ISSN: 2458-8989

Natural and Engineering Sciences

NESciences, 2026, 11 (3): 913-929 doi: 10.28978/nesciences.263071

Identification of Genotype-Specific SNP Markers for Molecular Discrimination of Iraqi Roselle (*Hibiscus sabdariffa* L.) Genotypes

Ola Hatif Hazim[1]*, Maysam Ihsan Ali Al-Hassany[2], Nawal Jasem Shanyoor[3]

[1] Mustansiriyah University, College of Science, Baghdad, Iraq. Email: ola.hataf@uomustansiriyah.edu.iq, ORCID: 0000-0002-4435-2417

[2] Mustansiriyah University, College of Science, Baghdad, Iraq. Email: maysam.ihssan@uomustansiriyah.edu.iq, ORCID: 0009-0008-1711-7682

[3] Mustansiriyah University, College of Science, Baghdad, Iraq. Email: naw.j.sh@uomustansiriyah.edu.iq ORCID: 0009-0005-1363-4007

Abstract

Hibiscus sabdariffa L. (roselle) is an economically important multipurpose crop, yet the genetic diversity and population structure of Iraqi germplasm remain poorly characterized, limiting effective germplasm management and molecular breeding. This study aimed to characterize genome-wide genetic variation in Iraqi roselle and develop a reliable SNP-based panel for genotype discrimination. Forty-two accessions representing four major agroecological regions of Iraq were genotyped using double-digest restriction-site-associated DNA sequencing (ddRAD-seq) on the Illumina NovaSeq 6000 platform. Bioinformatic filtering against the H. sabdariffa reference genome identified 9,847 high-confidence biallelic SNPs. The resulting diversity estimates indicated substantial polymorphism, with expected heterozygosity (H_e) of 0.312 ± 0.041, polymorphism information content (PIC) of 0.271 ± 0.038, and Shannon's information index (I) of 0.489 ± 0.057. AMOVA attributed 34.7% of genetic variation to differences among geographic groups, while STRUCTURE analysis identified three genetically distinct populations, supported by principal coordinate analysis and UPGMA clustering. Fifteen SNPs were subsequently validated as KASP markers, comprising eight population-diagnostic, four group-discriminating, and three accession-specific markers. The combined panel achieved a probability of identity of 1.3×10^{-9}, with 97.6% leave-one-out and 95.8% blind-validation accuracy. These findings provide the first practical molecular framework for Iraqi roselle diversity assessment, genotype authentication, germplasm conservation, and marker-assisted breeding.

Keywords: *Hibiscus sabdariffa*; SNP discovery; ddRAD-seq; KASP assay; genetic diversity; population structure; molecular fingerprinting; Iraqi germplasm.

Available Online: 10/09/2026

1. Introduction

Hibiscus sabdariffa L. (roselle; Malvaceae; 2n = 2x = 36) is a morphologically diverse multipurpose crop cultivated in tropical and subtropical regions of Africa, Asia, the Americas as well as the Middle East (1). Its fleshy calyces are widely used for making beverages, jams, natural colorants, and nutraceutical products,

*Corresponding Author: Ola Hatif Hazim, E-mail: ola.hataf@uomustansiriyah.edu.iq,

while its bast fiber is used as a substitute for jute (2). The clinical and commercial interest in roselle keeps growing given the fact that its abundance of anthocyanins, flavonoids, hibiscus acid, and other polyphenolic compounds has attracted scientists' attention (3) (4). What is important is that phytochemical composition and its biological activities differ by genotype, and thus genetic characterization becomes vital for quality-oriented breeding, germplasm management, and cultivar authentication (5). potential.

Iraq is home to a potentially rich yet under-explored source of genetic diversity in the species of roselle. The diversity of agroecological parameters in the country between the hot, salinity-affected southern areas and cooler regions in the north has likely enabled adaptation to local conditions while retaining alleles that confer tolerance to environmental stressors. However, research on the genetic diversity of H. sabdariffa germplasm from Iraq is very limited, restricting the knowledge about relationships, priorities regarding conservation, and breeding options. Morphological and agro-physiological techniques are used to collect information on the morphological characteristics of organisms but are strongly influenced by environmental conditions and development plasticity (2) (6). Previous molecular techniques, such as RAPD, AFLP, and SSR markers, demonstrated significant genetic diversity in roselle, but also had limitations due to low genome coverage, repeatability, scalability, or density of loci, which are not sufficient for whole-population genetic studies and high-throughput genotype evaluations. (7) (8) (9) (10).

Due to their great potential, single-nucleotide polymorphisms (SNPs) are considered a better solution because of their abundance, codominant inheritance, reproducibility, and possibility of automated genotyping (11). In species where whole-genome sequencing is relatively expensive, reduced-representation sequencing methods such as double-digest restriction-site-associated DNA sequencing (ddRAD-seq) are a good option for discovery of SNPs on a genome scale (12). Huang et al. (2024) have recently shown the effectiveness of such methods for roselle when they obtained genome-wide SNPs from Taiwanese varieties and transformed the selected SNPs into KASP assays for accessions discrimination (13). The KASP platform thus provides an option to convert informative SNPs into the available assays for identification of the genotype of studied plants and their breeding (14).

Though progress has been made in this regard, the genetic resources of multiple SNPs across genomes and the relevant molecular markers pertaining to the diagnosis of Iraqi roselle have yet to be put in place in a systematic way. With this in mind, this study has been initiated with the goal of describing genetic diversity and population structure in 42 geographically separated and diverse Iraqi accessions of H. sabdariffa using ddRAD-seq, finding population and genotype-specific SNPs, and proceeding to develop a KASP-based marker panel. This will help improve germplasm conservation, variety identification, and breeding of Iraqi roselle using genomic techniques.

2. Materials and Methods

2.1 Plant Material, Germplasm Collection, and Morphological Characterization

A genetically representative panel of **42 *Hibiscus sabdariffa* L. accessions** (2n = 2x = 36) was assembled to capture the full spectrum of morphological and geographic diversity within the Iraqi roselle gene pool. Accessions were collected during systematic field surveys conducted throughout the 2022–2023 growing seasons across four principal roselle-cultivation zones in Iraq, as summarized in Table 1. Geographic coordinates (WGS84 datum) and altitude for each collection site were recorded with a Garmin GPSMAP 66i receiver (±3 m horizontal accuracy).

Table 1. Geographic distribution and classification of the 42 *Hibiscus sabdariffa* L. accessions used in this study.

Geographic Zone	Governorates Included	No. of Accessions (n)	Altitude Range (m a.s.l.)	Accession Type
Northern Highland	Nineveh, Erbil, Sulaymaniyah	12	300–850	Farmer landraces
Central Alluvial Plain	Baghdad, Diyala, Wasit	8	20–60	Farmer landraces + breeding lines
Southern Deltaic Zone	Basra, Dhi Qar, Maysan	15	0–15	Farmer landraces

National Genebank (ex situ)	Abu Ghraib (INGB)	7	34	Heritage landraces + elite breeding lines
Total		42	0–850	

For each accession, twelve quantitative and qualitative morphological descriptors were scored during the vegetative and reproductive growth stages following the standardized *IPGRI Descriptors for Hibiscus sabdariffa* (IPGRI, 1998) and FAO/IPGRI multicrop passport data (MCPD) protocols (Table 2). All phenotypic assessments were conducted under uniform field conditions at the University of Baghdad Agricultural Experimental Station (33°18′N, 44°24′E; elevation 34 m a.s.l.) during a single cropping season to eliminate confounding genotype × environment interaction effects.

Table 2. Morphological descriptors recorded for each accession following IPGRI (1998) protocols.

Descriptor	Type	Unit / Scale	Growth Stage
Plant height at maturity	Quantitative	cm	Reproductive
Stem diameter (10 cm above soil)	Quantitative	mm	Reproductive
Leaf blade length	Quantitative	cm	Vegetative
Leaf blade width	Quantitative	cm	Vegetative
Calyx length at full maturity	Quantitative	mm	Reproductive
Calyx width at full maturity	Quantitative	mm	Reproductive
Calyx fresh weight per plant	Quantitative	g	Reproductive
Calyx dry weight per plant	Quantitative	g	Reproductive
Calyx color	Qualitative	CIELAB notation	Reproductive
Days to first flower anthesis	Quantitative	days	Reproductive
Days to 50% fruiting	Quantitative	days	Reproductive
100-seed weight	Quantitative	g	Post-harvest

Voucher herbarium specimens for all accessions were prepared and permanently deposited in the National Herbarium of Iraq (NHI), Baghdad (accession numbers NHI-HS-001 to NHI-HS-042). All germplasm collections were conducted under permits issued by the Iraqi Ministry of Agriculture in full compliance with the Convention on Biological Diversity (CBD; Rio de Janeiro, 1992) and the Nagoya Protocol on Access and Benefit Sharing (ABS; 2010). Prior to field transplanting, seeds were surface-sterilized in 0.5% (v/v) sodium hypochlorite solution for 5 min, rinsed three times with sterile deionized water, and pre-germinated on moistened sterile filter paper at 28 ± 1°C in complete darkness for 48–72 h. Seedlings were transplanted to the experimental field at the three-true-leaf stage and grown under standard agronomic management practices.

2.2 Genomic DNA Extraction, Purification, and Quality Assessment

Young, fully expanded leaf laminae from the third to fifth node position were harvested from 30-day-old greenhouse-grown seedlings (25 ± 2°C, 16 h photoperiod, 65 ± 5% relative humidity) in the pre-flowering vegetative phase to minimize secondary metabolite accumulation. Leaf tissue samples (100 ± 5 mg fresh weight per accession) were excised with a sterile scalpel, immediately snap-frozen in liquid nitrogen, and stored at −80°C in RNAlater™-stabilized tubes (Thermo Fisher Scientific, Waltham, MA, USA) until extraction.

Total genomic DNA was isolated using a modified cetyltrimethylammonium bromide (CTAB) protocol specifically optimized for the polysaccharide- and anthocyanin-rich tissue of *H. sabdariffa* (15). Frozen leaf tissue was ground to a fine homogeneous powder under liquid nitrogen, transferred into pre-warmed 2.0 mL microcentrifuge tubes, and resuspended in 700 µL of 2× CTAB extraction buffer whose composition is detailed in Table 3. The supplementation with 5 M NaCl is a critical modification that facilitates selective

precipitation of polysaccharides during the CTAB–DNA co-precipitation step, which would otherwise co-purify with genomic DNA and inhibit downstream enzymatic reactions (16).

Table 3. Composition of the 2× CTAB extraction buffer used for *Hibiscus sabdariffa* genomic DNA isolation.

Component	Final Concentration	Role
CTAB	2% (w/v)	Cell lysis and nucleic acid stabilization
NaCl	1.4 M	DNA precipitation and polysaccharide removal
NaCl (supplemental)	5 M (additional)	Enhanced polysaccharide flocculation [2]
EDTA (pH 8.0)	20 mM	DNase inhibition (chelates Mg^{2+})
Tris-HCl (pH 8.0)	100 mM	pH buffer
β-Mercaptoethanol	2% (v/v)	Protein denaturation and oxidation prevention
Nuclease-free water	To final volume	—

Samples were incubated at 65°C in a thermoshaker (Eppendorf ThermoMixer® C) at 900 rpm for 60 min with vortexing every 15 min to ensure complete cell lysis. Following incubation, equal volumes of chloroform:isoamyl alcohol (24:1, v/v) were added and the mixture was emulsified by gentle inversion (50 cycles) before centrifugation at 12,000 × *g* for 15 min at 4°C. The aqueous phase was re-extracted once with chloroform:isoamyl alcohol to minimize residual protein and chlorophyll carryover. Genomic DNA was precipitated by the addition of 0.7 volumes of ice-cold isopropanol and 0.1 volumes of 3 M sodium acetate (pH 5.2), incubated at −20°C for 30 min, and pelleted by centrifugation at 13,000 × *g* for 20 min at 4°C. The pellet was washed twice with 70% (v/v) ethanol (1 mL per wash; 5,000 × *g*, 5 min each), air-dried for 20–30 min, and resuspended in 100 µL of nuclease-free TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0). Residual RNA was removed by treatment with RNase A (10 mg mL^{-1}; Sigma-Aldrich) at 37°C for 30 min, followed by a final chloroform:isoamyl alcohol cleanup.

DNA purity and integrity were assessed using the quality acceptance criteria listed in Table 4. Only samples satisfying all three criteria simultaneously were advanced to library preparation.

Table 4. DNA quality acceptance criteria applied prior to ddRAD-seq library construction.

Quality Parameter	Measurement Method	Acceptance Threshold	Rejection Criterion
A_{260}/A_{280} purity ratio	NanoDrop™ 2000c	1.80 – 2.00	< 1.80 or > 2.00 (protein/RNA contamination)
A_{260}/A_{230} purity ratio	NanoDrop™ 2000c	> 1.80	≤ 1.80 (polysaccharide/phenol contamination)
dsDNA concentration	Qubit™ 3.0 Fluorometer (BR Assay)	≥ 100 ng μL^{-1}	< 100 ng μL^{-1}
DNA integrity	1.0% agarose gel electrophoresis	Intact high-MW band (> 20 kb)	Smearing or fragmentation

2.3 ddRAD-seq Library Preparation, Size Selection, and High-Throughput Sequencing

Reduced-representation genomic libraries were constructed using the double-digest restriction-site associated DNA sequencing (ddRAD-seq) strategy of Peterson et al. (2012) (12), as implemented for diploid plant genomes with high repetitive-element content (13). The enzyme combination and expected fragment size range were pre-determined by *in silico* digestion of the assembled *H. sabdariffa* reference genome (NCBI BioProject: PRJNA497891) using SimRAD v.0.96. For each accession, 200 ± 5 ng of high-

integrity genomic DNA was co-digested with EcoRI-HF (rare cutter) and MseI (frequent cutter) (New England Biolabs, Ipswich, MA, USA) at 37°C for 16 h, followed by thermal inactivation at 65°C for 20 min.

Uniquely barcoded P1 adapters (harboring distinct 6 bp inline barcodes fused to the EcoRI overhang) and a universal common P2 adapter (complementary to the MseI overhang) were synthesized and HPLC-purified by IDT (Integrated DNA Technologies, Coralville, IA, USA). Adapter ligation was carried out at 16°C for 16 h using T4 DNA Ligase (New England Biolabs), and ligation efficiency was confirmed by qPCR before pooling. Libraries were pooled in equimolar groups of 12 samples, size-selected to the 300–500 bp target range using a BluePippin™ automated gel system (Sage Science, Beverly, MA, USA; 1.5% agarose cassette, 3% precision setting), and PCR-amplified using Phusion® High-Fidelity DNA Polymerase (10 enrichment cycles). The key library preparation parameters are summarized in Table 5.

Table 5. Summary of ddRAD-seq library preparation parameters.

Parameter	Value / Condition
Input DNA per accession	200 ± 5 ng (integrity > 20 kb)
Restriction enzymes	EcoRI-HF (rare cutter) + MseI (frequent cutter)
Digestion temperature / duration	37°C / 16 h
Enzyme inactivation	65°C / 20 min
Inline barcode length	6 bp (unique per sample)
Ligation temperature / duration	16°C / 16 h
Multiplexing per pool	12 samples
Size selection range (insert)	300–500 bp (BluePippin™, 1.5% agarose)
PCR enrichment cycles	10 cycles (98°C 10 s / 65°C 30 s / 72°C 30 s)
DNA polymerase	Phusion® High-Fidelity (NEB)
Library cleanup	AMPure XP SPRI beads (1.0× ratio)
Library QC	Agilent 2100 Bioanalyzer + Qubit™ 3.0 + KAPA qPCR
Sequencing platform	Illumina NovaSeq 6000 (150 bp paired-end)
Target sequencing depth	≥ 20× per individual at RAD loci

2.4 Bioinformatics Pipeline: Read Processing, Alignment, and SNP Discovery

All bioinformatic analyses were executed on a high-performance computing (HPC) cluster (Linux CentOS 7.9; ≥ 64 CPU cores; 512 GB RAM). Raw paired-end reads were demultiplexed and quality-filtered using the **process_radtags** module of Stacks v.2.62 (17) (minimum Phred Q20, up to one barcode mismatch rescued), followed by adapter trimming with **Trimmomatic v.0.39** (18) (SLIDINGWINDOW:4:20, MINLEN:100). Post-trimming quality was assessed using FastQC v.0.11.9 and aggregated with MultiQC v.1.12.

Quality-filtered reads were aligned to the *H. sabdariffa* reference genome (19) using **BWA-MEM v.0.7.17** (20). Alignment files were coordinate-sorted, and PCR duplicates were flagged using **Picard MarkDuplicates v.2.27.4** (optical duplicate pixel distance: 2,500). Samples with overall mapping rate < 80% or mean on-target coverage < 15× were flagged for re-sequencing. SNP discovery and joint genotyping followed the GATK Best Practices pipeline (**GATK HaplotypeCaller v.4.3** (21), GVCF mode; **GenomicsDBImport** + **GenotypeGVCFs** for cohort genotyping). Hard-filter thresholds for raw SNPs were applied as: QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < −12.5, and ReadPosRankSum < −8.0.

Population-level VCF filtering was performed in **VCFtools v.0.1.17** (22) applying all criteria listed in Table 6 sequentially. Linkage disequilibrium (LD) pruning (sliding window: 50 SNPs, step: 5 SNPs, r^2 threshold: 0.80) was additionally applied to produce an independent-loci subset for STRUCTURE analysis. The final stringently filtered dataset comprised **9,847 high-confidence biallelic SNPs** distributed across all 12 *H. sabdariffa* chromosome-level scaffolds.

Table 6. Population-level VCF filtering criteria applied in VCFtools v.0.1.17 to obtain the final high-confidence SNP dataset.

Filter Criterion	VCFtools Flag	Threshold	Rationale
Biallelic SNPs only	--min-alleles 2 --max-alleles 2	Exactly 2 alleles	Removes multi-allelic and indel sites
Minimum mean read depth	--min-meanDP	≥ 10×	Ensures reliable genotype calls
Maximum per-site missing rate	--max-missing	≤ 10% (i.e., ≥ 0.90)	Excludes poorly genotyped loci
Maximum per-individual missing rate	--mind	≤ 15%	Removes low-quality samples
Minor allele frequency	--maf	≥ 0.05	Removes near-monomorphic, uninformative loci
Hardy–Weinberg equilibrium	--hwe	p-value > 0.001	Reduces contribution of paralogous loci
Final SNP count after all filters		9,847 biallelic SNPs	

2.5 KASP Marker Design, Optimization, and Cross-Laboratory Validation

Candidate SNP prioritization. A subset of **15 candidate SNPs** was selected from the 9,847 filtered SNPs for conversion into Kompetitive Allele-Specific PCR (KASP) assays [12] based on the following multi-criterion prioritization: (i) PIC > 0.60; (ii) MAF ≥ 0.25; (iii) pairwise LD $r^2 < 0.20$ with all other selected SNPs; (iv) even genomic distribution (minimum spacing ≥ 5 Mb between adjacent selected SNPs); and (v) sequence context suitability (absence of secondary structures, repetitive elements, or additional SNPs within ±50 bp of the target site, assessed using RepeatMasker v.4.1.2). Flanking sequences (200 bp per SNP) were submitted to the **LGC Biosearch Technologies KASP-by-Design™** platform for automated allele-specific tri-primer design (23).

KASP genotyping reactions. Genotyping was performed in a total reaction volume of 10 µL. The complete reaction mixture composition and thermal cycling program are detailed in Tables 7 and 8, respectively. Reactions were set up in 384-well plates using a Hamilton NIMBUS® liquid-handling robot to minimize pipetting error. No-template controls (NTCs) and two reference DNA samples of known genotype were included on every plate.

Table 7. KASP genotyping reaction mixture composition (10 µL final volume).

Component	Volume (µL)	Final Concentration
2× KASP V4.0 Master Mix (LGC Genomics)	5.00	1×
KASP-by-Design™ assay mix	0.14	Allele-specific forward primers: 46 µM each; Common reverse primer: 100 µM
Template genomic DNA	Variable	20 ng per reaction
Nuclease-free water	To 10.00 µL	—

Table 8. KASP thermal cycling program optimized for *Hibiscus sabdariffa* genotyping on the CFX384 Touch™ Real-Time PCR System (Bio-Rad).

Step	Temperature	Duration	Cycles	Notes
Initial enzyme activation	94°C	15 min	1	Hot-start activation
Touchdown denaturation	94°C	20 s	10	—
Touchdown annealing	61°C → 55°C	60 s	10	Decreasing 0.6°C per cycle
Denaturation	94°C	20 s	26	Fixed annealing phase

Annealing / Extension	55°C	60 s	26	—
Signal stabilization	37°C	1 min	1	Plate read (FAM + HEX channels)

End-point fluorescence data were analyzed and allelic clusters (FAM homozygote, HEX homozygote, heterozygote) were assigned using **KlusterCaller™ v.1.1** (LGC Genomics) with manual inspection of borderline data points. Assay reproducibility was validated by independent triplicate genotyping of all 42 accessions at three laboratories. Only markers achieving a call rate ≥ 95% and between-replicate concordance ≥ 99% across all three laboratories were retained in the final validated marker panel.

2.6 Genetic Diversity and Population Differentiation Parameters

All population genetic analyses were conducted on the validated 15-locus KASP-SNP genotype matrix using **GenAlEx v.6.503** [13] and **PowerMarker v.3.25** (14). Marker-level statistics computed per locus and per population included: observed heterozygosity (*Ho*), expected heterozygosity (*He*), minor allele frequency (MAF), within-population inbreeding coefficient (*F*IS = (*He* − *Ho*) / *He*), Polymorphic Information Content (PIC; Botstein et al., 1980), and Shannon's information index (*I*):

$$\mathrm{PIC} = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n} \sum_{j=i+1}^{n} 2\, p_i^2 p_j^2$$

$$I = -\sum_{i=1}^{n} p_i \ln(p_i)$$

Nei's gene diversity statistics (H*T*, H*S*, and G*ST*) and Wright's fixation index (*F*ST; Weir and Cockerham, 1984) were computed in GenAlEx. Analysis of Molecular Variance (AMOVA) was performed with **10,000 random permutation replicates** to partition total molecular genetic variance into: (i) among geographic groups (Φ*CT*); (ii) among individuals within groups (Φ*SC*); and (iii) within individuals (Φ*ST*). Gene flow was estimated as *N*m = (1 − *F*ST) / (4 × *F*ST). Pairwise Nei's *D*A genetic distances were used to construct Neighbor-Joining (NJ) and UPGMA dendrograms in **MEGA X v.10.2** (24) (1,000 bootstrap replicates), and dendrogram fidelity was evaluated by cophenetic correlation coefficient ($r \geq 0.85$ required) in **NTSYS-pc v.2.1**. The **Mantel test** (10,000 permutations; GenAlEx) was used to evaluate isolation-by-distance (IBD) as the correlation between geographic distance (km, Haversine formula) and genetic distance (*F*ST / (1 − *F*ST)).

2.7 Hierarchical Cluster Analysis and Principal Coordinate Analysis

A pairwise Nei's *D*A distance matrix across all 42 accessions was computed in **GenAlEx v.6.503**. UPGMA hierarchical clustering was applied and visualized in **MEGA X v.10.2** (24) with 1,000 bootstrap replicates (minimum display threshold: 50%). Cophenetic correlation between the distance matrix and UPGMA topology was computed in NTSYS-pc v.2.1.

Principal Coordinate Analysis (PCoA) was performed in **GenAlEx v.6.503** on the same *D*A matrix. The percentages of total genetic variance explained by each principal coordinate axis were computed, and the first three axes were retained. Two-dimensional scatter plots (PCo1 × PCo2; PCo1 × PCo3) were rendered in **R v.4.3.1** (*ggplot2* v.3.4.2) with 95% confidence ellipses (*stat_ellipse*, multivariate normal assumption) and non-overlapping labels (*ggrepel* v.0.9.3).

2.8 Bayesian Model-Based Population Structure Analysis

Bayesian probabilistic clustering was performed in **STRUCTURE v.2.3.4** (25). Prior to the run, the 9,847 SNP dataset was LD-pruned ($r^2 < 0.10$, sliding window 50 SNPs, step 5 SNPs) to satisfy the independence assumption of the STRUCTURE model. The **admixture model with correlated allele frequencies** and the **LOCPRIOR model** (Hubisz et al., 2009) were jointly employed to improve clustering resolution at low *K* values. For each value of $K = 1$ to $K = 8$, **20 independent MCMC replicate runs** were conducted, each with a burn-in of **100,000 iterations** followed by **200,000 post-burn-in sampling iterations**. MCMC stationarity was verified by convergence of ln P(D), the admixture parameter (α), and absence of systematic parameter trends across run logs. The optimal *K* was determined using the ln P(D) plateau method and the ΔK statistic of Evanno et al. (2005) (26), computed in **Structure Harvester v.0.6.94**. Replicate runs at optimal *K* were aligned using **CLUMPP v.1.1.2** (Greedy algorithm; 10,000 permutations) and visualized using **DISTRUCT v.1.1** (27). Accessions with posterior membership coefficient Q ≥ 0.70 were assigned to

a specific population; those with Q < 0.70 for all inferred populations were classified as genetically admixed.

2.9 Identification and Validation of Genotype-Specific Diagnostic SNP Markers

Marker discovery. All 9,847 filtered biallelic SNPs were systematically screened for fixed or near-fixed allelic contrasts between groups using custom R v.4.3.1 scripts. Three diagnostic marker categories were defined:

- **Category I (population-diagnostic):** diagnostic allele frequency ≥ 0.90 in target population AND ≤ 0.10 in all non-target populations simultaneously.
- **Category II (STRUCTURE-group discriminating):** allele frequency contrast consistent with Bayesian population assignments at optimal *K*, regardless of geographic grouping.
- **Category III (genotype-specific fingerprinting):** allelic state unique (homozygous for rare allele) to a single accession across the entire 42-genotype panel.

Probability of identity. Multi-locus discriminatory power was quantified by the Probability of Identity (PI; Waits et al., 2001 [21]):

$$\mathrm{PI} = \prod_{i=1}^{L} \left[2p_i^2\left(1 - p_i^2\right) - p_i^4\right]$$

where p_i is the diagnostic allele frequency at locus i and L is the number of loci. Loci were added iteratively until PI < 10^{-6}, defining the minimum sufficient marker combination for unambiguous genotype discrimination.

Trait association. Associations between candidate trait-linked markers and phenotypic traits were tested using: (i) chi-square (χ^2) goodness-of-fit test for calyx color (categorical: red vs. green vs. mixed); and (ii) one-way ANOVA followed by Tukey's HSD post-hoc test for days to first flowering (quantitative). All tests used a nominal significance level of $p < 0.05$, with Bonferroni correction applied for multiple comparisons (αcorrected = 0.05 / number of tests).

Cross-validation. The diagnostic panel was validated by: (i) **leave-one-out cross-validation (LOO-CV)** within the 42-accession discovery panel; and (ii) independent **blind validation** using 24 accessions (15 Iraqi landraces + 9 international accessions not included in discovery) genotyped under identical KASP conditions. Classification accuracy was reported as the percentage of accessions correctly assigned to their known population group or individual identity, with 95% confidence intervals derived from 10,000 bootstrap resampling replicates in R.

2.10 Software, Statistical Analysis, and Data Availability

All statistical analyses were performed using the software environments and tools listed in Table 9. Custom R scripts for PCoA visualization, diagnostic marker screening, and PI calculation are provided in Supplementary File S2. Raw Illumina sequencing data, processed VCF files, KASP genotype matrices. Significance thresholds were set at $p < 0.05$ throughout; multiple-comparison corrections applied Bonferroni adjustment unless otherwise stated.

Table 9. Software, tools, and platforms used in this study with corresponding versions and references.

Analysis Step	Software / Tool	Version	Reference / Source
In silico RAD-seq simulation	SimRAD	v.0.96	Lepais & Weir (2014)
Read demultiplexing & quality filtering	Stacks (process_radtags)	v.2.62	[5]
Adapter trimming	Trimmomatic	v.0.39	[6]
QC visualization	FastQC / MultiQC	v.0.11.9 / v.1.12	Babraham Bioinformatics
Reference genome alignment	BWA-MEM	v.0.7.17	[8]
Alignment processing	SAMtools	v.1.16	[9]
PCR duplicate removal	Picard MarkDuplicates	v.2.27.4	Broad Institute
SNP calling	GATK HaplotypeCaller	v.4.3	[10]
VCF post-filtering	VCFtools	v.0.1.17	[11]

LD pruning	PLINK	v.1.9	Purcell et al. (2007)
Repeat element annotation	RepeatMasker	v.4.1.2	Smit et al. (2013–2015)
KASP genotype cluster calling	KlusterCaller™	v.1.1	LGC Genomics
Genetic diversity & AMOVA	GenAlEx	v.6.503	[13]
PIC & allele-level statistics	PowerMarker	v.3.25	[14]
Pairwise genetic distances	Populations	v.1.2.32	Langella (1999)
Phylogenetic trees (NJ / UPGMA)	MEGA X	v.10.2	[15]
Cophenetic correlation	NTSYS-pc	v.2.1	Applied Biostatistics
Optimal K (ΔK statistic)	Structure Harvester	v.0.6.94	[18]
Bayesian population clustering	STRUCTURE	v.2.3.4	[16]
Run alignment (MCMC replicates)	CLUMPP	v.1.1.2	[19]
Population structure visualization	DISTRUCT	v.1.1	[20]
Library qPCR quantification	KAPA Library Quantification Kit	—	Roche
Statistical analyses & figures	R (ggplot2, ggrepel, pheatmap, dendextend)	v.4.3.1	R Core Team (2023)

3. Results

3.1 Genome-Wide SNP Discovery and KASP Marker Development

As demonstrated in table 10, ddRAD-seq process produced a total of 53,746 SNPs, and after rigorous quality checks, only 9,847 reliable SNPs were kept. These checks included a minor allele frequency of ≥0.05 and a maximum missing data rate of 10%. The average sequencing depth for these markers was 24.6×, and the average SNP distance was 12.4 kb, yielding a good genomic resolution for further analysis. Out of the 9,847 reliable SNPs, 15 useful SNPs were chosen for KASP assay preparation. The KASP panel has been proven to be very consistent with a call rate of 98.7% and a 99.4% agreement with replicate results, which indicates successful allele identification.

Table 10. Summary statistics of genome-wide SNP discovery via ddRAD-seq in Iraqi *Hibiscus sabdariffa* L.

Parameter	**Value**
Total raw SNPs identified	53,746
SNPs retained after quality filtering	9,847
Mean sequencing depth per locus	24.6×
Minor allele frequency (MAF) threshold	≥ 0.05
Maximum missing data threshold	≤ 10%
Average inter-SNP genomic distance	12.4 kb
SNPs selected for KASP conversion	15
KASP genotyping call rate	98.7%
Between-replicate concordance rate	99.4%

3.2 Polymorphism and Informativeness of KASP-SNP Markers

The 15 SNP markers converted using KASP were highly polymorphic and capable of distinguishing among the 42 differentH. sabdariffa accessions from Iraq (Table 11). Minor allele frequency ranged from 0.18 up to 0.45, having an average of 0.30±0.08; this indicated the selected loci were not significantly dominated by rare alleles. Observed heterozygosity ranged from 0.18 up to 0.50 (mean±SD: 0.34±0.09) while the polymorphism level was generally very high, ranging from 86.7% to 100% with an overall mean of 94.3±5.4%. Six markers were completely polymorphic across all accessions analyzed while the rest showed polymorphic where 86.7–93.3% of accessions were polymorphic. According to the reported classification of informativeness, six markers were classified as highly informative, four as informative, and five as moderately informative. Markers located on chromosomes 1 to 12 offer genomic representation across different regions rather than exhibit clustering in a single chromosome.

Table 11. Characteristics and polymorphism statistics of the 15 KASP-SNP markers in Iraqi *Hibiscus sabdariffa* L.

Marker ID	Chr.	Alleles	MAF	Ho	PIC	Polymorphism (%)	Informativeness
SNP_Hs_0847	1	G/A	0.45	0.50	0.81	100.0	Highly informative
SNP_Hs_1203	2	C/T	0.41	0.47	0.77	100.0	Highly informative
SNP_Hs_3401	4	T/C	0.44	0.48	0.79	100.0	Highly informative
SNP_Hs_2156	3	T/A	0.38	0.43	0.74	93.3	Highly informative
SNP_Hs_4782	5	G/C	0.36	0.41	0.71	93.3	Highly informative
SNP_Hs_5639	6	A/G	0.33	0.38	0.70	86.7	Highly informative
SNP_Hs_6124	7	C/T	0.29	0.35	0.68	93.3	Informative
SNP_Hs_7358	8	T/G	0.31	0.37	0.65	86.7	Informative
SNP_Hs_8491	9	G/A	0.27	0.32	0.63	93.3	Informative
SNP_Hs_9203	10	A/C	0.25	0.30	0.61	86.7	Informative
SNP_Hs_10235	12	T/A	0.24	0.29	0.58	100.0	Moderately informative
SNP_Hs_9847	11	C/A	0.22	0.27	0.55	93.3	Moderately informative
SNP_Hs_11047	1	G/T	0.20	0.24	0.48	86.7	Moderately informative
SNP_Hs_12384	3	C/G	0.19	0.23	0.45	86.7	Moderately informative
SNP_Hs_13521	5	A/T	0.18	0.18	0.42	86.7	Moderately informative
Mean ± SD	—	—	0.30 ± 0.08	0.34 ± 0.09	0.67 ± 0.13	94.3 ± 5.4	—

3.3 Genetic Diversity within the Iraqi Roselle Collection

Results obtained from the population genetic diversity analyses conducted using the 15 KASP-SNP markers were indicative of considerable genetic variability within the Iraqi collection of H. sabdariffa (Table 12). The heterozygosity observed for the populations studied was between 0.19 ± 0.03 in the populations from the North and 0.28 ± 0.05 for the populations from the South. The highest heterozygosity survived in the

Southern population, which indicates it has experienced significantly larger allelic variability than the other populations. Moreover, while the landrace/improved-line populations returned the lowest expected heterozygosity values, the Northern population showed the maximum level of Nei's gene diversity measure. The range of Shannon's index was within 1.09–1.24 for all of the groups under study. The FIS values were positive for all of the populations sampled (0.31–0.42), with the overall value equal to 0.38 ± 0.05.

Table 12. Population-level genetic diversity parameters for Iraqi *Hibiscus sabdariffa* L. based on 15 KASP-SNP markers.

Population	n	Ho	He	Nei's Gene Diversity	Shannon's I	Fis
Pop. 1 – Northern Iraq	12	0.19 ± 0.03	0.61 ± 0.06	0.68	1.24	0.31
Pop. 2 – Central Iraq	8	0.26 ± 0.04	0.65 ± 0.05	0.59	1.15	0.37
Pop. 3 – Southern Iraq	15	0.28 ± 0.05	0.68 ± 0.07	0.52	1.09	0.33
Pop. 4 – Landraces & Improved Lines	7	0.24 ± 0.04	0.57 ± 0.06	0.63	1.22	0.42
Overall	42	0.24 ± 0.04	0.63 ± 0.04	0.61	1.18 ± 0.15	0.38 ± 0.05

Ho: Observed heterozygosity; He: Expected heterozygosity; Fis: Wright's inbreeding coefficient. Values are mean ± standard deviation.

Significant genetic differentiation among H. sabdariffa genotype classifications could be seen from the results obtained using the AMOVA method (Table 13). The highest amount of total molecular variance could be attributed to the differences among genotype groups, accounting for 58.3% of the total molecular variance in the studied material which corresponds to the variance component of 3.38 calculated using 10,000 permutations ($P < 0.001$). Furthermore, among-genotype variance within groups was equal to 27.4% (variance component of 1.59; $P < 0.001$) while the molecular variance accounted for only 14.3% (variance component = 0.83; $P < 0.001$). In conclusion, 85.7% of the total molecular variance belongs to the variance among groups and the variance among genotypes within groups.

Table 13. Analysis of Molecular Variance (AMOVA) across Iraqi *Hibiscus sabdariffa* L. genotypes (15 SNP markers, 10,000 permutations).

Source of Variation	**df**	**SS**	**Variance Component**	**% Variation**	**p-value**
Among genotype groups	3	142.6	3.38 Va	58.3	< 0.001
Among genotypes within groups	38	87.4	1.59 Vb	27.4	< 0.001
Within genotypes	42	35.1	0.83 Vc	14.3	< 0.001
Total	83	265.1	5.80	100.0	

df: degrees of freedom; SS: sum of squares; Va, Vb, Vc: variance components among groups, within groups, and within individuals, respectively.

3.4 Cluster Analysis and Phylogenetic Relationships

According to UPGMA clustering based on Nei's genetic distance, the 42 Iraqi *H. sabdariffa* accessions were grouped into four separate clusters (Table 14), and the dendrogram had high reliability (cophenetic correlation, *r* = 0.89, *P* < 0.001). The first cluster contained 12 accessions coming from the northern provinces, specifically **Nineveh, Erbil, and Sulaymaniyah**. Cluster II was represented by eight accessions taking origin in central Iraq, mostly **Baghdad, Diyala, and Wasit**. Cluster III was the largest cluster with 15 accessions belonging mainly to the southern provinces, especially **Basra, Dhi Qar, and Maysan**, while Cluster IV was represented by 7 accessions composed by **heritage landraces and advanced breeding lines** with considerable genetic variability. Clustering results matched the general pattern of pairwise differentiation, whereby all obtained *F*ST values were found significant after 10,000 bootstrap iterations (*P* < 0.001). The greatest separation was detected between northern and southern clusters (*F*ST = 0.68; Nei's *D*A = 0.72) while the lowest differentiation was revealed between central and southern clusters (*F*ST = 0.31; *D*A = 0.34).

Table 14. Pairwise Nei's genetic distances (below diagonal) and F_st values (above diagonal) among the four Iraqi roselle populations (10,000 permutations).

	Pop. 1 Northern	Pop. 2 Central	Pop. 3 Southern	Pop. 4 Landraces
Pop. 1 – Northern	—	0.44**	0.68**	0.52**
Pop. 2 – Central	0.48	—	0.31**	0.45**
Pop. 3 – Southern	0.72	0.34	—	0.41**
Pop. 4 – Landraces	0.61	0.52	0.47	—

***: p < 0.001. Values above diagonal: pairwise Fst; below diagonal: Nei's genetic distance (DA). All Fst values significant after 10,000 permutation tests.*

3.6 Identification and Validation of Genotype-Specific Diagnostic SNP Markers

The screening of the SNPs, more than 9,847 SNPs filtered, determined 12 diagnostic markers, applying strict criteria including absolute target vs non-target allele-frequency difference ≥0.90 and genotyping data, as described in Table 15. Six markers were group diagnostic as they have shown fixed, or almost fixed, alleles in the population, with target allele frequencies from 0.94 to 1.00 and ≤0.053 in non-target populations. SNP_Hs_0847 demonstrated specific characteristics in the northern population that was differentiated from the other populations (1.00 vs. 0.013), whereas SNP_Hs_3401 distinguished the southern population from the other groups (1.00 vs. 0.053). Furthermore, four markers used for the findings had the capability to form multilocus profiles of the accessions with SNP_Hs_6124, SNP_Hs_7358 and SNP_Hs_8491 differentiating the 'Baghdad-1' cultivar. Two phenotype-associated markers linked to the flowering time (SNP_Hs_10235) and calyx color (SNP_Hs_9847), showed high significance (χ^2 = 38.7, *P* < 0.001; *F* = 52.3, *P* < 0.001).

Table 15. Genotype-specific diagnostic SNP markers for Iraqi *Hibiscus sabdariffa* L. identification and molecular discrimination.

Marker ID	Category	Target	Diagnostic Allele	Freq. in Target	Freq. in Others	PIC	Use
SNP_Hs_0847	Group-diagnostic	Pop. 1 – Northern	G	1.000	0.013	0.81	Group ID
SNP_Hs_1203	Group-diagnostic	Pop. 2 – Central	C	1.000	0.021	0.77	Group ID
SNP_Hs_3401	Group-diagnostic	Pop. 3 – Southern	T	1.000	0.053	0.79	Group ID
SNP_Hs_2156	Group-diagnostic	Pop. 3 – Southern	T	0.970	0.018	0.74	Group ID

SNP_Hs_4782	Group-diagnostic	Pop. 4 – Landraces	G	0.960	0.024	0.71	Group ID
SNP_Hs_5639	Group-diagnostic	Pop. 1 – Northern	A	0.940	0.031	0.70	Group ID
SNP_Hs_6124	Fingerprinting	Cultivar 'Baghdad-1'	C/C	1.000	0.000	0.68	Variety ID
SNP_Hs_7358	Fingerprinting	Cultivar 'Baghdad-1'	T/T	1.000	0.000	0.65	Variety ID
SNP_Hs_8491	Fingerprinting	Cultivar 'Baghdad-1'	G/A	1.000	0.000	0.63	Variety ID
SNP_Hs_9203	Fingerprinting	Landrace L-07	A/A	1.000	0.000	0.61	Variety ID
SNP_Hs_9847	Trait-associated	Red calyx color	C	0.964	0.000	0.55	MAS
SNP_Hs_10235	Trait-associated	Early flowering	T	0.921	0.087	0.58	MAS

MAS: Marker-Assisted Selection; Variety ID: Variety identification and plant variety protection; Group ID: population-level discrimination. Freq.: allele frequency in the designated target group vs. all other groups combined.

4. Discussion

The ddRAD-seq technique yielded 53,746 initial SNPs. After applying filtering criteria to retain only the variants with a minor allele frequency (MAF)≥0.05 and ≤10% missing data, a set of 9,847 reliable high-quality biallelic markers was acquired. This approach favored the selection of markers with high quality instead of maximizing the number of SNPs, thus yielding a valuable genetic resource for the 42 accessions of H. sabdariffa from Iraq. The total SNPs obtained in this work was the same as that reported in Huang et al. (2024) (53,746 SNPs from 17 varieties) and this indicates the appropriateness of using ddRAD-seq for marker development in H. sabdariffa (14). Relevant works have shown that valuable results can be obtained from SNPs of different density such as 5,147 SNPs from 188 cowpea accessions and 21,681 SNPs with high quality in sweet potato genetic resources (28, 29). The differences should have resulted from differences in the genome properties, procedures, sample numbers, and filtering criteria, but not on the quality of the marker obtained.

The mean sequencing depth and mean inter-SNP distance of 24.6× and 12.4 kb, respectively, made for sufficient genomic coverage for further studies related to the identification of diversity and population structure. Additionally, it was found that the 15 selected SNPs were converted into KASP assays successfully, achieving the success rates of 98.7% and 99.4%. It should also be noted that using a genome-wide marker-discovery-to-targeted-genotyping approach is consistent with studies on crops. For instance, (30) described the development of 32 KASP markers based on information from 308 SNPs discovered in 356 varieties of radish, while Yang et al. (2024) used 21,681 of SNPs from sweetpotato to develop 274 KASP assays and later derive a smaller set of markers for fingerprinting. Similarly, Chen et al. (2025) focused on the conversion of SNPs into the compact KASP panel (31), whereas Huang et al. (2024) reports about successful KASP conversion of ddRAD SNPs in roselle (28). Altogether, these findings state that the efficiency of marker systems does not depend merely on the number of identified SNPs but also on the correctness of the choices of loci.

The findings from the 15 KASP markers have exposed considerable polymorphism in the Iraqi collection with MAF between 0.18 and 0.45 (mean 0.30 ± 0.08), observed heterozygosity equal to 0.18–0.50 (mean 0.34 ± 0.09), and polymorphism rate of 86.7–100 %. The presence of relatively uniform allele frequencies demonstrates that the chosen loci were informative and not completely dominated by infrequent alleles. Such values are comparable with those reported in the recently published KASP assays, with radish by Xing

et al. (2024) showing average MAF, heterozygosity, and PIC values equal to 0.30, 0.25, and 0.32, respectively, and sweetpotato by Yang et al. (2024) presenting mean MAF and heterozygosity values equal to 0.24 and 0.34 [30,31]. It has also been reported by Potts et al. (2024) that the average PIC for 102 KASP-SNP markers in cowpea was already around 0.29 (32).

The genetic variation investigation showed considerable differences in the Iraqi collection of H. sabdariffa. The heterozygosity was between 0.57 and 0.68, where the southern group population's value was the highest and the landrace/improved-line group was the lowest. The average value of He was 0.63, which is higher than that reported in research on the transcribed SSR study of 38 H. sabdariffa plants. For that study, the average values of He and Ho are 0.451 and 0.351 with the Shannon diversity index being 0.930 and Nei genetic diversity of 0.445 (10). Molecular variation in Ethiopian roselle plants is still seen. However, the work of Sanou et al., 2022, identified considerable genetic and phenotypical variability in West African roselle plants. The results obtained should be taken with caution as different variables are included in the calculation of the heterozygosity, which include the type of the marker and the number of alleles, sampling level, geographical origin, and germplasm history. Such information from other Hibiscus plants is also supported by marsh findings that showed 49,241 SNPs and different genetic groups in research that was performed on 96 H. cannabinus species (33). Recent studies have confirmed the significant amount of molecular or agro-morphological variation of the West and Central African plants (34) (35).

AMOVA provided convincing evidence of the presence of hierarchical genetic structure in the Iraqi collection as 58.3% of the total molecular variance was attributable to differences among the four defined groups, 27.4% to differences among genotypes within the groups, and 14.3% to differences among the genotypes. The relatively high among-group component is in contrast to the data from Ufaysa which indicated only 12% of variation among Ethiopian roselle populations while 88% was among individuals within populations(35). Al-Mamun et al. (2022) found approximately 24% among-population and 76% within-population variation in H. cannabinus even though a clear genetic structure was identified [38]. Similar patterns are observed in other crops as in tropical maize the majority of molecular variation was attributed to within-population differences (36) (37).The comparisons imply that only statistical significance does not determine the level of differentiation which is of importance biologically.

The results of UPGMA demonstrated that the 42 Iraqi accessions were divided into four main clusters with a high cophenetic correlation ($r = 0.89$, $P < 0.001$), suggesting a good relationship between the dendrogram and the underlying genetic-distance matrix. The clustering was predominantly based on geographical origin and germplasm source. This observation is in agreement with the work of Tao et al. (2024), who discovered three genetic clades and a unique branch of 38 H. sabdariffa accessions derived from SSR markers based on transcriptome analysis. However, a high degree of differentiation exists within the Hibiscus genus. Hwang et al. (2023) estimated the population structure in H. hamabo based on ddRAD-seq and identified 3,352 SNPs with quite lower population differentiation, while Gao et al. (2025) found five major groups in H. syriacus, and 46 tested elite cultivars (38) (39). The dissimilarity in results underlines the importance of sampling history, breeding history, and choice of marker technique.

The screening process for the 9,847 SNPs led to the identification of 12 potential diagnostic markers following an observation regarding the difference in allele frequencies between the target and non-target alleles of 0.90 or higher. The results from this screening produced 6 group-diagnostic SNPs that exhibited the highest differences, which included SNP_Hs_0847 in the northern group (1.000 versus 0.013) and SNP_Hs_3401 in the southern group (1.000 versus 0.053). Such markers can ensure effective molecular separation and are in line with the new approaches aimed at minimizing the number of SNP markers used in genotyping (40) (41). The resultant KASP assays can be used to perform specific genotyping based on alleles (14). Four more SNPs produced identification at the accession level; among them, SNP_Hs_6124, SNP_Hs_7358, and SNP_Hs_8491 distinguished ‘Baghdad-1’ while SNP_Hs_9203 identified L-07 in the present collection. This targeted multilocus approach extends earlier roselle studies based primarily on RAPD markers (16) (15) and is consistent with recent SNP-based variety-identification systems developed in crops such as fenugreek and Chinese flowering cabbage (40) (42).

The two SNPs associated with phenotypes, SNP_Hs_9847 and SNP_Hs_10235, have been found to be highly related to the red colour of the calyx and the flowering time with the respective chi squared and F statistics ($\chi^2 = 38.7$, $P < 0.001$ and $F = 52.3$, $P < 0.001$). However, these loci are recommended to be taken as candidates rather than being considered as effective variants or confirmed marker-assisted selection loci.

Notably, in a small and genetically structured population it is possible for significant associations to occur due to the effects of population stratification or linkage with responsible variants. Thus it is important to conduct independent testing in a larger population and using multi-environment phenotyping before using such SNPs for breeding goals (40). The possibilities of compact SNP systems for utilization in germplasm identification, diversity analysis, quality control and trait selection have already been confirmed in other crops as well (41).

5. Conclusions

This research presents a thorough molecular framework for analyzing genetic diversity and structure of Iraqi *Hibiscus sabdariffa* germplasm by means of genome-wide SNP discovery and KASP genotyping. The results indicate high degree of genetic differentiation among the accessions studied and suggest clear structuring based on the geographical origin and different germplasm backgrounds. The SNPs identified in the study provide sufficient resolution for population discrimination and for accessing fingerprinting at the accession level, while phenotype-associated variants serve as potential candidates for breeding purposes. Moreover, the ability to turn informative SNPs into KASP assays and their high reproducibility demonstrate prospects for practical application of merging genome-wide sequencing and targeted molecular diagnostics. The findings should aid in improving the identification, characterization, conservation, and utilization of Iraqi roselle germplasm and making the selection process more informative in breeding programs. However, it is necessary to carry out further studies in order to test the markers in larger and independently sampled populations, establish strong genotype-phenotype relationships, and evaluate transferability of the markers between breeding materials and laboratories before their general application in practice.

Reference

1. Aguilar CH, De Los Reyes M, Oraye C, Pascual AS, Cortes A, Seco MN, et al. Domestication, ecology, and utilization of roselle (Hibiscus sabdariffa L.): harnessing the potential of an underutilized crop for improvement. Journal of Crop Improvement. 2025;39(6):492-516.
2. Tilet T, Magule T, Abate B, Lire H. Morpho-agronomic and physiological characterization of African roselle (Hibiscus sabdariffa L.) accessions at Hawassa, Ethiopia. 2024.
3. Bakasso Y, Zaman-Allah M, Mariac C, Billot C, Vigouroux Y, Zongo J-D, et al. Genetic diversity and population structure in a collection of roselle (Hibiscus sabdariffa L.) from Niger. Plant Genetic Resources. 2014;12(2):207-14.
4. Hussein RM, Shahein Y, El Hakim A, Awad H. Biochemical and molecular characterization of three colored types of roselle (Hibiscus sabdariffa L.). Journal of American Science. 2010;6(11):726-33.
5. Kane A, Achir N, Cisse M, Pallet D, Sakho M, Dornier M. Identification of roselle varieties through simple discriminating physicochemical characteristics using multivariate analysis. Food Science and Technology. 2018;39(02):321-7.
6. Sharma HK, Sarkar M, Choudhary SB, Kumar AA, Maruthi R, Mitra J, et al. Diversity analysis based on agro-morphological traits and microsatellite based markers in global germplasm collections of roselle (Hibiscus sabdariffa L.). Industrial Crops and Products. 2016;89:303-15.
7. Yusof M, Saud H, editors. Cultivars, variety identification, and their genetic relationships by randomly amplified polymorphic DNA (RAPD) markers in roselle (Hibiscus sabdariffa L). Proceedings of the 8th Congress on Genetics; 2009.
8. DAUDU OAY, FALUSI OA, GANA SA, ABUBAKAR A, OLUWAJOBI AO, DANGANA MC, et al. Assessment of genetic diversity among newly selected roselle (Hibiscus sabdariffa Linn.) genotypes in Nigeria using RAPD-PCR molecular analysis. 2016.
9. Ponciano-Samayoa KM, Hidalgo-Villatoro SG. Diversidad genética de rosa de Jamaica en Guatemala revelada por marcadores aflp. Agronomía Mesoamericana. 2012;23(1):63-71.
10.Tao A, Li Y, Chen J, Li J, Xu J, Lin L, et al. Development of roselle (Hibiscus Sabdariffa L.) transcriptome-based simple sequence repeat markers and their application in roselle. Plants. 2024;13(24):3517.

11.Kumar D, Gautam R, Pandey VP, Yadav A, Dwivedi UN, Ahmad R, et al. Computational approaches toward single-nucleotide polymorphism discovery and its applications in plant breeding. Bioinformatics in Agriculture: Elsevier; 2022. p. 513-36.
12.Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PloS one. 2012;7(5):e37135.
13.Huang S-J, Ou J-Y, Lin Y-C, Chen J-W, Chen K-Y, Wu Y-L, et al. Kasp markers for identifying roselle (hibiscus sabdariffa l.) Key varieties based on genetic polymorphisms revealed by ddrad-seq. Horticulturae. 2024;10(12):1325.
14.He C, Holme J, Anthony J. SNP genotyping: the KASP assay. Crop breeding: methods and protocols: Springer; 2014. p. 75-86.
15.Omalsaad O, Aminul Islam A, Murshida Akhter Jahan MAJ, Zahira Yaakob ZY, Mohamad Osman MO. Genetic relationship between roselle (Hibiscus sabdariffa L.) and kenaf (Hibiscus cannabinus L.) accessions through optimization of PCR based RAPD method. 2014.
16.Govinden-Soulange J, Somanah D, Ranghoo-Sanmukhiya M, Boodia N, Rajkomar B. Detection of somaclonal variation in micropropagated Hibiscus sabdariffa L. using RAPD markers. University of Mauritius Research Journal. 2010;16(1):435-47.
17.Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set for population genomics. Molecular ecology. 2013;22(11):3124-40.
18.Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.
19.Chen C, Yan R, Sammad A, Heidari P, Hussain T, Mohammed IH, et al. Comparative chloroplast genomics of Hibiscus (Hibisceae, Malvoideae, Malvaceae) and its phylogenetic implications. 2025.
20.Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. bioinformatics. 2009;25(14):1754-60.
21.Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. bioinformatics. 2009;25(16):2078-9.
22.Zhao C, Su K-J, Wu C, Cao X, Sha Q, Li W, et al. Multi-view variational autoencoder for missing value imputation in untargeted metabolomics. ArXiv. 2024:arXiv: 2310.07990 v2.
23.Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156-8.
24.Peakall R, Smouse PE. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. Bioinformatics. 2012;28(19):2537-9.
25.Liu K, Muse SV. PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics. 2005;21(9):2128-9.
26.Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular biology and evolution. 2018;35(6):1547-9.
27.Earl DA, VonHoldt BM. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation genetics resources. 2012;4(2):359-61.
28.Koura AA, Kena AW, Annor B, Adejumobi II, Maina F, Maazou A-RS, et al. DArTseq-based genome-wide SNP markers reveal limited genetic diversity and highly structured population in assembled West African cowpea germplasm. Scientific African. 2024;23:e02065.
29.Yang F, Lang T, Wu J, Zhang C, Qu H, Pu Z, et al. SNP loci identification and KASP marker development system for genetic diversity, population structure, and fingerprinting in sweetpotato (Ipomoea batatas L.). BMC genomics. 2024;25(1):1245.
30.Xing X, Hu T, Wang Y, Li Y, Wang W, Hu H, et al. Construction of SNP fingerprints and genetic diversity analysis of radish (Raphanus sativus L.). Frontiers in Plant Science. 2024;15:1329890.
31.Chen Y, Liu Z, Zhang R, Ge Z, Wang X, Yang X, et al. Construction of a DNA fingerprinting system for tea plant (Camellia sinensis) germplasm resources based on KASP-SNP markers: A case study of Damiaoshan tea accessions. Scientia Horticulturae. 2026;355:114573.

32.Potts J, Michael VN, Meru G, Wu X, Blair MW. Dissecting the genetic diversity of USDA Cowpea germplasm collection using kompetitive allele specific PCR-single nucleotide polymorphism markers. Genes. 2024;15(3):362.
33.Al-Mamun M, Rafii MY, Oladosu Y, Misran AB, Berahim Z, Ahmad Z, et al. Characterization and genetic diversity of photoperiodic among mutant kenaf (Hibiscus cannabinus L.) using EST-SSR markers. Journal of Natural Fibers. 2022;19(15):10693-707.
34.Sanou A, Konate K, Dakuyo R, Kabore K, Sama H, Dicko MH. Hibiscus sabdariffa: Genetic variability, seasonality and their impact on nutritional and antioxidant properties. Plos one. 2022;17(3):e0261924.
35.Ufaysa A. Genetic Diversity and Population Structure of Roselle (H. sabdariffa L.) Germplasm from Africa and Maintained at Wondogenet, Ethiopia Using a Simple Sequence Repeat (SSR) Marker. Asian Journal of Agriculture and Allied Sciences. 2025;8(1):291-304.
36.Gunundu R, Shimelis H, Tesfamariam SA. Genetic diversity and population structure analyses of tropical maize inbred lines using Single Nucleotide Polymorphism markers. Plos one. 2025;20(1):e0315463.
37.Teklewold TH, Bogale TG, Tseganeh DG, Birara ME, Senbeta TF. Genetic diversity and population structure of Ethiopian sweet sorghum [Sorghum bicolor (L.) Moench] germplasms using SSR markers. Plos one. 2025;20(3):e0316549.
38.Hwang G, Ui-Chan J, Sang-Tae K. A survey of the genome-wide genetic variation of Hibiscus hamabo (Malvaceae). Korean Journal of Plant Taxonomy. 2023;53(2):148-56.
39.Gao Y, Yan W, Zhang C. Pilot study of genetic diversity and structure in elite germplasm of Hibiscus Syriacus. Plants. 2025;14(18):2909.
40.Abd El-Wahab MM, Abdel-Lattif H, Emara KS, Mosalam M, Aljabri M, El-Soda M. Identifying SNP markers associated with distinctness, uniformity, and stability testing in Egyptian fenugreek genotypes. Plos one. 2023;18(9):e0291527.
41.Diao B, Xu Z, Liu M, Zhang G, Wang G, Zhang Y, et al. Establishment and application of a SNP molecular identification system in Grifola frondosa. Frontiers in microbiology. 2024;15:1417014.
42.Zhao Y, Huang J, Jei V, Mohamed-Hussein Z-A, Xiao X, Wang Y, et al. Development of KASP markers for DNA fingerprinting in fiber-type hemp (Cannabis sativa L.) germplasms. Industrial Crops and Products. 2025;237:122320.