classification of 45 Selenicereus accessions.

ID	Fruit Fresh color	Fruit Skin color	Fruit spines Present	Species (Previously classified)	Species (Currently classified)
HLG	Red	Red	Yes	*S. polyrhizus*	*S. megalanthus*
HQL	Red	Red	Yes	Undefined	*S. megalanthus*
YWG	White	Yellow	Yes	*S. megalanthus*	*S. megalanthus*
GXDH	White	Red	No	*S. undatus*	*S. polyrhizus*
YXX	Red	Red	No	*S. polyrhizus*	*S. polyrhizus*
FSH	Red	Red	No	*S. polyrhizus*	*S. polyrhizus*
JML	Violet-red	Red	No	*S. costaricensis*	*S. polyrhizus*
TNSH	Red	Red	No	*S. polyrhizus*	*S. polyrhizus*
DH	Red	Red	No	*S. polyrhizus*	*S. polyrhizus*
DH2	Red	Red	No	*S. polyrhizus*	*S. polyrhizus*
JDYH	Red	Red	No	*S. polyrhizus*	*S. polyrhizus*
LHH	Red	Red	No	*S. polyrhizus*	*S. polyrhizus*
MLG	Red	Red	No	*S. polyrhizus*	*S. polyrhizus*
RZDH	Red	Red	No	*S. polyrhizus*	*S. polyrhizus*
XMEH	Red	Red	No	*S. polyrhizus*	*S. polyrhizus*
ZL	Violet-red	Red	No	*S. costaricensis*	*S. polyrhizus*
HJ	White	Red	No	Undefined	*S. undatus*
GZHL	Violet-red	Red	No	*S. polyrhizus*	*S. undatus*
WCHL	White	Yellow	No	Undefined	*S. undatus*
GHB	White	Red	No	*S. undatus*	*S. undatus*
HBSS	Pink	Red	No	Undefined	*S. undatus*
YGBR	White	Red	No	*S. undatus*	*S. undatus*
HSJ	Red	Red	No	*S. polyrhizus*	*S. megalanthus*×*S. polyrhizus*
HHDG	Red	Red	No	Undefined	*S. megalanthus*×*S. polyrhizus*
GBSJ	White	Red	Yes	*S. undatus*	*S. megalanthus*×*S. undatus*
XMYH	Red	Red	No	*S. polyrhizus*	*S. megalanthus*×*S. undatus*
FR	Pink	Red	No	Undefined	*S. megalanthus*×*S. undatus*
QP	White	Green	No	Undefined	*S. megalanthus*×*S. undatus*
WC	Red	Red	No	*S. polyrhizus*	*S. megalanthus*×*S. undatus*

Note: Previously assigned species are compared with current classifications based on genomic data. Hybrid origins are indicated by "×" (e.g., S. megalanthus × S. undatus), denoting admixed ancestry without specifying parental direction.

ID	Fruit Fresh color	Fruit Skin color	Fruit spines Present	Species (Previously classified)	Species (Currently classified)
XJTY	Red	Red	No	*S. polyrhizus*	*S. megalanthus*×*S. undatus*
XLSJ	White	Red	No	*S. undatus*	*S. megalanthus*×*S. undatus*
YH3	Pink	Red	No	Undefined	*S. megalanthus*×*S. undatus*
CXYH	Red	Red	No	*S. polyrhizus*	*S. megalanthus*×*S. undatus*
GXFR	Pink	Red	No	Undefined	*S. polyrhizus s*×*S. undatus*
GFHL	Pink	Red	No	Undefined	*S. polyrhizus s*×*S. undatus*
GHL1	Red	Red	No	*S. polyrhizus*	*S. polyrhizus s*×*S. undatus*
SHQ	Red	Red	No	*S. polyrhizus*	*S. polyrhizus s*×*S. undatus*
HFJR	Pink	Red	No	Undefined	*S. megalanthus*×*S. polyrhizus*×*S. undatus*
QLTH	White	Green	No	*S. undatus*	*S. megalanthus*×*S. polyrhizus*×*S. undatus*
DGXL	Red	Red	No	*S. polyrhizus*	*S. megalanthus*×*S. polyrhizus*×*S. undatus*
DQEH	Red	Red	No	*S. polyrhizus*	*S. megalanthus*×*S. polyrhizus*×*S. undatus*
HHDK	Red	Red	No	Undefined	*S. megalanthus*×*S. polyrhizus*×*S. undatus*
LHDG	White	Green	No	*S. undatus*	*S. megalanthus*×*S. polyrhizus*×*S. undatus*
XJYQ	Red	Red	No	*S. polyrhizus*	*S. megalanthus*×*S. polyrhizus*×*S. undatus*
PB	White	Red	No	*S. undatus*	*S. megalanthus*×*S. polyrhizus*×*S. undatus*

Note: Previously assigned species are compared with current classifications based on genomic data. Hybrid origins are indicated by "×" (e.g., S. megalanthus × S. undatus), denoting admixed ancestry without specifying parental direction.

(A–C) Representative morphological features of three main cultivated *Selenicereus* genotypes: (A) *S. undatus* ("PB"), (B) *S. polyrhizus* ("HSJ"), and (C) *S. megalanthus* ("YWG").

(D–G) Unique phenotypes observed in specific accessions: (D) green fruit peel of "QP", (E) yellow fruit peel of "WCHL", (F) pink/bicolor flesh of P1, (G) yellow stem of "HJ".

2.2. Reference Genome Selection and Restriction Digestion Strategy Assessment

The previously published genome of *Selenicereus undatus* (formerly *Hylocereus undatus*) was chosen as the reference genome (http://www.pitayagenomic.com/). To optimize the restriction endonuclease combination, electronic restriction digestion was conducted following four core principles of SLAF-seq: (1) minimizing digestion events in repetitive genomic regions; (2) ensuring uniform distribution of digested fragments across the genome; (3) verifying that fragment lengths post-digestion meet the designed specifications; and (4) ensuring the number of generated fragments meets experimental requirements. Additional considerations included the reference genome size (1.47 Gb) and GC content to enhance digestion efficiency.

2.3. DNA Extraction, SLAF Library Construction, and Sequencing

Fresh stem tissues of the selected plant materials were rapidly frozen in liquid nitrogen and manually ground into a fine powder. Genomic DNA was extracted using a polysaccharide-polyphenol plant genomic DNA extraction kit (Tiangen Biotech, Beijing, China). DNA purity and concentration were assessed using a NanoDrop™ One spectrophotometer (Thermo Fisher Scientific, USA). Qualified DNA samples underwent restriction digestion using the optimized enzyme combination (*RsaI* + *HaeIII*). Subsequently, SLAF library construction and high-throughput sequencing (paired-end 126 bp) were conducted by Biomarker Technologies Corporation (Beijing, China) on the Illumina HiSeq™ 2500 platform.

2.4. SLAF Tag Development and SNP Marker Identification

Sequencing reads were identified and demultiplexed for each sample using dual-index information. Raw reads underwent stringent quality control to remove adaptor sequences, low-quality reads (those with inadequate quality scores), and reads with a high proportion of unknown bases (N). Clean reads were clustered based on sequence similarity to group reads from the same genomic locus, discarding low-depth fragments to ensure data accuracy. All high-quality clean reads were aligned to the reference genome using BWA software (Li and Durbin, 2009). SNP markers were identified through two independent tools, GATK (Bathke and Lühken, 2021) and SAMtools (Danecek and Bonfield et al., 2021). Due to variations in algorithms, the results were not fully congruent; thus, only the intersection of findings from both methods was retained as a reliable SNP marker dataset. SNPs with an integrity greater than 0.8 and a minor allele frequency (MAF) exceeding 0.05 were selected for further analyses.

2.5. Genetic Diversity and Evolutionary Relationship Analyses

Genetic diversity and evolutionary relationships among the three main cultivated *Selenicereus* genotypes were analyzed using a high-consistency SNP dataset (1,206,498 markers), defined by a MAF threshold of 0.05 and site integrity of 0.8. Three complementary analytical approaches were employed:

1. **Population Structure Analysis**: Conducted using ADMIXTURE software(Alexander and Lange, 2011), with the assumed number of population clusters (K value) ranging from 1 to 10. Cross-validation error rates were calculated for each K value to determine the optimal number of clusters.
2. **Phylogenetic Tree Construction**: Performed using MEGA X software(Kumar and Stecher et al., 2018) with the neighbor-joining method, employing the Kimura 2-parameter model and 1000 bootstrap replicates for branch support assessment.
3. **Principal Component Analysis (PCA)**: Conducted using PLINK v2.0(Chen and Meng et al., 2019) to visualize genetic relationships and population stratification among samples.

2.6. Development of Hierarchical Core SNP Panels for Genetic Background Discrimination

A filtered VCF file containing 368,185 high-quality SNPs from the 45 accessions served as the initial dataset. Diagnostic SNPs were identified from 34 non-admixed individuals assigned by ADMIXTURE ($K = 2$, $Q \geq 0.8$), which were divided into two groups: Group 1 ($n = 20$, red-fleshed) and Group 2 ($n = 14$, white-fleshed). Candidate diagnostic SNPs were selected based on three stringent criteria: (1) pairwise Weir & Cockerham's FST > 0.5; (2) absolute allele frequency difference between groups > 0.6; and (3) population specificity (allele frequency > 0.7 in one group and < 0.3 in the other). The application of these filters resulted in 10,025 candidate SNPs, which underwent linkage disequilibrium (LD)-based pruning (window size = 50 kb, step size = 5 SNPs) to reduce redundancy. Pruned SNPs were ranked by a composite score integrating FST, allele frequency divergence, and population specificity, which allowed the extraction of hierarchical core SNP panels (10, 20, 50, 100, 200, and 500 SNPs).

2.7. Performance Evaluation of Core SNP Panels

The discriminatory power of each hierarchical core SNP panel was evaluated using three key metrics:

1. **Mean FST**: The average pairwise FST across all SNPs in the panel (± standard deviation), reflecting genetic differentiation between groups.
2. **Separation Score**: The mean absolute allele frequency difference between Group 1 (red-fleshed) and Group 2 (white-fleshed) across all loci.
3. **Mean Inter-Group PCA Distance**: Euclidean distance between the centroids of the two groups in the PC1–PC2 space derived from PCA of the panel.

PCA validation was performed for the entire set of 10,025 candidate SNPs as well as each hierarchical core panel using PLINK v2.0, employing the proportion of variance explained by PC1 and the spatial separation of the two groups as primary performance indicators.

2.8. Haplotype Reconstruction and Co-Segregation Analysis

For each hierarchical SNP panel, haplotypes were reconstructed across the 34 non-admixed samples. Genotypes were converted to alternative allele counts (0 = homozygous reference, 1 = heterozygous, 2 = homozygous alternative) and visualized as heatmaps using Seaborn (Waskom, 2021) (Python v3.10). Samples were ordered by phenotype (red-fleshed first), with row labels colored to distinguish red-fleshed (red) and white-fleshed (black) accessions. Complete co-segregation was defined as all red-fleshed samples sharing nearly identical haplotypes distinct from those of all white-fleshed samples.

2.9. Experimental Validation of the Ultra-Core 10-SNP Panel

To establish a robust molecular diagnostic standard, we first performed an internal validation using the 45 accessions from the discovery population. These accessions were selected based on their confirmed genetic backgrounds. Subsequently, an independent validation panel of 15 accessions, completely randomly selected from the germplasm nursery and not included in the discovery population, was used to test the established standard. This panel comprised 5 red-fleshed (R1–R5), 5 white-fleshed (W1–W5), and 5 pink or bicolor-fleshed accessions (P1–P5) confirmed as hybrids from crosses between red and white-fleshed parents. For both validations, genomic DNA was amplified by PCR using primers flanking each of the top 10 SNPs (Supplementary Table S1), and amplicons were subjected to Sanger sequencing for genotyping (Lalitha, 2000). Genotypes were classified as homozygous reference (0), heterozygous (1), or homozygous alternative (2) based on chromatogram inspection.

2.10. Identification of Flesh Color-Associated SNPs via Phenotype-Driven Screening

To dissect the genetic basis underlying flesh color variation, a targeted phenotype-driven screening was performed.

From the initial set of 1,206,498 high-consistency SNPs, markers specifically associated with the flesh color phenotype (Red vs. White) were identified using a case-control association analysis. A total of 437 SNPs showing significant association (based on a stringent p-value threshold and FST criterion) with the flesh color phenotype were retained for further analysis.

To evaluate the discriminative power of these 437 associated SNPs and to infer the genetic architecture (oligogenic vs. polygenic) of the trait, Principal Component Analysis (PCA) was performed using subsets of top-ranked SNPs (Top 5, 10, 50) and the entire set (All 437). This approach aimed to determine whether the red/white flesh color dichotomy is primarily governed by a few major loci or a broader polygenic background. Finally, the 437 flesh color-associated SNPs were mapped to the *S. undatus* reference genome to identify adjacent or overlapping candidate genes, and functional annotation was performed to predict their potential roles in metabolic pathways, particularly those related to betalain biosynthesis.

2.11. Transcriptomic Profiling and Functional Annotation of Candidate Genes

To validate the function of candidate genes identified from the SNP analysis, we utilized published RNA-seq data (NCBI Accession number, PRJNA704510) (Hua and Chen et al., 2021) from dragon fruit flesh tissues (Red flesh dragon fruit) at

different developmental stages (17, 23, 25, and 32 days after artificial pollination), with three biological replicates for each genotype.

The expression levels (TPM) of candidate genes located near the 437 SNPs were extracted. Subsequently, a correlation analysis was calculated between the expression levels of these candidate genes and two key betalain biosynthesis genes, *HU03G00480.1* (*CYP76AD1-1*) and *HU03G01342.1* (*DODA1*)(Hua and Chen et al., 2021). Genes showing significant differential expression and high correlation coefficients ($|r| > 0.6$) with the two key betalain biosynthesis genes were selected as key regulators.

3. Results

3.1. Optimal Restriction Enzyme Selection and SLAF Tag Definition

Based on the principles of SLAF-seq and a comprehensive assessment of the reference genome size (1.47 Gb) and GC content, the optimal restriction enzyme combination identified for this study was *RsaI* and *HaeIII*. To facilitate effective subsequent sequencing and marker development, SLAF tags were defined as digested fragments with lengths ranging from 414 to 464 bp.

3.2. Sequencing Data Quality and Alignment Performance

High-throughput sequencing was executed using a paired-end 126 bp strategy, yielding a total of 248.64 Mb of raw sequencing data. The number of reads per sample ranged from 2,540,192 to 12,363,790, ensuring adequate coverage for subsequent analyses. Quality assessment revealed that the GC content of the sequencing data varied from 39.79% to 41.78%, with an average of 40.71%, consistent with the GC characteristics of the dragon fruit genome. Q30 values (proportion of bases with Phred quality score ≥ 30) ranged from 89.03% to 94.23%, with an average of 92.13%, indicating high sequencing accuracy.

Subsequently, clean reads (after the removal of adaptor sequences and low-quality reads) were aligned to the *S. undatus* reference genome using BWA software (Li and Durbin, 2009). The alignment rates ranged from 96.24% to 99.42%, demonstrating a high degree of concordance between the sequencing data and the reference genome. These findings collectively confirm that the sequencing data are of high quality and that the reference genome selection is appropriate, providing a reliable foundation for subsequent SLAF tag development and SNP identification. Detailed quality and alignment statistics for each sample are presented in Supplementary Table S2.

3.3. Development and Distribution of SLAF Tags and SNP Markers

A total of 643,349 SLAF tags were successfully developed from the 45 dragon fruit accessions, of which 384,400 were polymorphic, accounting for 59.75% of the total tags. The average sequencing depth of the SLAF tags was 10.20×, ensuring the reliability of allele calling. Using the highest-depth sequence within each SLAF tag as the reference, a total of 5,368,737 potential SNP markers were identified across the entire population. After stringent filtering based on integrity (> 0.8) and minor allele frequency (MAF > 0.05), 1,206,498 high-consistency SNP markers were retained for subsequent genetic analyses.

The chromosomal distribution of SLAF tags and SNP markers is illustrated in Fig. 2, exhibiting a relatively uniform distribution across the genome, thereby validating the effectiveness of the SLAF-seq strategy. Detailed statistics for SLAF tags and SNP markers in each sample are provided in Supplementary Tables S3 and S4, respectively.

(A) Density of SLAF tags per 1-Mb window, showing uneven but broad coverage across all chromosomes. (B) Density of SNP markers per 1-Mb window, revealing higher variation and clustering in certain regions. Color scale indicates tag/SNP count; darker red = higher density.

3.4. Genetic and Evolutionary Relationships Analysis

Utilizing the high-consistency SNPs, a genomic-level analysis of population structure was performed for the 45 dragon fruit individuals using ADMIXTURE software (Alexander and Lange, 2011). The analysis indicated that the optimal number of

clusters was 2 (K = 2), corresponding to the lowest cross-validation error rate (Fig. 3). This suggests that the 45 dragon fruit individuals can be classified into two primary groups predominantly aligning with the red-fleshed and white-fleshed categories, with a few exceptions. When K = 3, the cross-validation error rate was only slightly higher than that at K = 2; at this level, the Q1 group from K = 2 was further divided into two subclusters. Specifically, the red-fleshed group separated into a distinct subcluster, while another included variety such as "YWG" of *S. megalanthus*, characterized by stems and fruit skin covered in dense, hard spines. These three clusters correspond to the three main cultivated genetic groups of dragon fruit reported in botanical literature (Fig. 4).

The lowest CV error occurs at K = 2, indicating optimal clustering of the population into two genetic groups. Error rates increase with higher K values, suggesting overfitting.

Clades a, b, and c represent *S. undatus*, *S. polyrhizus*, and *S. megalanthus* cultivated genetic groups, respectively. Color bars show ancestry proportions; samples are ordered by phylogenetic position.

Similarly, using the same high-consistency SNPs, we analyzed the genetic relationships among the samples through the neighbor-joining method in MEGA X(Kumar and Stecher et al., 2018). The analysis revealed that the 45 samples could be categorized into two major clusters, with the first cluster further divided into two subclusters, consistent with the population structure results. Consequently, the samples were clearly delineated as corresponding to *S. polyrhizus*, *S. undatus*, *S. megalanthus* cultivated genetic groups, and various interspecific hybrids (Fig. 4).

3.5. Genetic Diversity and Evolutionary Relationship Analyses

Population structure analysis of the 45 accessions was performed using the high-consistency SNP dataset with ADMIXTURE software. Cross-validation error rates were calculated for assumed cluster numbers (K) ranging from 1 to 10, with the lowest error rate observed at K = 2 (Fig. 3), indicating that the accessions can be primarily divided into two genetic groups. These groups largely correspond to the red-fleshed and white-fleshed phenotypic categories, with only a few exceptions. When K = 3, the cross-validation error rate was slightly higher than that at K = 2; this clustering pattern further subdivided the Q1 group (from K = 2) into two distinct subclusters. Specifically, red-fleshed accessions were separated into one subcluster (*S. polyrhizus*) and another subcluster comprising *S. megalanthus* accessions (e.g., "YWG"), characterized by dense, hard spines on stems and fruit skin. Thus, the K = 3 clustering result clearly distinguishes the three main cultivated genetic group of dragon fruit: *S. undatus*, *S. polyrhizus*, and *S. megalanthus* (Fig. 4).

Phylogenetic analysis was conducted using the neighbor-joining (NJ) method in MEGA X with 1000 bootstrap replicates. Consistent with the population structure results, the 45 accessions were clustered into two major clades, with the first major clade further divided into two subclades (Fig. 4). The first subclade included all *S. polyrhizus* accessions, the second subclade contained *S. megalanthus* accessions (including "YWG", "HLG", and "HQL"), and the second major clade consisted entirely of *S. undatus* accessions. Interspecific hybrids were distributed at the junctions of the major clades, reflecting their mixed genetic backgrounds. The bootstrap values for the major clades were generally high, confirming the reliability of the phylogenetic tree.

PCA was performed to further visualize the genetic relationships among the accessions. The results indicated that the first two principal components (PC1 and PC2) explained 22.38% and 11.17% of the total genetic variation, respectively (Supplementary Figure S1). The 45 accessions were distinctly separated into three groups corresponding to *S. undatus*, *S. polyrhizus*, and *S. megalanthus*, which is consistent with the population structure and phylogenetic tree results. This further confirms the clear genetic differentiation among these three main cultivated genotypes.

3.6. Core SNP Panel Development and Performance Evaluation

To develop diagnostic SNP panels for practical applications, 34 non-admixed individuals were selected based on the ADMIXTURE results (K = 2, Q ≥ 0.8), including 20 red-fleshed accessions (Group 1) and 14 white-fleshed accessions (Group 2). Candidate diagnostic SNPs were filtered using three stringent criteria: pairwise Weir & Cockerham's FST > 0.5, absolute allele frequency difference > 0.6, and population specificity (frequency > 0.7 in one group and < 0.3 in the other). The intersection of these filters yielded a total of 10,025 candidate SNPs. After LD-based pruning (window = 50 kb, step = 5 SNPs), hierarchical

core SNP panels (10, 20, 50, 100, 200, and 500 SNPs) were extracted based on a composite score integrating FST, allele frequency divergence, and specificity.

Three key metrics were utilized to evaluate the discriminatory power of each core panel (Table 2):

1. **Mean FST**: The ultra-core 10-SNP panel exhibited the highest mean FST (1.000 ± 0.000), indicating fixed allelic differences between the two groups at all 10 loci. As panel size increased, the mean FST gradually decreased to 0.705 for the 500-SNP panel due to the inclusion of less differentiated SNPs.
2. **Separation Score**: The ultra-core 10-SNP panel achieved the highest separation score (0.995), reflecting nearly perfect allele frequency divergence between the two groups. The separation score remained high (> 0.800) even for the 500-SNP panel.
3. **Mean Inter-Group PCA Distance**: The Euclidean distance between the group centroids in the PC1–PC2 space remained consistently around 0.346–0.348 across all panels, confirming stable genetic separation regardless of the number of markers.

Table 2
Performance metrics of hierarchical core SNP panels for discriminating two major genetic groups in *Selenicereus*

SNP Panel	No. of SNPs	Mean FST(SD)	Separation Score	Mean Inter-Group PCA Distance
ultra_core_10	10	1.000(0.000)	0.995	0.348
ultra_core_20	20	0.994(0.014)	0.981	0.347
core_50	50	0.949(0.045)	0.949	0.347
core_100	100	0.889(0.070)	0.914	0.346
extended_200	200	0.821(0.087)	0.873	0.345
extended_500	500	0.705(0.114)	0.801	0.346
Note: Separation Score = mean absolute allele frequency difference between Group 1 and Group 2.				

PCA validation demonstrated that all core panels maintained excellent discriminatory power (Supplementary Figure S2). For the ultra-core 10-SNP panel, PC1 explained 97.8% of the total variation, achieving complete non-overlapping clustering of the two groups (Fig. 5B). For larger panels, PC1 explained 97.12% (20 SNPs), 94.82% (50 SNPs), 95.28% (100 SNPs), 94.92% (200 SNPs), and 93.33% (500 SNPs), with no intermingling of individuals in any case. These results indicate that the core SNP panels, particularly the ultra-core 10-SNP panel, can effectively recapitulate the population structure using minimal markers.

Samples are colored by flesh color phenotype: red = red-flesh, cyan = white-flesh. (A) Shows clear separation along PC1 (91.57% variance), with minor overlap. (B) Reveals near-perfect clustering along PC1 (97.8% variance), indicating strong genetic differentiation driven by the ultra-core SNPs.

3.7. Haplotype Reconstruction and Phenotype Co-Segregation Analysis

Haplotype analysis was conducted for each core SNP panel across the 34 non-admixed samples. Genotypes were converted to alternative allele counts (0 = homozygous reference, 1 = heterozygous, 2 = homozygous alternative) and visualized as heatmaps. For the ultra-core 10-SNP panel, all 20 red-fleshed samples shared a nearly identical haplotype dominated by the alternative allele (red cells), while all 14 white-fleshed samples were uniformly homozygous for the reference allele (white cells) (Fig. 6). This perfect genotype-phenotype co-segregation pattern was consistently observed across all six core panels, with increased SNP density revealing only minor within-group genetic variation, but maintaining the fundamental red/white flesh dichotomy (Supplementary Figure S3). These results confirm that the selected core SNPs are tightly linked to the genetic basis of fruit flesh color.

(A) Genotype heatmap of 10 ultra-core SNPs showing strong allele frequency divergence between red-flesh (top, n = 20) and white-flesh (bottom, n = 14) samples. (B) Genotype heatmap of 500 extended SNPs confirming consistent genetic differentiation across a broader genomic region. Each cell represents ALT allele count (0: white, 1: pink, 2: red). Samples are ordered by phenotype.

3.8. Experimental Validation of the Ultra-Core 10-SNP Panel

To establish a reliable molecular diagnostic criterion, we first analyzed the genotypes of the 45 discovery accessions with confirmed pure genetic backgrounds at the 10 ultra-core SNP loci. The analysis revealed that among the 14 white-fleshed accessions, 11 were homozygous for the reference allele (0) at all 10 loci, 1 was homozygous at 9 loci, and the remaining 2 were homozygous at 8 loci. Thus, the minimum number of 0 genotypes observed was 8. Among the 20 red-fleshed accessions, 16 were homozygous for the alternative allele (2) at all 10 loci, and the other 4 were homozygous at 9 loci, with a minimum of 9 such genotypes. The remaining 11 admixed accessions displayed a mixture of 0, 1, and 2 genotypes across the loci.

To balance discriminatory accuracy with tolerance for potential minor genotyping errors or rare recombination events, we established the following molecular typing criteria:

1. If a sample exhibits the homozygous reference genotype (0) at ≥ 7 SNP loci, it is classified as having a white-flesh genetic background.
2. If a sample exhibits the homozygous alternative genotype (2) at ≥ 7 SNP loci, it is classified as having a red-flesh genetic background.
3. If a sample shows heterozygous genotypes (1) at multiple loci, or a combination of homozygous genotypes that does not meet either of the above pure-line criteria, it is classified as a hybrid between red- and white-fleshed types.

The complete genotype data for these 45 accessions are provided in Supplementary Table S5.

We then applied this established criterion to an independent, completely randomized validation panel of 15 accessions (5 phenotypically white-fleshed: W1–W5; 5 phenotypically red-fleshed: R1–R5; and 5 phenotypically pink/bicolor-fleshed: P1–P5). Genotyping at the 10 ultra-core SNP loci revealed notable discrepancies between flesh color phenotype and underlying genetic background (Table 3):

Among the 5 phenotypically white-fleshed accessions, only 3 met the molecular criterion for a white-flesh genetic background (≥ 7 loci of 0/0): W1 was homozygous reference (0/0) at all 10 loci, W2 at 9 loci, and W3 at 7 loci. The remaining two (W4 and W5) exhibited mixed genotypes inconsistent with pure white ancestry and were classified as red–white hybrids.

Among the 5 phenotypically red-fleshed accessions, 3 met the criterion for a red-flesh genetic background (≥ 7 loci of 1/1, coded as “2” in Table 3): R1 and R2 each had 9 loci of 1/1, and R3 had 8 loci. The other two (R4 and R5) displayed hybrid-like genotype patterns and were classified as red–white hybrids.

Table 3
Genotypes at the 10 ultra-core SNP loci in 15 validation accessions with distinct flesh colors.

Position	Ref/Alt	W1	W2	W3	W4	W5	R1	R2	R3	R4	R5	P1	P2	P3	P4	P5
chr04:101320703	A/T	0	0	0	0	1	2	1	2	2	2	1	1	1	1	1
chr05:27106429	G/A	1	1	0	2	0	2	2	0	2	2	0	0	1	2	1
chr05:81783223	A/G	0	2	0	1	1	1	2	2	1	2	2	2	1	1	2
chr07:35138978	A/G	0	2	0	0	0	2	2	1	1	2	2	0	1	0	1
chr07:56371701	T/C	1	0	0	1	0	2	1	0	2	1	0	2	2	1	1
chr07:58485677	G/A	0	2	0	0	0	2	2	2	0	2	0	0	1	1	2
chr07:113215963	G/A	0	1	0	0	0	2	2	1	2	2	0	2	0	1	0
chr08:71145652	G/A	0	0	0	2	0	2	2	2	1	2	2	1	1	1	1
chr09:60199790	C/T	0	1	0	2	0	2	2	0	0	2	2	0	0	2	1
chr11:43495628	T/G	1	0	0	0	1	2	2	2	2	2	1	2	1	1	1

Note: "Ref/Alt" denotes the reference and alternative alleles, respectively. Samples are grouped by flesh color: W1–W5 = white-fleshed, R1–R5 = red-fleshed, and P1–P5 = pink or bicolor-fleshed accessions derived from crosses between red- and white-fleshed dragon fruit. Genotype codes: 0 = homozygous for the reference allele (associated with white flesh), 1 = heterozygous, 2 = homozygous for the alternative allele (associated with red flesh). All samples were genotyped via Sanger sequencing after PCR amplification.

All 5 phenotypically pink/bicolor-fleshed accessions (P1–P5) showed complex genotype combinations with multiple heterozygous (0/1) or mixed homozygous states across loci, fully consistent with their expected status as F1 hybrids.

These results demonstrate that visual assessment of flesh color alone can be misleading for inferring true genetic background, whereas the ultra-core 10-SNP panel provides an objective and accurate molecular diagnosis.

The experimental validation confirmed that the ultra-core 10-SNP panel is a robust tool for distinguishing the genetic background (ancestry) of pitaya accessions. However, these background-informative markers are not necessarily linked to the causal loci controlling the flesh color phenotype.To uncover the specific genetic architecture underlying flesh color, a phenotype-driven screening approach was therefore employed, shifting the focus from population structure to trait association.

3.9. Genetic Architecture of Flesh Color Revealed by Targeted SNP Analysis

Unlike the core SNP panels developed for germplasm identification (which rely on population structure), we sought to identify SNPs specifically driving the flesh color phenotype. From the genome-wide dataset, we filtered 437 SNPs significantly associated with flesh color.

PCA was performed using different numbers of these top-associated SNPs (Top 5, 10, 50, and All 437) to observe the separation between red- and white-fleshed samples. A “signal dilution” effect was observed. While the Top 5 SNPs provided the clearest separation, the result remained unreliable. As the number of SNPs increased (e.g., Top 10 or 50), the separation effect deteriorated, and the samples from the two phenotypic groups became severely intermingled (Supplementary Figure S4). This finding suggests that flesh color in dragon fruit is likely controlled by a few major genes (oligogenic inheritance) rather than a highly polygenic background. Consequently, using a small set of unvalidated SNPs to distinguish between the two flesh color groups is impossible. Therefore, investigating the function of the candidate genes may be more effective.

Annotation of the 437 flesh color-associated SNPs revealed 51 candidate genes, which are potential regulators or structural enzymes involved in the betacyanin biosynthesis pathway.

3.10. Identification of Key Regulators via Transcriptomic Correlation

To pinpoint the causal genes among the 51 candidates, we analyzed public transcriptome data, focusing on genes whose expression levels correlated with key dragon fruit betalain biosynthesis genes. Four key genes showed high correlation coefficients with the key betalain biosynthesis genes *HU03G00480.1* (*CYP76AD1-1*) and *HU03G01342.1* (*DODA1*) (r = 0.63–0.78) (Supplementary Table S6) were identified as strong candidates:

bHLH Transcription Factor (HU10G00331.1, **SNP:chr10/3558278, T/C)**: Likely acts as a master switch (part of the MBW complex) to initiate the transcription of the betacyanin pathway.

Cytochrome P450 (HU06G00398.1, **SNP: chr06/4340309, G/A)**: Likely the functional enzyme (similar to CYP76AD1) catalyzing the key hydroxylation step in pigment synthesis.

Peroxidase 12 (HU11G01046.1, **SNP: chr11/85685778, G/C)**: Likely protects unstable betacyanins from oxidative degradation by scavenging hydrogen peroxide.

MATE Transporter (HU05G00589.1, **SNP: chr05/14715116, C/T)**: Likely responsible for pumping pigments into the vacuole for storage.

In summary, through an integrated analysis of population genetics, transcriptomics, and comparative genomics, we have successfully narrowed down the genetic control of flesh color in dragon fruit from thousands of SNPs to four high-confidence candidate genes. These genes represent compelling targets for future functional validation and may serve as valuable markers for marker-assisted breeding of fruit color in *Selenicereus*.

4. Discussion

The population structure and phylogenetic analyses presented in this study establish a robust framework for understanding the genetic classification and evolutionary relationships among cultivated dragon fruit genotypes. The results from the ADMIXTURE analysis identified K = 2 as the optimal number of clusters, correlating with the lowest cross-validation error rate (Fig. 3). This analysis suggests that *Selenicereus polyrhizus* and *Selenicereus megalanthus* are grouped together in a single cluster, while *Selenicereus undatus* forms a distinct cluster, indicating a closer genetic relationship between the *S. polyrhizus* and *S. megalanthus* compared to their relationship with *S. undatus*. Notably, the cross-validation error rate at K = 3 was only slightly higher than at K = 2. At K = 3, the three main cultivated genetic groups (*S. polyrhizus*, *S. undatus*, and *S. megalanthus*) were clearly differentiated into independent clusters (Fig. 4), aligning closely with the traditional phenotypic classification of cultivated dragon fruit (Oo and Lynn et al., 2023).

Phylogenetic analysis further supports these findings, revealing two primary groupings among the 45 accessions. The first group subdivided into two subclusters corresponding to the *S. polyrhizus* and *S. megalanthus* cultivated genotypes, while the second group represented *S. undatus* cultivated genotype (Fig. 4). From an evolutionary perspective, *S. polyrhizus* and *S. undatus* exhibit a shorter genetic distance relative to one another compared to either genotype's distance from *S. megalanthus*. This evolutionary pattern is reinforced by phenotypic evidence: *S. megalanthus* is distinctively characterized by robust, hard spines on its flowers, fruits, and stems (Al Bkhetan and Wang, 2024), whereas *S. polyrhizus* and *S. undatus* share similarities in stem morphology, differing mainly in flesh color (red for *S. polyrhizus* and white for *S. undatus*) (Al Bkhetan and Wang, 2024). Collectively, these results highlight the complex interplay between genetic and phenotypic traits in dragon fruit, underscoring the necessity for an integrated germplasm classification approach that merges molecular data with morphological observations.

A key insight from our study is the inadequacy of traditional phenotype-based classification, which has historically relied on skin and flesh color as primary diagnostic traits (Hernández and Salazar, 2012; Mercado-Silva, 2018; Al Bkhetan and Wang, 2024). For example, *S. megalanthus* has commonly been associated with the label "yellow dragon fruit" due to its characteristic yellow skin (Hernández and Salazar, 2012; Mercado-Silva, 2018; Al Bkhetan and Wang, 2024). However, our molecular analyses

reveal that the red-skinned, red-fleshed accessions 'HLG' and 'HQL' cluster closely with the yellow-skinned, white-fleshed 'YWG' within the *S. megalanthus* clade (Fig. 4). Conversely, 'WCHL'—a yellow-skinned, white-fleshed accession (Fig. 1E) traditionally defined as *S. megalanthus*—was genetically grouped with *S. undatus*. Similarly, 'HJ', which exhibits entirely yellow stems (Fig. 1G), was also classified as *S. undatus* through population structure and phylogenetic analyses.

These observations lead us to hypothesize that yellow pigmentation in dragon fruit (whether in skin, flesh, or stems) may result from single or multiple gene mutations affecting betaxanthin synthesis (Danecek and Bonfield et al., 2021), rather than being a cultivar specific trait. This hypothesis is further substantiated by the green-skinned accession 'QP' (Fig. 1D), identified as an interspecific hybrid between *S. megalanthus* and *S. undatus*—its unique skin color likely reflecting differential regulation of betalain biosynthesis. These findings challenge the long-standing reliance on single phenotypic traits for dragon fruit germplasm classification and highlight the importance of incorporating molecular markers to complement morphological characteristics in germplasm identification.

Additionally, our study addresses several long-standing germplasm classification ambiguities. The accession 'ZL' was previously assigned to *Selenicereus costaricensis*, a phenotypic variant of red-fleshed dragon fruit first described in Costa Rica (Kumar and Stecher et al., 2018; Bathke and Lühken, 2021) and characterized by deep red to violet-red flesh. Despite its morphological similarities to other red-fleshed dragon fruits, 'ZL' was tentatively classified as a distinct phenotypic type without molecular evidence, often regarded as a variant of *S. polyrhizus*. Our phylogenetic and population structure analyses provide the necessary evidence to confirm that 'ZL' is indeed a member of the *S. polyrhizus* genetic group, clarifying the genetic affiliation of the *S. costaricensis* phenotypic variant with *S. polyrhizus*. This resolution simplifies the germplasm classification framework for red-fleshed dragon fruit and reduces confusion in germplasm management.

Another significant finding pertains to the ploidy of the commercially valuable tetraploid variety 'YWG', known for its superior fruit quality (Kumar and Stecher et al., 2018). Traditionally hypothesized to be an allotetraploid, 'YWG' was hypothesized to result from hybridization between *S. polyrhizus* and *S. undatus*, followed by chromosomal doubling (Hernández-Langford and Mandujano et al., 2024). However, our molecular data contradict this hypothesis: 'YWG' clusters closely with 'HLG' and 'HQL' within the *S. megalanthus* genetic group, demonstrating significant genetic divergence from both *S. polyrhizus* and *S. undatus* (Fig. 4). Population structure, phylogenetic, and PCA analyses consistently support the genetic grouping of these three accessions as *S. megalanthus*. We thus infer that 'YWG' is a tetraploid genotype with a genetic background closely related to *S. megalanthu*. This finding carries important implications for breeding, as tetraploids with different genetic backgrounds exhibit distinct genetic characteristics and breeding potentials—genotypes with a single genetic background may retain more traits of their ancestral diploids, while interspecific hybrid tetraploids often display increased heterosis. Our finding offers vital guidance for the rational utilization of 'YWG' in polyploid breeding programs.

A major strength of this study is the development and validation of hierarchical core SNP panels for practical dragon fruit identification. While genome-wide SNP markers are powerful for population structure inference (Browning and Browning, 2011), their high density can exceed the requirements for routine diagnostics(Browning and Browning, 2011). To address this, we developed a tiered diagnostic SNP panel set, filtering 368,185 high-quality SNPs down to 10,025 candidate diagnostic SNPs based on stringent criteria (FST > 0.5, allele frequency difference > 0.6, and population specificity). After LD-based pruning, we derived hierarchical panels of 10, 20, 50, 100, 200, and 500 SNPs.

Performance evaluations indicated that the ultra-core 10-SNP panel achieved maximal discriminatory power, with a mean FST of 1.000 (indicating fixed allelic differences between red-fleshed and white-fleshed groups) (Da Silva Ribeiro and Galván et al., 2022) and a separation score of 0.995 (Table 2) (Chagné and Vanderzande et al., 2019). PCA validation confirmed that this ultra-core panel recapitulated the population structure observed with the complete set of 10,025 SNPs, with PC1 explaining 97.8% of the total variation and achieving complete non-overlapping clustering (Fig. 5B). Haplotype analysis further substantiated perfect co-segregation between the ultra-core SNPs and flesh color: all red-fleshed samples shared a distinct haplotype dominated by the alternative allele, while white-fleshed samples were uniformly homozygous for the reference allele (Fig. 6).

Experimental validation using an independent, completely randomized panel of 15 accessions—comprising 5 phenotypically white-fleshed, 5 red-fleshed, and 5 pink/bicolor-fleshed individuals—revealed both the power and necessity of the ultra-core 10-SNP panel. Contrary to the assumption of perfect genotype–phenotype concordance, molecular profiling uncovered significant misalignment: among the five accessions visually classified as white-fleshed, two (W2, W4) were genetically identified as red–white hybrids; similarly, two of the five red-fleshed accessions (R3, R4) also exhibited hybrid genotypes. Only three white- and three red-fleshed accessions met the strict molecular criteria for pure genetic backgrounds (≥ 7 loci of Ref/Ref or Alt/Alt, respectively). In contrast, all five pink/bicolor-fleshed accessions displayed expected hybrid genotype patterns with extensive heterozygosity or mixed homozygous states across the 10 loci (Table 3).

These findings critically reinforce our earlier observation that flesh color—like other vegetative traits such as stem or peel pigmentation—can be a poor proxy for true genetic background, likely influenced by a limited number of regulatory genes, epistatic interactions, or environmental modulation. Relying solely on visual phenotyping may therefore lead to erroneous conclusions in breeding decisions or germplasm curation. The ultra-core 10-SNP panel overcomes this limitation by providing an objective, DNA-based diagnostic standard. Its ability to unambiguously distinguish pure lines from admixed individuals makes it a rapid, cost-effective, and reliable tool not only for variety identification and seed purity testing but also for correcting phenotypic misclassifications—a crucial advancement for precision breeding and accurate germplasm management in dragon fruit.

Despite these advancements, our study has notable limitations that warrant future investigation. First, while we hypothesize that yellow pigmentation in dragon fruit arises from gene mutations regulating betalain biosynthesis, functional genetic studies were not performed to validate this mechanism. Future research should integrate transcriptomic and metabolomic analyses to identify key genes controlling skin and flesh color, as well as stem morphology, thus providing a molecular basis for our phenotypic observations. Second, our germplasm classification revisions are based on 45 accessions; expanding the sample size to include more diverse germplasm (e.g., wild relatives, accessions from other regions, and additional tetraploid varieties) will further refine the germplasm classification system and enhance its generalizability. Third, while the ultra-core 10-SNP panel effectively discriminates between red-fleshed and white-fleshed accessions, developing species-specific panels for *S. megalanthus*, *S. polyrhizus*, and *S. undatus* will augment the utility of molecular markers in germplasm conservation and breeding.

Another area for future exploration is the genetic basis of hybrid vigor in dragon fruit. Hybrid breeding has emerged as the primary method for dragon fruit improvement (Chagné and Vanderzande et al., 2019), with both intraspecific and interspecific hybridization contributing to enhanced yield, quality, and stress tolerance(Ha and Oanh et al., 2018). However, the genetic mechanisms underlying hybrid vigor in dragon fruit remain poorly understood. Our high-consistency SNP dataset presents a valuable resource for identifying quantitative trait loci (QTLs) associated with key agronomic traits (e.g., fruit size, sugar content, and disease resistance) and investigating the role of heterozygosity in hybrid performance.

In conclusion, this study significantly enhances our understanding of dragon fruit genetic diversity, germplasm classification, and evolutionary relationships. Leveraging SLAF-seq technology, we identified 1,206,498 high-consistency SNPs that facilitated accurate germplasm classification correction, including the genetic assignment of 'HLG', 'HQL', and 'ZL', as well as the clarification of 'YWG"s genetic background. We also developed a validated ultra-core 10-SNP panel that serves as a practical tool for rapid variety identification and hybrid authentication. These findings challenge traditional phenotype-based germplasm classification and provide essential technical support for future genetic improvement and breeding programs. The integrated approach presented here—combining population genomics, germplasm genetic analysis, and diagnostic marker development—sets a benchmark for research on understudied horticultural crops and promises to drive the sustainable development of the global dragon fruit industry. Nevertheless, our phenotype-driven screening identified four candidate genes (bHLH, CYP450, POD12, and MATE transporter) that strongly correlate with betalain biosynthesis genes, offering promising targets for future functional validation of flesh color regulation. However, given that SLAF-seq is a reduced-representation sequencing method, the 437 associated SNPs and 51 candidate genes likely represent only a portion of the genetic architecture underlying this trait. Whole-genome resequencing with larger populations will be necessary to capture additional causal variants and provide a complete mechanistic understanding.

5. Conclusion

This study systematically examined the genetic diversity, phylogenetic relationships, and germplasm classification of 45 Chinese dragon fruit (*Selenicereus* spp.) accessions using SLAF-seq technology. A total of 1,206,498 highly consistent SNP markers were identified, providing a high-resolution genomic framework for resolving long-standing ambiguities in dragon fruit germplasm characterization.

Our findings confirm that the 45 accessions can be clearly clustered into three distinct genetic groups corresponding to *S. undatus*, *S. polyrhizus*, and *S. megalanthus*. Key germplasm classification corrections were achieved: (1) Red-skinned, red-fleshed accessions 'HLG' and 'HQL' were assigned to the *S. megalanthus* genetic group, expanding the phenotypic scope of this cultivated group beyond the traditional yellow-skinned, white-fleshed definition; (2) The previously misidentified 'ZL' (annotated as *S. costaricensis*) was confirmed as the *S. polyrhizus* genetic group, clarifying the genetic affiliation of the *S. costaricensis* phenotypic variant with *S. polyrhizus*; (3) The commercially valuable tetraploid 'YWG' was confirmed to have a close genetic background with *S. megalanthus*, providing key guidance for polyploid breeding.

Notably, our findings challenge the reliability of traditional phenotype-based germplasm classification. Traits such as skin color, flesh color, and stem pigmentation were shown to be influenced by genetic mutations or regulatory variations (e.g., betalain biosynthesis) rather than being cultivar specific characteristics. This underscores the necessity of integrating molecular markers with morphological observations for accurate germplasm identification.

Additionally, we developed and validated a tiered set of diagnostic SNP panels. The ultra-core 10-SNP panel not only enables rapid and accurate discrimination among red-fleshed, white-fleshed, and hybrid dragon fruit accessions, but also corrects misclassifications arising from phenotypic plasticity or non-Mendelian expression of flesh color, thereby providing a robust molecular standard for germplasm authentication and breeding selection. This low-cost, high-efficiency tool enables rapid variety authentication, seed purity testing, and hybrid identification. Furthermore, our phenotype-driven analysis revealed four candidate genes (bHLH, CYP450, POD12, and MATE transporter) strongly correlated with betalain biosynthesis, providing valuable targets for future molecular studies on flesh color regulation.

In summary, this study significantly advances our understanding of dragon fruit's genetic architecture and germplasm classification system. The corrected germplasm classifications, high-quality SNP dataset, and diagnostic marker panels provide essential technical support for genetic improvement, parent selection in hybrid breeding, and rational utilization of germplasm resources. These outcomes will facilitate the development of improved cultivation strategies and accelerate the sustainable development of the global dragon fruit industry while establishing a benchmark for molecular characterization of understudied horticultural crops. The core SNP panel established in this study can be widely used in germplasm identification, variety registration, property protection and parental combination prediction in dragon fruit breeding, which will effectively promote the molecular breeding process of this horticultural crop.

Declarations

Ethics approval and consent to participate

Human and animal subjects were not included in this study.

Consent for publication

Publication consent: All authors, affiliated institutions, and individuals involved in this study have provided consent for the publication of this article. The manuscript does not contain any disputed or unauthorized data, patents, or any intellectual property disputes.

Availability of data and materials

The raw SLAF-seq sequence data generated in this study have been deposited into the Genome Sequence Archive (GSA) of the National Genomics Data Center under BioProject accession number PRJCA032254 and are publicly accessible at https://ngdc.cncb.ac.cn/gsa. The hierarchical core SNP panels (including the 10,025-SNP set and six reduced diagnostic panels of 10–500 SNPs) derived from this study are available in Variant Call Format (VCF) via Figshare under the DOI: 10.6084/m9.figshare.30984250. The published RNA-seq data used for transcriptomic correlation analysis are available in NCBI under accession number PRJNA704510. All other data supporting the findings of this study are included within the paper and its Supplementary Information files.

Competing interests

The authors declare no competing interests.

Clinical trial number

Not applicable.

Funding

This research was funded in part by the National Science Foundation of China (Grant No.32102341), Project of Collaborative Innovation Center of GDAAS (XTXM202203), Guangdong Provincial Rural Revitalization Program (2022-NPY-00-034), Science and Technology Program of Guangzhou (Grant No. 202201010527).); "14th Five-Year Plan" Fujian Seed Industry Innovation and Industrialization Project (2021-2025) — Agricultural Fine Variety Breeding and Industrialization Development Project "Introduction, Innovation and Industrialization Development of Characteristic and High-Quality Fruit Varieties" (No. zycxny2021010)

Authors' contributions

Q.M. Sun conceptualized and designed the study, analyzed data, and critically reviewed the manuscript. J.C. Li performed data analysis, conducted bioinformatics analysis, prepared figures, and drafted the manuscript. M.F. Lin contributed to data analysis and revised the manuscript. Y.J. Liu assisted in completing the molecular biology validation work. All authors have read and approved the final version of the manuscript.

Author information

Authors and Affiliations

Institute of Fruit Tree Research, Guangdong Academy of Agricultural Sciences; Key Laboratory of South Subtropical Fruit Biology and Genetic Resource Utilization, Ministry of Agriculture and Rural Affairs; Guangdong Provincial Key Laboratory of Science and Technology Research on Fruit Tree, Guangzhou, 510640, China. Juncheng Li & Qingming Sun

Jiangxi Provincial Key Laboratory of Plantation and High Valued Utilization of Specialty Fruit Tree and Tea Institute of Biological Resources, Jiangxi Academy of Sciences, Nanchang, China. Mengfei Lin

Fruit Science Institute, Fujian Academy of Agricultural Sciences Fuzhou,350013 Fujian China

Youjie Liu

Corresponding author

Correspondence to Qingming Sun.

References

1. Al Bkhetan Z, Wang S (2024) mgikit: demultiplexing toolkit for MGI fastq files. Bioinformatics 40(9):btae554

2. Alexander DH, Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12:1–6
3. Allemailem KS, Almatroudi A et al (2021) Single nucleotide polymorphisms (SNPs) in prostate cancer: its implications in diagnostics and therapeutics. Am J translational Res 13(4):3868
4. Armstrong EE, Li C et al (2025) A Pipeline and Recommendations for Population and Individual Diagnostic SNP Selection in Non-Model Species. Mol Ecol Resour 25(3):e14048
5. Bathke J, Lühken G (2021) OVarFlow: a resource optimized GATK 4 based Open source Variant calling workFlow. BMC Bioinformatics 22:1–18
6. Berry ST, Leon AJ et al (1995) Molecular marker analysis of Helianthus annuus L. 2. Construction of an RFLP linkage map for cultivated sunflower. Theor Appl Genet 91:195–199
7. Browning SR, Browning BL (2011) Population structure can inflate SNP-based heritability estimates. Am J Hum Genet 89(1):191–193
8. Chagné D, Vanderzande S et al (2019) Validation of SNP markers for fruit quality and disease resistance loci in apple (Malus× domestica Borkh.) using the OpenArray® platform. Horticulture research 6
9. Chen JY, Xie FF et al (2021) A chromosome-scale genome sequence of pitaya (Hylocereus undatus) provides novel insights into the genome evolution and regulation of betalain biosynthesis. Horticulture research 8
10. Chen ZL, Meng JM et al (2019) A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. Nat Commun 10(1):3404
11. Chen Z, He Y et al (2022) Investigation of genetic relationships within three Miscanthus species using SNP markers identified with SLAF-seq. BMC Genomics 23:1–11
12. Croft J, Cross N et al (1999) Plant names for the 21st century: the International Plant Names Index, a distributed data source of general accessibility. Taxon 48(2):317–324
13. Da Silva Ribeiro T, Galván JA et al (2022) Maximum SNP FST outperforms full-window statistics for detecting soft sweeps in local adaptation. Genome Biol Evol 14(10):evac143
14. Danecek P, Bonfield JK et al (2021) Twelve years of SAMtools and BCFtools. Gigascience 10(2):giab008
15. Ha TD, Oanh LTK et al (2018) Flowering phenology and mating system of a red skin pitaya (Hylocereus spp.) germplasm collection in Taiwan. Asian J Adv Agricultural Res 7(3):1–8
16. Hasan N, Choudhary S et al (2021) Recent advancements in molecular marker-assisted selection and applications in plant breeding programmes. J Genetic Eng Biotechnol 19(1):128
17. Hernández YDO, Salazar JAC (2012) Pitahaya (Hylocereus spp.): a short review. Comunicata Sci 3(4):220–237
18. Hernández-Langford DG, Mandujano MAC et al (2024) Genetic relationships and patterns of genetic variation in cultivated and crop wild relatives of pitahayas: edible fruits in Selenicereus (Cactaceae). Genetic Resources and Crop Evolution: 1–16
19. Hua Q, Chen C et al (2021) A genome-wide identification study reveals that HmoCYP76AD1, HmoDODA α1 and HmocDOPA5GT involved in betalain biosynthesis in Hylocereus. Genes 12(12):1858
20. Jiang T, Wen C et al (2022) SNP molecular markers development and genetic diversity analysis of Forsythia suspensa based on SLAF-seq technology. Med Plant Res 12
21. Korotkova N, Borsch T et al (2017) A phylogenetic framework for the Hylocereeae (Cactaceae) and implications for the circumscription of the genera. Phytotaxa 327(1):1–46
22. Kumar S, Stecher G et al (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35(6):1547–1549
23. Lalitha S (2000) Primer premier 5. Biotech Software & Internet Report: The Computer Software Journal for Scient 1 (6): 270–272
24. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. bioinformatics 25 (14): 1754–1760

25. Mercado-Silva EM (2018) Pitaya—Hylocereus undatus (haw). Exotic fruits: 339–349
26. Oo KT, Lynn ZM et al (2023) In vitro Propagation of Three Pitaya Varieties (Hylocereus undatus, Hylocereus polyrhizus and Hylocereus megalanthus) with the Use of Different BAP Concentrations. J Sci Innov Res 12:33–39
27. Pan L, Fu J et al (2017) Genetic diversity among germplasms of Pitaya based on SSR markers. Sci Hort 225:171–176
28. Powell W, Morgante M et al (1996) The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Mol Breeding 2:225–238
29. Tamby Chik C, Bachok S et al (2011) Quality characteristics and acceptability of three types of pitaya fruits in a consumer acceptance test. J Tourism Hospitality Culin Arts (JTHCA) 3(1):89–98
30. Tel-Zur N (2022) Breeding an underutilized fruit crop: A long-term program for Hylocereus. Hortic Res 9:uhac078
31. Waskom ML (2021) Seaborn: statistical data visualization. J open source Softw 6(60):3021
32. Xu X, Lu L et al (2015) QTL mapping of cucumber fruit flesh thickness by SLAF-seq. Sci Rep 5(1):15829
33. Zhao D, Xiao S et al (2024) Construction of high-density genetic map based on SLAF-seq and QTL analysis of major traits in sweetpotato [Ipomoea batatas (L.) Lam]. Plant Physiol Biochem 211:108647
34. Zhou Q, Zhou C et al (2017) Genome-wide SNP markers based on SLAF-seq uncover breeding traces in rapeseed (Brassica napus L). Front Plant Sci 8:648

Figures

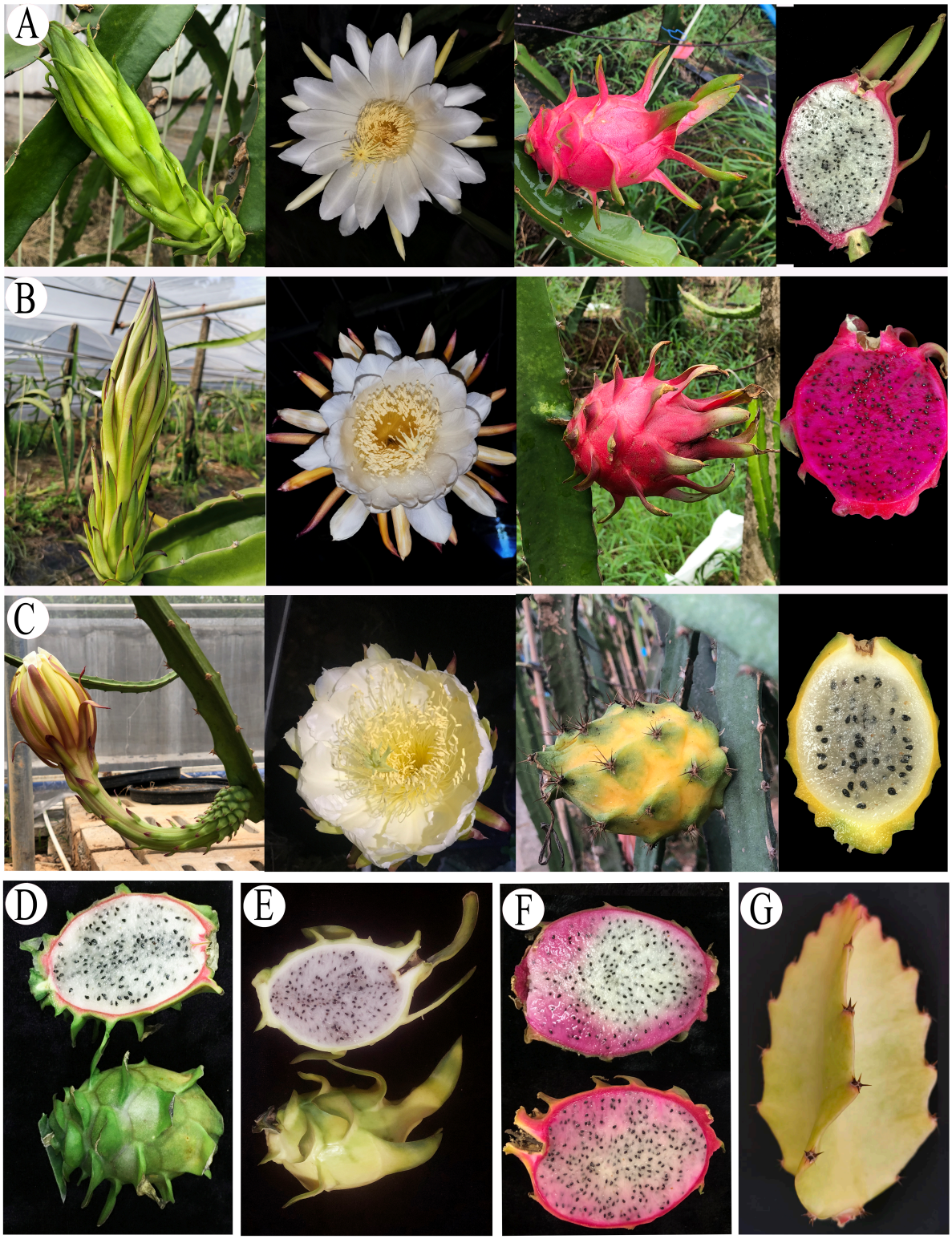


Figure 1

Phenotypic diversity of Selenicereus genotypes and selected accessions.

(A–C) Representative morphological features of three main cultivated *Selenicereus* genotypes: (A) *S. undatus* ("PB"), (B) *S. polyrhizus* ("HSJ"), and (C) *S. megalanthus* ("YWG").

(D–G) Unique phenotypes observed in specific accessions: (D) green fruit peel of "QP", (E) yellow fruit peel of "WCHL", (F) pink/bicolor flesh of P1, (G) yellow stem of "HJ".

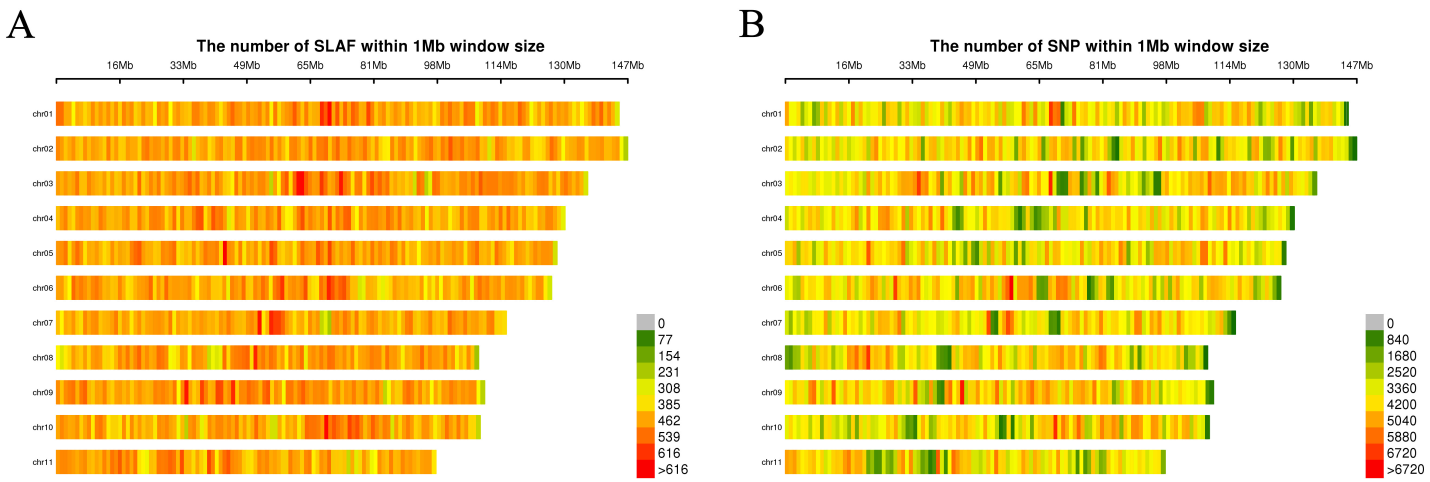


Figure 2

Genomic distribution of SLAF tags and SNP markers across 11 chromosomes.

(A) Density of SLAF tags per 1-Mb window, showing uneven but broad coverage across all chromosomes. (B) Density of SNP markers per 1-Mb window, revealing higher variation and clustering in certain regions. Color scale indicates tag/SNP count; darker red = higher density.

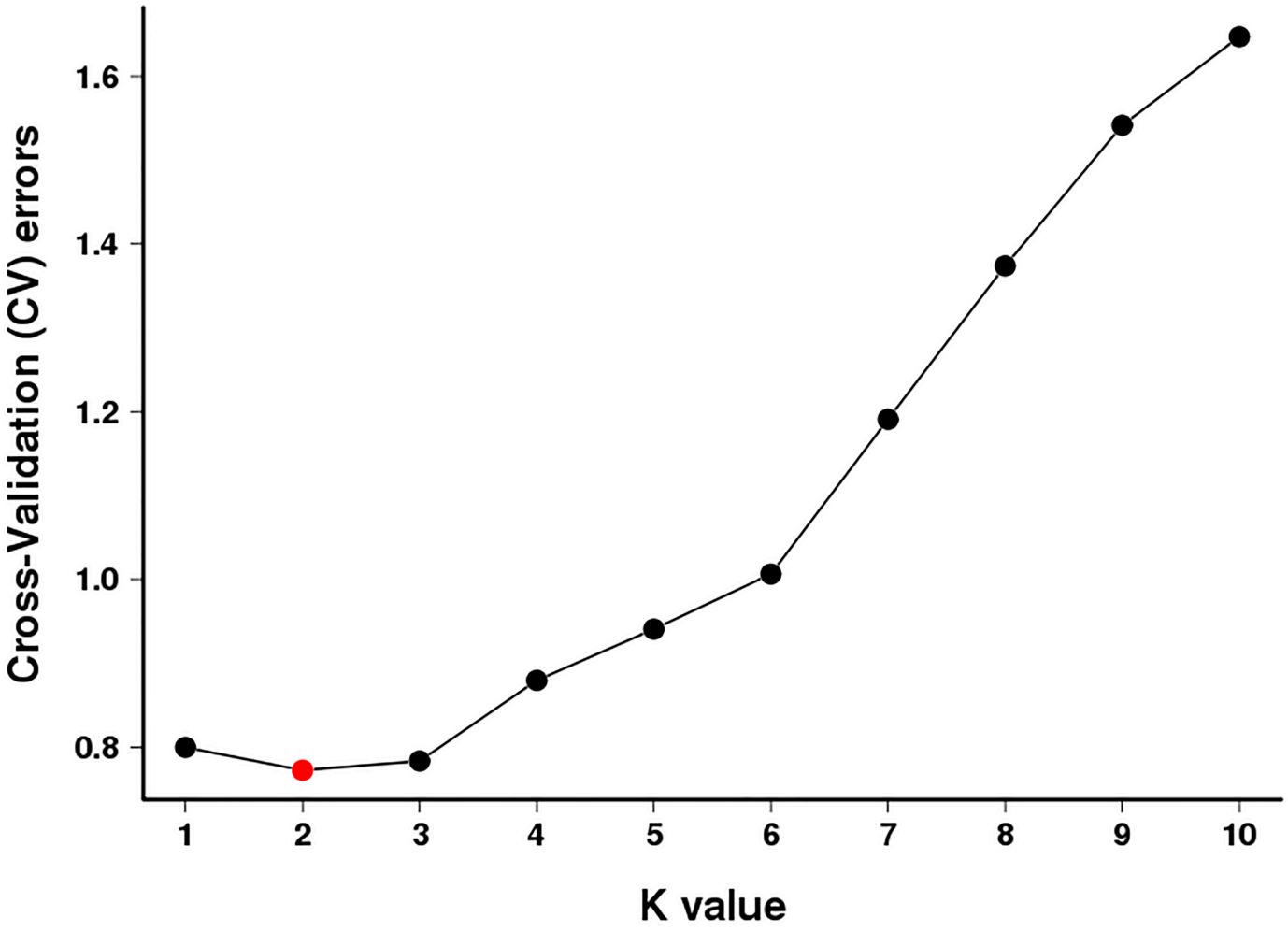


Figure 3

Cross-validation (CV) error rates across K values in admixture analysis.

The lowest CV error occurs at K = 2, indicating optimal clustering of the population into two genetic groups. Error rates increase with higher K values, suggesting overfitting.

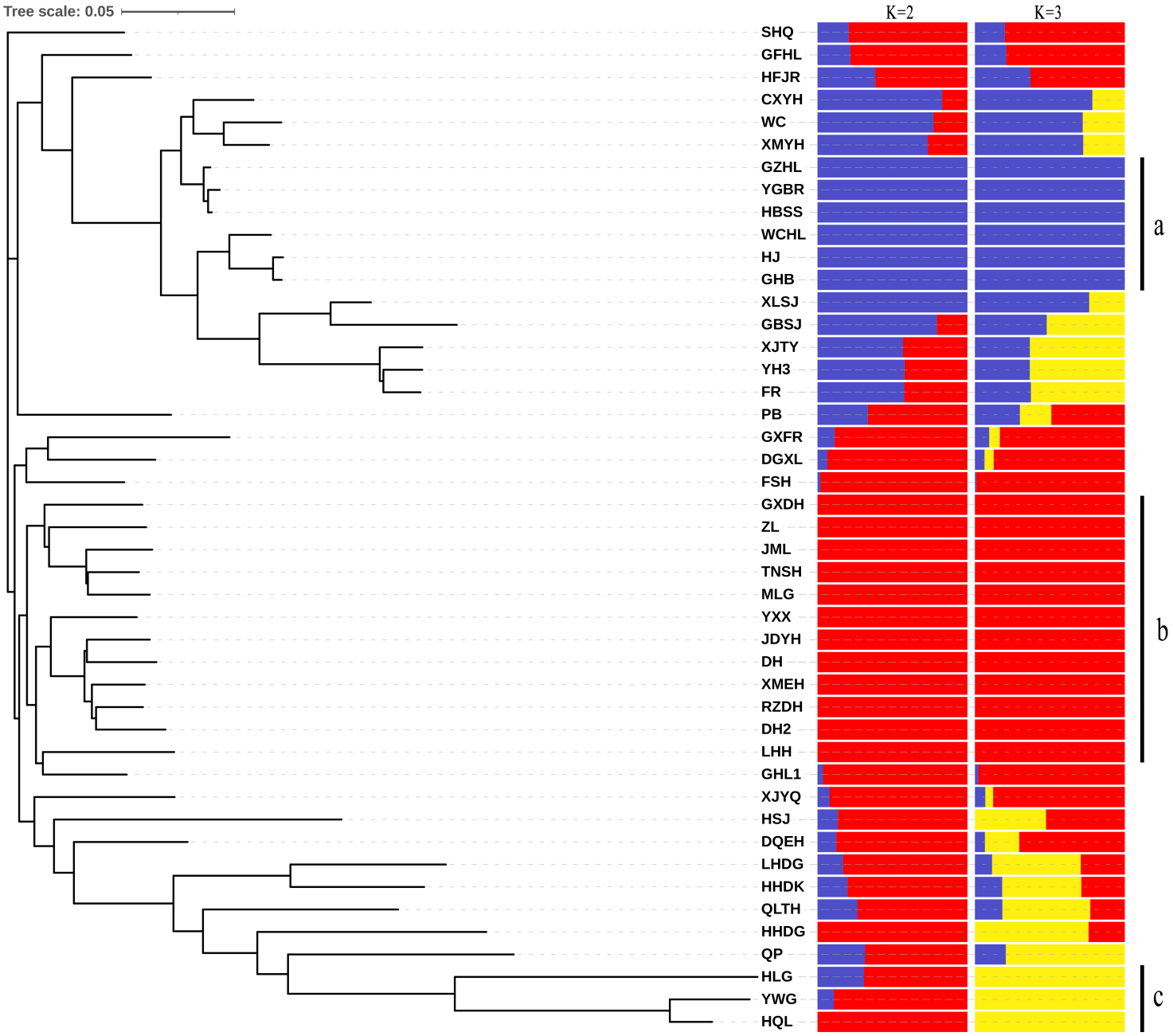


Figure 4

Phylogenetic tree and admixture clustering (K=2 and K=3) of 45 dragon fruit accessions.

Clades a, b, and c represent *S. undatus*, *S. polyrhizus*, and *S. megalanthus* cultivated genetic groups, respectively. Color bars show ancestry proportions; samples are ordered by phylogenetic position.

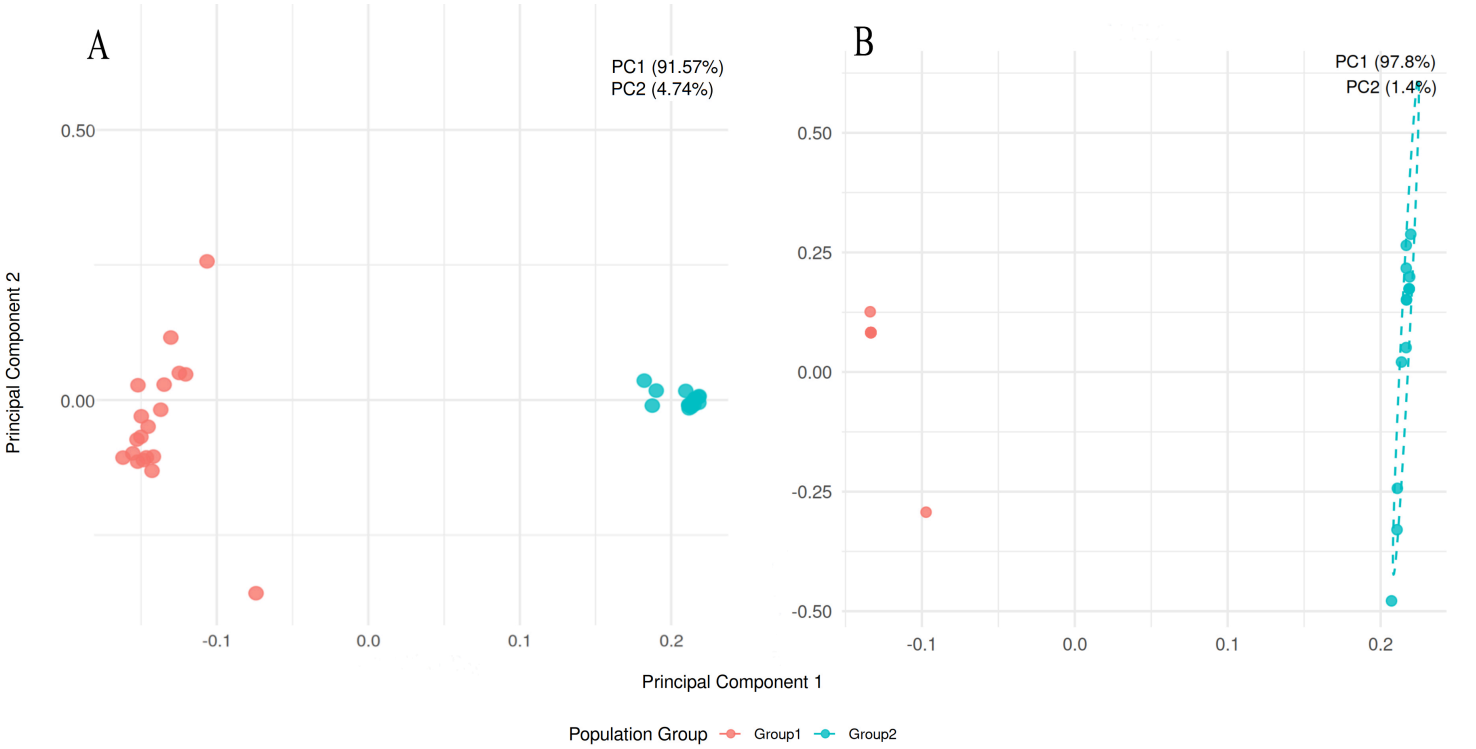


Figure 5

Principal component analysis (PCA) of 34 dragon fruit accessions using key SNPs (A) and ultra_core_10 SNPs (B).

Samples are colored by flesh color phenotype: red = red-flesh, cyan = white-flesh. (A) Shows clear separation along PC1 (91.57% variance), with minor overlap. (B) Reveals near-perfect clustering along PC1 (97.8% variance), indicating strong genetic differentiation driven by the ultra-core SNPs.

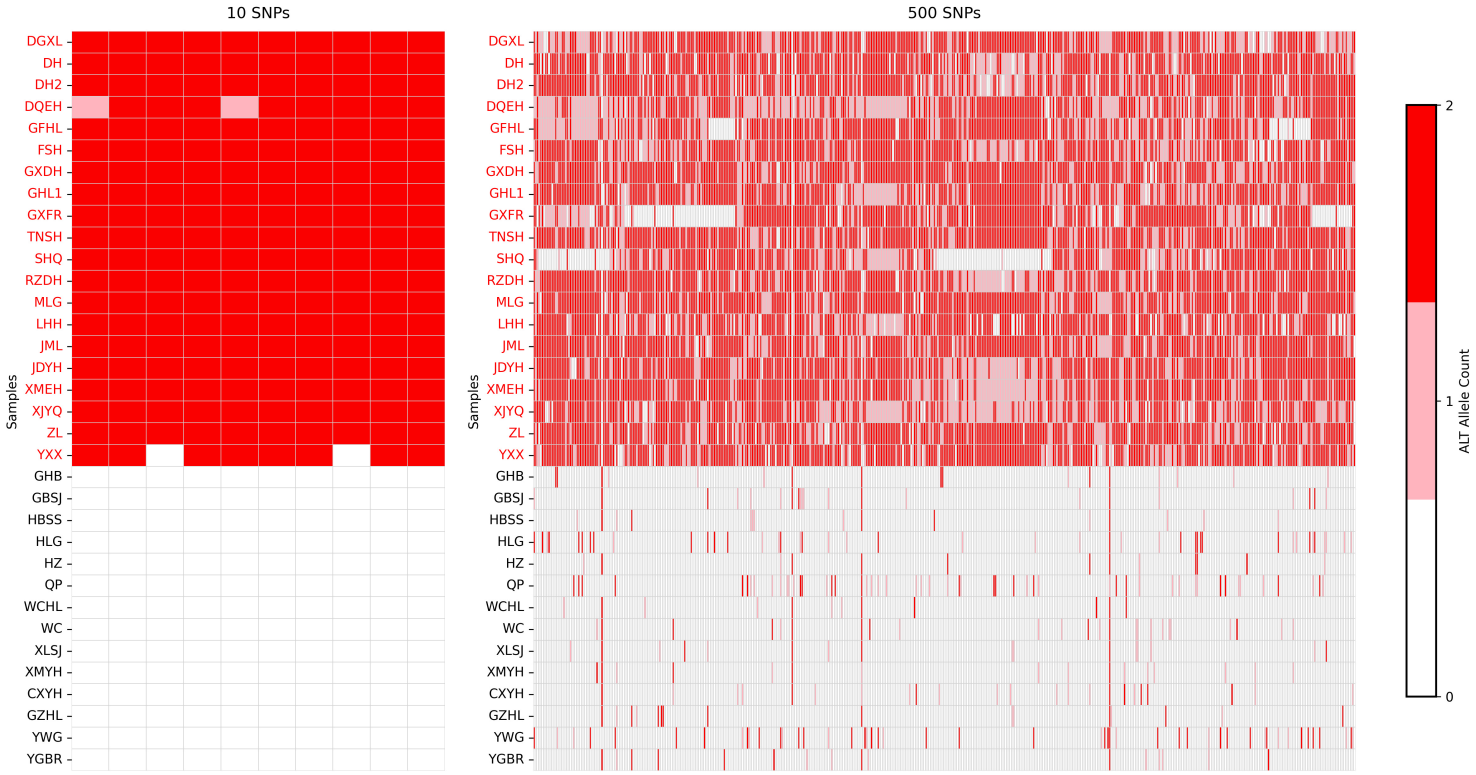


Figure 6

Haplotype patterns of ultra-core and extended SNP sets reveal complete co-segregation with flesh color phenotype in dragon fruit.

(A) Genotype heatmap of 10 ultra-core SNPs showing strong allele frequency divergence between red-flesh (top, n=20) and white-flesh (bottom, n=14) samples. (B) Genotype heatmap of 500 extended SNPs confirming consistent genetic differentiation across a broader genomic region. Each cell represents ALT allele count (0: white, 1: pink, 2: red). Samples are ordered by phenotype.

Supplementary Files

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

- TableS3.docx
- TableS1.docx
- TableS5.docx
- TableS4.docx
- TableS2.docx
- TableS6.docx
- FigureS3.png
- FigureS4.png
- FigureS1.jpg
- FigureS2.jpg
- Supplementarycaptions.docx