Shahid Bahonar University of Kerman
Agricultural Biotechnology Journal
Iranian Biotechnology Society
Print ISSN: 2228-6705 Online ISSN: 2228-6500

Molecular identification and genetic diversity analysis of *Ocimum basilicum* L. using chloroplast *matK* sequences and Genome-wide SNP markers

Ola Hatif Hazim
* Corresponding author. College of Science, Mustansiriyah University, Baghdad, Iraq. E-mail: ola.hataf@uomustansiriyah.edu.iq

Maysam Ihsan Ali Al-Hassany
College of Science, Mustansiriyah University, Baghdad, Iraq. E-mail: maysam.ihssan@uomustansiriyah.edu.iq

Zinah Ameer Abbood
College of Science, Mustansiriyah University, Baghdad, Iraq. E-mail: Zina_ameer@uomustansiriyah.edu.iq

Abstract

Objective

The economically significant aromatic crop sweet basil (*Ocimum basilicum* L.) needs thorough genetic analysis for breeding optimization and germplasm conservation. To clarify genetic diversity and population organization in farmed basil, we combined genome-wide SNP markers with chloroplast DNA barcoding (*matK*).

Materials and methods

A total of thirty different cultivated *O. basilicum* accessions were analyzed (10 accessions from three distinct geographic locations) in Baghdad, Iraq. The *matK* gene from each accession was amplified and sequenced using the Sanger sequencing. Genomic SNPs were generated using the genotyping-by-sequencing (GBS) method. After passing a quality control assessment, 6234 of the SNPs were considered valid high-quality SNPs. Genetic diversity, population structure, and differentiation were analyzed using haplotype analysis, STRUCTURE analysis, principal coordinates analysis (PCA), and statistical evaluation methods, including analysis of molecular variance (AMOVA) and Fst pairwise statistics.

Results

MatK sequencing revealed moderate genetic diversity among eight haplotypes was observed (Hd = 0.685 ± 0.045; π = 0.0076 ± 0.0012), whereas an excess of heterozygotes was identified in the SNP analysis (Ho = 0.092 ± 0.038; He = 0.251 ± 0.072; Fis = 0.634 ± 0.089; P < 0.001). Bayesian clustering suggests three genetically distinct populations (K = 3), while AMOVA results indicate 24.3% of the overall genetic variation is attributed to the population (Φst = 0.243, P < 0.001).

Collectively, SNP markers exhibit relatively greater resolution than *matK* and demonstrate a moderate degree of correlation with *matK*-based genetic relationships (Mantel r = 0.523, P = 0.002).

Conclusions

Independent conservation is required for three genetically distinct populations with limited gene flow. The pronounced heterozygote deficit suggests that crosses between genetically distinct populations may enhance heterosis in breeding programs. Our results provide new biological insights into the genetic architecture of cultivated basil. These insights include the presence of three well-defined genetic clusters, a strong but structured lack of heterozygosity, and partial concordance between the chloroplast and nuclear genomes. These findings could improve our understanding of gene flow, domestication effects, and population differentiation in *O. basilicum*. Furthermore, they could provide a genomic framework for future breeding and conservation strategies.

Keywords: conservation genomics, DNA barcoding, heterozygote deficiency, population genetics, SNP genotyping

Paper Type: Research Paper.

Citation: Hazim, O. H., Al-Hassany, M. I. A., & Abbood, Z. A. (2026). Molecular identification and genetic diversity analysis of *Ocimum basilicum* L. using chloroplast *matK* sequences and Genome-wide SNP markers. *Agricultural Biotechnology Journal,* 18(4), 259-276.

Agricultural Biotechnology Journal, 18(4), 259-276. DOI: 10.22103/jab.2026.27125.1888

Received: April 04, 2026. Received in revised form: June 07, 2026.

Accepted: June 08, 2026. Published online: August 30, 2026.

Publisher: Shahid Bahonar University of Kerman & Iranian Biotechnology Society.

Introduction

Medicinal and aromatic plants are among the most important natural resources used for disease prevention and treatment (Amirteymoori et al., 2021). They are widely distributed across tropical, subtropical, and temperate regions of the world and continue to attract scientific interest. Because they have biologic active compounds that used in pharmaceutical, food, and cosmetic industries (Zhakipbekov et al., 2024). Basil belongs to the mint family (Lamiaceae), a family rich in medicinally important plant species, comprising over 7,000 species distributed across more than 230 genera. Basil is known for its significant genetic and phenotypic diversity, with numerous varieties and strains differing in leaf shape, flower colour, essential oil aroma, and the chemical composition of its active compounds (Avetisyan et al., 2017). In order to maximize this plant's productivity and guarantee the quality of its aromatic and therapeutic products, there is a growing interest in researching its genetic traits. A variety of secondary compounds, including volatile oils (linalool, eugenol, and methyl chavicol), flavonoids, terpenes, and phenolic

compounds, are present in *Ocimum basilicum L.* and have highlighted effectiveness as antioxidants, antibacterial, and antifungal agents, as well as anti-inflammatory and some anti-tumor activities (Patel et al., 2015; Letsiou et al., 2024). Numerous studies have demonstrated that the variability in biological activity across basil types is mostly related to genetic variation between various strains (Abdelaziz et al., 2024). Plants, like basil, are traditionally categorized using morphological and anatomical traits such as leaf form, flower type, stem length, and phenological traits. However, environmental factors like soil type, temperature, humidity, and light intensity frequently influence these plant features, resulting in overlapping traits and making it challenging to identify between genetically related species and cultivars (Letsiou et al., 2024). Morphological traits remain valuable for taxonomic and breeding studies. However, they may be influenced by environmental factors, which can limit their reliability in distinguishing closely related genotypes. Because molecular approaches are based on DNA variation and are largely unaffected by environmental conditions (Arabpour et al., 2021). They have become reliable tools for plant identification and genetic diversity analysis (Farahvashi et al., 2026). Plant taxonomy and the study of genetic linkages among them have undergone a qualitative change as a result of significant advancements in molecular biology tools like gene sequencing and polymerase chain reaction (PCR) (Kassem, 2025). Because of their relatively autonomous genome, favorable mutation rate, and monogenic inheritance in the majority of flowering plants, chloroplasts are one of the most significant genetic resources employed in molecular plant studies. Several plastid genes, including rbcL, *matK*, and trnH-psbA, have been employed as molecular markers in genetic diversity study and plant identification (Du et al., 2025). Because of its placement within the trnK intron and its greater evolutionary rate than other conserved plastid genes, the *matK* gene (maturase K) is one of the most employed plastid genes in plant DNA barcoding research. Because of this, it provides moderate to high discriminatory power for differentiating many plant taxa while remaining relatively conserved within species (Kirankumar et al., 2023, Kumar et al., 2025). The *matK* gene is one of the standard genes for molecular identification of flowering plants, according to a number of international organizations, including the Consortium for the Barcode of Life (CBOL) (CBOL, 2009). The effectiveness of the *matK* gene in assessing genetic diversity and examining evolutionary links in aromatic and medicinal plants, including members of the mint family (Lamiaceae), has been shown in numerous studies. These investigations have highlighted the gene's capacity to identify minute genetic variations among different cultivars and strains, advancing our knowledge of the genetic makeup and evolutionary history of plants (Shah et al., 2026). Recent research has discovered significant genetic variation across local and global types, evident in differences in chemical composition, volatile oils, and biological activity. In order to assist plant breeding programs, preserve genetic resources, and guarantee the best possible use of plants in industrial and medical domains, it is crucial to study this genetic variety. Furthermore, considering the significant overlap between Ocimum species-which are frequently challenging to differentiate based only on morphological traits-a precise molecular definition of *Ocimum basilicum L.* is especially crucial. The dependability of scientific study can be impacted by definition errors, which might provide erroneous results in pharmacological and chemical investigations (Govindaraj et al., 2015). In addition to their role in preventing local varieties from

going extinct and supporting biodiversity conservation strategies, particularly in regions under increasing environmental pressure, genetic diversity studies based on the *matK* gene also help create accurate genetic databases that can be consulted in future research (Kirankumar et al., 2023). Although many studies have been conducted on genetic diversity in *Ocimum basilicum*, many questions remain unanswered. Because these studies have focused either on chloroplast markers or limited sets of nuclear markers, these cannot fully capture genome-wide diversity. Also, direct comparisons between maternally and biparentally inherited genomes are not possible. Furthermore, many of them have analyzed limited germplasm collections that often lack representation of locally adapted cultivated populations. In the present study, chloroplast *matK* barcoding was integrated with genome-wide SNP markers to address these limitations. This work provides a more comprehensive assessment of genetic diversity and population structure. The samples studied in this study represent three geographically distinct cultivated populations that have not been jointly analyzed in previous studies using a combined cpDNA and SNP approach.

Materials and methods

Plant material and DNA extraction: During the 2023 growing season, thirty *Ocimum basilicum* L. accessions were gathered from three geographically separate places (10 samples each location). Fresh leaf tissue was kept at -20°C after being dehydrated in silica gel. The DNeasy Plant Mini Kit (Qiagen, Germany) was used to extract total genomic DNA in accordance with the manufacturer's instructions. A260/A280 ratios of 1.8–2.0 were deemed acceptable when evaluating DNA quality and concentration using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA). For use in subsequent processes, DNA samples were diluted to 20 ng/µL.

***matK* gene amplification and sequencing:** The universal primers *matK*-3F_KIM (5'-CGTACAGTACTTTTGTGTTTACGAG-3') and *matK*-1R_KIM (5'-ACCCAGTCCATCTGGAAATCTTGGTTC-3') were used to amplify the chloroplast *matK* gene. [17]. 12.5 µL of GoTaq Green Master Mix (Promega, USA), 1.0 µL of each primer (10 µM), and 2.0 µL of template DNA (20 ng/µL) were used in PCR reactions (25 µL). 95℃ for five minutes, 35 cycles of 94°C for thirty seconds, 52°C for forty-five seconds, 72°C for one minute, and a final extension at 72°C for seven minutes were the thermal cycling conditions. The QIAquick PCR Purification Kit (Qiagen) was used to purify the PCR products, and an ABI 3730xl DNA Analyzer (Applied Biosystems, USA) was used for bidirectional sequencing. Geneious Prime 2023.1.2 was used to assemble and modify the sequences, producing 798 bp alignments.

Genotyping-by-sequencing (GBS): GBS libraries were prepared using a double-digest restriction-site associated DNA (ddRAD) protocol (Peterson et al., 2012). Genomic DNA (200 ng) was digested with PstI-HF and MspI (New England Biolabs, USA), ligated with barcoded adapters, and size-selected (300-500 bp) using a Pippin Prep system (Sage Science, USA). The multiplexed library was sequenced on an Illumina NovaSeq 6000 platform (150 bp paired-end). Raw reads were demultiplexed using Stacks v2.62 (Rochette et al., 2019) and quality-filtered with Trimmomatic v0.39 (Bolger et al., 2014) (LEADING:20, TRAILING:20, MINLEN:100). Filtered reads were aligned to the *O. basilicum* reference genome (GCA_019657835.1) using BWA-MEM v0.7.17 (Li, 2013), and SNPs were called using GATK v4.3.0.0 HaplotypeCaller

(McKenna et al., 2010). Raw SNPs (n = 12,456) were filtered using VCFtools v0.1.16 (Danecek et al., 2011) with criteria: minimum quality Q ≥ 30, maximum missing data ≤ 20%, minimum minor allele frequency (MAF) ≥ 0.05, and depth 10-100×, retaining 6,234 high-quality SNPs (50.0% retention rate).

Genetic diversity analysis-*matK* sequence analysis: *matK* sequences (n = 29) were aligned using ClustalW in MEGA X v10.2.6 (Kumar et al., 2018). Haplotype diversity (Hd), nucleotide diversity (π), and transition/transversion ratio (Ts/Tv) were calculated using DnaSP v6.12.03 (Rozas et al., 2017). Haplotypes were identified based on nucleotide polymorphisms across the 798 bp alignment.

SNP-based diversity indices: For the 6,234 SNPs, observed heterozygosity (Ho), expected heterozygosity (He), inbreeding coefficient (Fis), polymorphic information content (PIC), Shannon's information index (I), and effective number of alleles (Ne) were calculated using GenAlEx v6.503 (Peakall & Smouse, 2006). Standard deviations were computed across loci, and the significance of Fis was tested using 1,000 permutations.

Population structure analysis: Bayesian clustering was performed using STRUCTURE v2.3.4 (Pritchard et al., 2000) with the admixture model and correlated allele frequencies. K (number of genetic clusters) was tested from 1 to 5 with 10 independent runs per K, each consisting of 100,000 MCMC iterations after 50,000 burn-in. Optimal K was determined using the Evanno method (Evanno et al., 2005) (ΔK statistic) in Structure Harvester v0.6.94 (Earl & vonHoldt, 2012). Membership coefficients were averaged across replicates using CLUMPP v1.1.2 (Rosenberg, 2004; Jakobsson & Rosenberg, 2007) and visualized with Distruct v1.1 (R Core Team, 2014). Principal component analysis (PCA) was performed on the SNP genotype matrix using R v4.3.1 (R Core Team, 2014). Missing genotypes were imputed using mean allele frequencies. The proportion of variance explained by each principal component was calculated from eigenvalues.

Population differentiation: Hierarchical analysis of molecular variance (AMOVA) was conducted using GenAlEx v6.503 to partition genetic variance among and within populations. Significance was assessed using 1,000 permutations. Pairwise Fst values were calculated in Arlequin v3.5.2.2 (Excoffier & Lischer, 2010) with 10,000 permutations and Bonferroni correction (α = 0.0167). Gene flow (Nm) was estimated as $Nm = (1 - Fst) / 4Fst$.

Comparative analysis: Correlation between genetic distance matrices from *matK* (Tamura-Nei model) and SNPs (Nei's distance) was assessed using the Mantel test with 9,999 permutations in the R package ade4 v1.7-22 (Dray & Dufour, 2007). Discriminatory power was compared using probability of identity (PI) calculated from haplotype/allele frequencies (Waits et al., 2001). All statistical analyses used an α = 0.05 significance threshold. Data visualization was performed using R packages ggplot2 v3.4.2 (Wickham, 2016) and pheatmap v1.0.12 (Kolde, 2019).

Results

Molecular characterization-*matK* gene sequencing and diversity: The chloroplast *matK* gene was successfully amplified and sequenced from 29 of 30 *Ocimum basilicum* L. accessions, yielding a 96.7% success rate. The aligned sequences were 798 bp in length and contained 28

variable sites (3.51% polymorphism). The transition/transversion ratio was 2.14. Eight distinct haplotypes were identified across the three populations (Table 1 and Figure 1). Population 2 exhibited the highest genetic diversity with Hd = 0.711 ± 0.076 and π = 0.0081 ± 0.0014, while Population 3 showed the lowest diversity (Hd = 0.578 ± 0.091, π = 0.0063 ± 0.0011). Overall haplotype diversity across all populations was moderate (Hd = 0.685 ± 0.045).

Table 1. Genetic diversity indices based on *matK* sequences

Population	N	Length (bp)	Variable Sites	Haplotypes	Hd ± SD	π ± SD	Ts/Tv
Pop 1	10	798	12	4	0.667 ±0.082	0.0074 ±0.0015	2.18
Pop 2	10	798	15	4	0.711 ± 0.076	0.0081 ±0.0014	2.12
Pop 3	9	798	9	3	0.578 ±0.091	0.0063 ±0.0011	2.10
Overall	29	798	28	8	0.685 ±0.045	0.0076 ±0.0012	2.14

N = number of samples; Hd = haplotype diversity; π = nucleotide diversity; Ts/Tv = transition/transversion ratio.

SNP discovery and filtering: Genotyping-by-sequencing generated 12,456 raw SNPs across all 30 accessions. After quality filtering (minimum Phred quality score Q30, maximum 20% missing data, and minimum minor allele frequency MAF ≥ 0.05), 6,234 high-quality SNPs were retained, representing a 50.0% retention rate. A mean minor allele frequency of 0.172 ± 0.088 and a transition/transversion ratio of 2.11 were found in the filtered SNP dataset.

Genetic diversity analysis-SNP-based diversity indices: Analysis of 6,234 SNPs demonstrated high genetic variation among the three groups (Table 2). Overall observed heterozygosity (Ho = 0.092 ± 0.038) was significantly lower than expected heterozygosity (He = 0.251 ± 0.072), resulting in a positive inbreeding coefficient (Fis = 0.634 ± 0.089, P < 0.001). This pattern indicates substantial heterozygote deficiency, likely resulting from inbreeding and/or population substructure. Ho stands for observed heterozygosity, He for expected heterozygosity, Fis for inbreeding coefficient, and PIC for polymorphic information content; I is the diversity index of Shannon; Ne is the effective number of alleles. Genetic diversity was highest in Population 2 (He = 0.289, Shannon's I = 0.456), then in Population 1 (He = 0.267, I = 0.412), and lowest in Population 3 (He = 0.197, I = 0.326). The SNP markers were moderately informative, as shown by the polymorphic information content (PIC), which varied from 0.191 to 0.245.

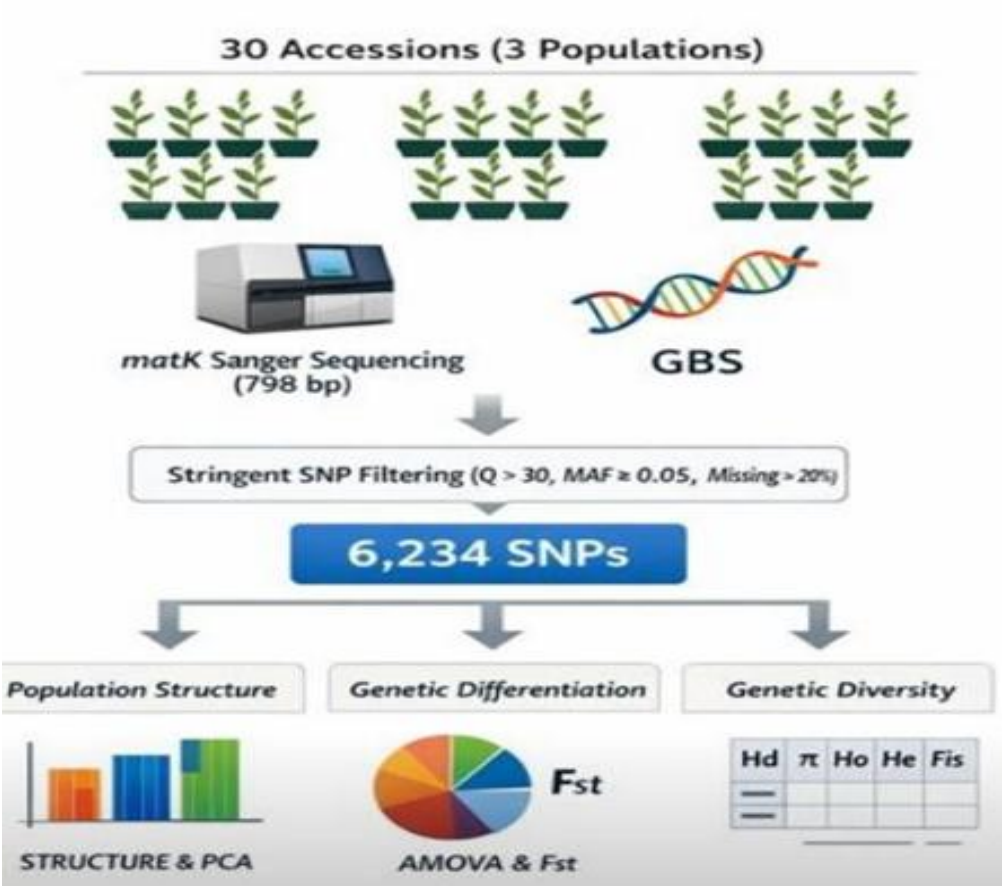

Figure 1. Workflow of molecular and population genetic analyses, including *matK* sequencing, ddRAD-based SNP discovery, population structure analysis, and diversity estimation in *Ocimum basilicum* populations

Table 2. SNP-based genetic diversity indices

Population	N	Ho ± SD	He ± SD	Fis ± SD	PIC ± SD	Shannon's I ± SD	Ne
Pop 1	10	0.098 ±0.041	0.267 ±0.068	0.633 ±0.092	0.221 ±0.061	0.412 ±0.098	1.42
Pop 2	10	0.105 ±0.038	0.289 ±0.071	0.637 ±0.086	0.245 ±0.067	0.456 ±0.102	1.51
Pop 3	10	0.073 ±0.035	0.197 ±0.064	0.629 ±0.089	0.191 ±0.058	0.326 ±0.089	1.28
Overall	30	0.092 ±0.038	0.251 ±0.072	0.634 ±0.089	0.219 ±0.065	0.398 ±0.104	1.40

Population genetic structure- Bayesian clustering analysis: To determine the ideal number of genetic clusters (K), STRUCTURE analysis was used. With the largest ΔK value of 412.8, K = 3 was determined to be optimal by the Evanno technique (Table 3). The three populations exhibited low mixing and distinct genetic divergence at K = 3. Cluster 1 membership was 90% in Population 1, Cluster 2 membership was 88% in Population 2, while Cluster 3 membership was 86% in Population 3. The upper limit of K was selected based on the number of sampled populations and expected genetic structure.

Table 3. Structure analysis for determining optimal K

K	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	Ln''(K)	ΔK	Optimal
1	-19,234.5	45.2	-	-	-	-
2	-18,892.3	38.7	342.2	99.8	2.58	-
3	-18,456.3	32.1	436.0	193.6	412.8	✓
4	-18,214.1	56.8	242.2	51.4	0.90	-
5	-18,023.3	89.3	190.8	-	-	-

K = number of clusters; Ln'(K) = first derivative; Ln''(K) = absolute value of second derivative; ΔK = Evanno's delta K.

Principal component analysis: Principal component analysis, based on 6,234 SNPs, indicated a clear separation among the three populations (Table 4). The first two principal components (PC1 and PC2) together explained 54.1% of the total genetic variance. While PC2 (22.7% variance) discriminated between Populations 1 and 2, PC1 (31.4% variance) separated Populations 3 from Populations 1 and 2.

Table 4. Principal component analysis variance based on 6,234 SNPs

Component	Eigenvalue	% Variance	Cumulative %
PC1	1,954.7	31.4%	31.4%
PC2	1,412.8	22.7%	54.1%
PC3	687.3	11.0%	65.1%
PC4	523.9	8.4%	73.5%
PC5	412.6	6.6%	80.1%

Population differentiation- Analysis of molecular variance (AMOVA): A significant genetic difference between the three groups was shown using hierarchical AMOVA (Table 5). 24.3% of the total genetic diversity was divided among populations and 75.7% within populations, both of which were highly significant ($P < 0.001$) according to the analysis. Moderate genetic differentiation was suggested by the total fixation index ($\Phi st = 0.243$).

Table 5. Analysis of molecular variance (AMOVA) based on SNPs

Source of Variation	df	Sum of Squares	Mean Squares	Variance Component	% Variance	Φ-statistic	P-value
Among populations	2	3,456.8	1,728.4	18.42	24.3%	Φst = 0.243	<0.001
Within populations	27	4,892.3	181.2	57.34	75.7%	-	<0.001
Total	29	8,349.1	-	75.76	100%	-	-

df = degrees of freedom; 1,000 permutations.

Pairwise population differentiation: All comparisons were highly significant ($P < 0.001$), with pairwise Fst values between populations ranging from 0.156 to 0.312 (Table 6). Populations 2 and 3 showed the greatest differentiation (Fst = 0.312), whilst Populations 1 and 2 reflected the least (Fst = 0.156). Restricted gene exchange was indicated by the inverse correlation between

gene flow estimates (Nm) and Fst values, which ranged from 0.55 (Pop2-Pop3) to 1.35 (Pop1-Pop2) (Figure 2).

Table 6. Pairwise genetic differentiation (Fst) and gene flow (Nm)

	Pop 1	Pop 2	Pop 3
Pop 1	-	0.156***	0.287***
Pop 2	0.156***	-	0.312***
Pop 3	0.287***	0.312***	-

*** P < 0.001. Above diagonal = Fst values. Pairwise gene flow estimates (Nm) are provided below the diagonal. Gene flow (Nm): Pop1-Pop2 = 1.35, Pop1-Pop3 = 0.62, Pop2-Pop3 = 0.55.

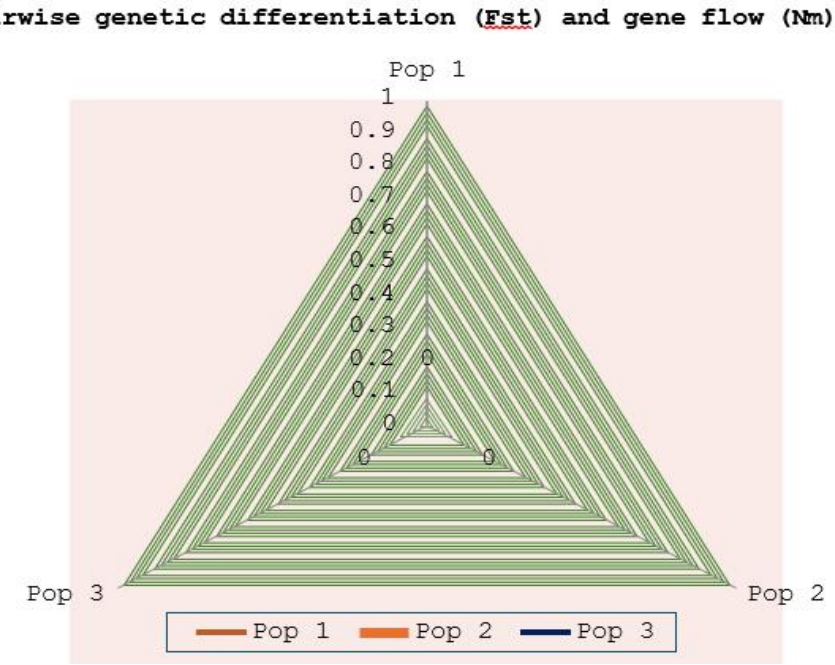

Figure 2. Pairwise genetic differentiation (Fst) and gene flow (Nm)

Comparative analysis of markers-Concordance between *matK* and SNP markers: Moderate concordance (71.4% agreement in population designations) was found when genetic inferences from *matK* sequences and SNP markers were compared. The Mantel test revealed a moderate correlation between the genetic distance matrices obtained from the two marker systems ($r = 0.523$, $P = 0.002$), suggesting that the patterns of the nuclear and chloroplast genomes are generally consistent. However, SNP markers highlighted substantially superior discriminatory power compared to *matK* (Table 7). The probability of identity (PI) for SNPs was 2.8×10^{-26}, compared to 0.094 for *matK*, representing a 3.4×10^{24}-fold improvement. SNPs also indicated higher population differentiation ($\Phi st = 0.243$) than *matK* ($\Phi st = 0.168$), with 1.45-fold greater variance partitioned among populations. While SNP markers provide vastly superior resolution

for population genetic studies, *matK* sequencing offers a cost-effective and time-efficient alternative for basic genetic identification and preliminary diversity assessment.

Table 7. Comparative performance of *matK* and SNP markers

Parameter	*matK*	SNPs	Ratio (SNP/*matK*)
Success rate	96.7% (29/30)	100% (30/30)	1.03×
Polymorphic loci	28 sites	6,234 loci	222.6×
Discriminatory power (PI)	0.094	2.8×10^{-26}	3.4×10^{24}×
Population differentiation (Φst)	0.168	0.243	1.45×
Variance among populations	16.8%	24.3%	1.45×
Concordance in clustering	71.4%	71.4%	-
Mantel correlation (r)	-	0.523 (P = 0.002)	-
Approximate cost per sample (USD)	~15	~45	3.0×
Time for analysis (days)	~5	~10	2.0×

PI = probability of identity; Φst = fixation index.

Discussion

Genetic diversity assessment: Using both *matK* sequences (Hd = 0.685, π = 0.0076) and genome-wide SNPs (He = 0.251), the current study found moderate genetic diversity in *Ocimum basilicum L.* Our He value indicates more allelic variation in our sample collection than the mean genetic diversity (GD = 0.13) found in Ethiopian basil using DArTseq markers. In both marker systems (*matK*: Hd = 0.711; SNPs: He = 0.289), Population 2 consistently reflected the greatest variety, suggesting important genetic resources for breeding initiatives. A significant heterozygote deficit (Fis = 0.634, P < 0.001) was found by SNP analysis, and observed heterozygosity (Ho = 0.092) was significantly lower than predicted (He = 0.251). This result is consistent with recent genomic investigations that found farmed basil to have similarly low Ho (0.01–0.02) (Hedrick, 2005). The Wahlund effect from population substructure, selection pressures during cultivation, and inbreeding through vegetative propagation and restricted seed exchange are probably the causes of the higher Fis (Carović-Stanko et al., 2010). Instead of being a sampling artifact, this heterozygote loss is typical of populations of cultivated basil. The high inbreeding coefficient (Fis = 0.634) and significant heterozygosity deficiency (Ho < He) observed in the results of this study could indicate non-random mating patterns in the studied populations. However, this pattern may not exclusively reflect true biological inbreeding. It could also be influenced by the Wahlund effect due to population structure. As can be seen, this is supported by the STRUCTURE and AMOVA results. Furthermore, we cannot ignore and reject the possible contribution of clonal propagation, limited seed exchange and potential technical factors such as genotyping errors or the presence of paralogous loci in the GBS data. Therefore, the observed Fis values cannot be interpreted as direct evidence of inbreeding alone, but rather as the result of a combination of biological and methodological factors.

Population structure and differentiation: K = 3 was shown to be optimal (ΔK = 412.8) by Bayesian clustering, with a clear population assignment (86–90% membership). This is in contrast to K = 2 found in Ethiopian accessions, which probably reflects variations in breeding

history, geographic sampling, and the complicated polyploid structure of basil (Li et al., 2013). The very low genetic differentiation seen in Ethiopian germplasm was significantly exceeded by the moderate-to-high genetic differentiation indicated by pairwise Fst values (0.156–0.312). AMOVA confirmed significant structure by partitioning 24.3% of variation among populations ($\Phi st = 0.243$, $P < 0.001$). In Ethiopian basil, gene flow estimates ($Nm = 0.55–1.35$) were less than $Nm \approx 2.01$ and below the $Nm > 1$ threshold needed to prevent genetic drift (Hedrick, 2005). Geographic isolation and small seed exchange networks are reflected in this limited gene flow (Carović-Stanko et al., 2010). For maximum genetic variety and possible heterosis exploitation in breeding programs, the results suggest that conservation of all three populations may help preserve existing allelic diversity. Because the moderate-to-high differentiation shows that each population contains distinct allelic combinations (Sloan et al., 2018).

Comparative performance of markers: SNPs demonstrated vastly superior discriminatory power ($PI = 2.8 \times 10^{-26}$) compared to *matK* ($PI = 0.094$), representing a 3.4×10^{24}-fold difference. SNPs also indicated higher population differentiation ($\Phi st = 0.243$ vs 0.168), partitioning 1.45-fold more variance among populations. The substantially higher resolution of SNP markers reflects their genome-wide distribution and biparental inheritance compared with the single maternally inherited chloroplast locus capturing biparental inheritance and recombination (Hedrick, 2005). Despite differences, moderate concordance (71.4%; Mantel $r = 0.523$, $P = 0.002$) between marker systems is reasonable given the distinct evolutionary dynamics of chloroplast and nuclear genomes. Discordance arises from pollen-mediated gene flow affecting only nuclear markers, hybridization and polyploidy decoupling cytoplasmic and nuclear histories (Carović-Stanko et al., 2010), and differential selection pressures (Sloan et al., 2018). While *matK* provides cost-effective preliminary screening ($45/sample, 10 days), it is essential for high-resolution population genetics and breeding applications (Varshney et al., 2005).

Implications for breeding and conservation: The 6,234 SNPs provide a high-resolution toolkit for genomic selection (Jannink et al., 2010), marker-assisted backcrossing, and GWAS to identify QTLs for agronomic traits (Palumbo et al., 2019). The pronounced heterozygote deficit and moderate population differentiation suggest breeding strategies should: (1) design crosses between divergent populations to maximize heterosis; (2) implement controlled pollination to reduce inbreeding depression; (3) use genomic information to optimize genetic diversity in breeding populations (Woolliams et al., 2015). For conservation, the three distinct genetic clusters with restricted gene flow require separate preservation strategies: (1) establish representative core collections capturing maximum allelic diversity, (2) implement *ex situ* conservation through seed banks with periodic regeneration; (3) prioritize high-diversity individuals (particularly Population 2) for long-term preservation. A tiered strategy is advised: high-density SNP genotyping for populations of breeding or conservation concern comes after the initial *matK* screening for species verification (Schnable & Springer, 2013). Although the sample size was relatively limited, the combined use of chloroplast and genome-wide markers provided consistent patterns of population differentiation.

Conclusions: To investigate the genetic diversity and population structure of *Ocimum basilicum* L., this study combined chloroplast DNA barcoding using the *matK* gene with genome-

wide SNP markers. The results showed that there was moderate genetic diversity among the studied populations. Clear genetic differentiation was also obtained between the three geographical groups. For detecting population structure, genetic differentiation and heterozygosity deficiency, SNP-based analyses revealed much higher resolution than *matK* sequences. However, *matK* remained useful as a rapid and cost-effective tool for molecular identification and initial assessment of diversity. The significant lack of heterozygotes and moderate to high population differentiation could indicate limited gene flow and possible effects of inbreeding, population splitting and cultivation history. Population 2 could be considered as a valuable genetic resource for future breeding and conservation programs because it consistently showed the highest genetic diversity. All three Bayesian clustering analyses, PCA and AMOVA consistently supported the existence of three genetically distinct populations. These results emphasize the need to maintain genetically distinct basil populations to maintain their allelic diversity. It can be concluded that the identified SNP dataset provides a valuable genomic resource for marker-assisted selection, genomic improvement, and future association studies that can target agronomic and medicinal traits in *O. basilicum*. Future studies should use larger samples, conduct wider geographical sampling, and consider additional genomic approaches to further elucidate the evolutionary history and adaptive diversity of basil germplasm.

Limitations: Using STRUCTURE to analyze this dataset was accomplished with a limited sample size (n=30). This restriction on sample size could affect the overall ability to detect small demographic substructure in this dataset. Nonetheless, clusters were clearly evident from the analyses, with assignment probabilities between 86-90%, and strong support was detected for ΔK (K=3). Thus, these results should be viewed as providing valid results and showing clear evidence of a distinct cultivated population substructure, but they should be interpreted with caution. Future studies with larger sample sizes and/or expanding the range of sampling locations will help to more accurately assess the population structure of *O. basilicum*.

Author Contributions

Ola Hatif Hazim designed the study, conducted laboratory work, analyzed the data, and drafted the manuscript. Maysam Ihsan Ali Al-Hassany and Zinah Ameer Abbood supervised the research, contributed to data interpretation, and reviewed and edited the manuscript.

Data availability statement

The data contributing to the findings of this study are available from the investigating researcher upon request.

Acknowledgements

The authors would like to thank the Department of Biology, College of science, Mustansiriyah University.

Funding

No funding was received for this study.

Conflict of Interest

The authors declare no conflict of interest.

References

Abdelaziz, S. A., Khaled, K. A. M., Younis, R. A. A., Al-Kordy, M. A., El-Domyati, F. M., & Moghazee, M. M. (2024). Comparison of four DNA barcoding loci to distinguish between some Apiaceae family species. *Beni-Suef University Journal of Basic and Applied Sciences*, *13*, Article 12. https://doi.org/10.1186/s43088-023-00457-7

Amirteymoori, E., Khezri, A., Dayani, O., Mohammadabadi, M., Khorasani, S., Mousaie, A., & Kazemi-Bonchenari, M. (2021). Effects of linseed processing method (ground versus extruded) and dietary crude protein content on performance, digestibility, ruminal fermentation pattern, and rumen protozoa population in growing lambs. *Italian Journal of Animal Science*, *20*(1), 1506–1517. https://doi.org/10.1080/1828051X.2021.1984324

Arabpour, Z., Mohammadabadi, M., & Khezri, A. (2021). The expression pattern of p32 gene in femur, humeral muscle, back muscle and back fat tissues of Kermani lambs. *Agricultural Biotechnology Journal*, *13*(4), 183–200. https://doi.org/10.22103/jab.2022.18782.1371

Avetisyan, A., Markosian, A., Petrosyan, M., Sahakyan, N., Babayan, A., Aloyan, S., & Trchounian, A. (2017). Chemical composition and some biological activities of the essential oils from basil (*Ocimum*) different cultivars. *BMC Complementary and Alternative Medicine*, *17*(1), Article 60. https://doi.org/10.1186/s12906-017-1587-5

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Carović-Stanko, K., Liber, Z., Besendorfer, V., Javornik, B., Bohanec, B., Kolak, I., & Satovic, Z. (2010). Genetic relations among basil taxa (*Ocimum* L.) based on molecular markers, nuclear DNA content, and chromosome number. *Plant Systematics and Evolution*, *285*, 13–22. https://doi.org/10.1007/s00606-009-0251-z

CBOL Plant Working Group. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(31), 12794–12797. https://doi.org/10.1073/pnas.0905845106

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dray, S., & Dufour, A. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software*, *22*(4), 1–20. https://doi.org/10.18637/jss.v022.i04

Du, C., Xiong, T., Dong, Y., Gao, H., Zhang, Y., & Yue, J. (2025). Complete chloroplast genome structural and phylogenetic analysis of *Physostegia virginiana* (L.) Benth. 1930

(Lamiaceae). *Mitochondrial DNA Part B: Resources*, *10*(8), 736–741. https://doi.org/10.1080/23802359.2025.2528568

Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*, 359–361. https://doi.org/10.1007/s12686-011-9548-7

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, *14*(8), 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, *10*(3), 564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

Farahvashi, M., Mohammadabadi, M., Askari-Hesni, M., Ghanatsaman, Z. A., & Nanaee, H. A. (2026). Population structure of hawksbill turtles (*Eretmochelys imbricata*) nesting along the Persian Gulf coastline revealed by inter-simple sequence repeat (ISSR) markers. *Scientific Reports*, *16*(1), Article 4753. https://doi.org/10.1038/s41598-025-34749-y

Govindaraj, M., Vetriventhan, M., & Srinivasan, M. (2015). Importance of genetic diversity assessment in crop plants and its recent advances: An overview of its analytical perspectives. *Genetics Research International*, *2015*, Article 431487. https://doi.org/10.1155/2015/431487

Hedrick, P. W. (2005). A standardized genetic differentiation measure. *Evolution*, *59*(8), 1633–1638.

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, *23*(14), 1801–1806. https://doi.org/10.1093/bioinformatics/btm233

Jannink, J. L., Lorenz, A. J., & Iwata, H. (2010). Genomic selection in plant breeding: From theory to practice. *Briefings in Functional Genomics*, *9*(2), 166–177. https://doi.org/10.1093/bfgp/elq001

Kassem, M. A. (2025). Comparative analysis of chloroplast genomes across 20 plant species reveals evolutionary patterns in gene content, codon usage, and genome structure. *International Journal of Plant Biology*, *16*(3), Article 105. https://doi.org/10.3390/ijpb16030105

Kirankumar, S. I., Balaji, R., Tanuja, & Parani, M. (2023). The complete chloroplast genome of *Ocimum basilicum* L. var. *basilicum* (Lamiaceae) and its phylogenetic analysis. *Mitochondrial DNA Part B: Resources*, *8*(11), 1169–1173. https://doi.org/10.1080/23802359.2023.2275835

Kolde, R. (2019). *pheatmap: Pretty heatmaps* (Version 1.0.12) [R package]. https://CRAN.R-project.org/package=pheatmap

Kumar, A., Rodrigues, V., Saxena, A., Mishra, P., Shukla, A. K., Shasany, A. K., Nazar, N., Sgamma, T., Slater, A., & Sundaresan, V. (2025). Evaluation of the plastid and nuclear

DNA barcodes in genus *Ocimum* towards quality assurance in herbal industry. *Industrial Crops and Products*, *224*, Article 120399. https://doi.org/10.1016/j.indcrop.2024.120399

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

Letsiou, S., Madesis, P., Vasdekis, E., Montemurro, C., Grigoriou, M. E., Skavdis, G., Moussis, V., Koutelidakis, A. E., & Tzakos, A. G. (2024). DNA barcoding as a plant identification method. *Applied Sciences*, *14*(4), Article 1415. https://doi.org/10.3390/app14041415

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM* (arXiv:1303.3997). https://doi.org/10.48550/arXiv.1303.3997

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Palumbo, F., Qi, P., Pinto, V. B., Devos, K. M., & Barcaccia, G. (2019). Construction of the first SNP-based linkage map using genotyping-by-sequencing and mapping of the male-sterility gene in leaf chicory. *Frontiers in Plant Science*, *10*, Article 276. https://doi.org/10.3389/fpls.2019.00276

Patel, H. K., Fougat, R. S., Kumar, S., Mistry, J. G., & Kumar, M. (2015). Detection of genetic variation in *Ocimum* species using RAPD and ISSR markers. *3 Biotech*, *5*(5), 697–707. https://doi.org/10.1007/s13205-014-0269-y

Peakall, R., & Smouse, P. E. (2006). GENALEX 6: Genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, *6*(1), 288–295. https://doi.org/10.1111/j.1471-8286.2005.01155.x

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLOS ONE*, *7*(5), Article e37135. https://doi.org/10.1371/journal.pone.0037135

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

R Core Team. (2014). R: A language and environment for statistical computing. *MSOR Connections*, *1*.

Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*, *28*(21), 4737–4754. https://doi.org/10.1111/mec.15253

Rosenberg, N. A. (2004). DISTRUCT: A program for the graphical display of population structure. *Molecular Ecology Notes*, *4*(1), 137–138. https://doi.org/10.1046/j.1471-8286.2003.00566.x

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sánchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of

large data sets. *Molecular Biology and Evolution*, *34*(12), 3299–3302. https://doi.org/10.1093/molbev/msx248

Schnable, P. S., & Springer, N. M. (2013). Progress toward understanding heterosis in crop plants. *Annual Review of Plant Biology*, *64*, 71–88. https://doi.org/10.1146/annurev-arplant-042110-103827

Shah, P., Jain, N., Gawande, N., Sharma, T., Devanathan, K., Sankaranarayanan, S., & Balaji, R. (2026). Advancing plant DNA barcoding: Integrating chloroplast genome sequencing, cryptic diversity discovery and machine learning. *Molecular Biology Reports*, *53*(1), Article 550. https://doi.org/10.1007/s11033-026-11736-8

Sloan, D. B., Warren, J. M., Williams, A. M., Wu, Z., Abdel-Ghany, S. E., Chicco, A. J., & Havird, J. C. (2018). Cytonuclear integration and co-evolution. *Nature Reviews Genetics*, *19*(10), 635–648. https://doi.org/10.1038/s41576-018-0035-9

Varshney, R. K., Graner, A., & Sorrells, M. E. (2005). Genomics-assisted breeding for crop improvement. *Trends in Plant Science*, *10*(12), 621–630. https://doi.org/10.1016/j.tplants.2005.10.004

Waits, L. P., Luikart, G., & Taberlet, P. (2001). Estimating the probability of identity among genotypes in natural populations: Cautions and guidelines. *Molecular Ecology*, *10*(1), 249–256. https://doi.org/10.1046/j.1365-294x.2001.01185.x

Wickham, H. (2016). Data analysis. In *ggplot2: Elegant graphics for data analysis* (pp. 189–201). Springer. https://doi.org/10.1007/978-3-319-24277-4

Woolliams, J. A., Berg, P., Dagnachew, B. S., & Meuwissen, T. H. (2015). Genetic contributions and their optimization. *Journal of Animal Breeding and Genetics*, *132*(2), 89–99. https://doi.org/10.1111/jbg.12148

Zhakipbekov, K., Turgumbayeva, A., Akhelova, S., Bekmuratova, K., Blinova, O., Utegenova, G., Shertaeva, K., Sadykov, N., Tastambek, K., Saginbazarova, A., Urazgaliyev, K., Tulegenova, G., Zhalimova, Z., & Karasova, Z. (2024). Antimicrobial and other pharmacological properties of *Ocimum basilicum*, Lamiaceae. *Molecules*, *29*(2), Article 388. https://doi.org/10.3390/molecules29020388

مجله بیوتکنولوژی کشاورزی
شاپا چاپی: ۶۷۰۵-۲۲۲۸ شاپا الکترونیکی: ۶۵۰۰-۲۲۲۸

شناسایی مولکولی و تحلیل تنوع ژنتیکی *Ocimum basilicum L.* با استفاده از توالی‌های کلروپلاستی *matK* و نشانگرهای SNP در سطح کل ژنوم

اُلا هاتف حازم

* نویسنده مسئول. دانشکده علوم، دانشگاه المستنصریه، بغداد، عراق. ایمیل: ola.hataf@uomustansiriyah.edu.iq

میثم احسان علی الحسنی

دانشکده علوم، دانشگاه المستنصریه، بغداد، عراق. ایمیل: maysam.ihssan@uomustansiriyah.edu.iq

زینه امیر عبود

دانشکده علوم، دانشگاه المستنصریه، بغداد، عراق. ایمیل: Zina_ameer@uomustansiriyah.edu.iq

تاریخ دریافت: ۱۴۰۵/۰۱/۱۵ تاریخ دریافت فایل اصلاح شده نهایی: ۱۴۰۵/۰۳/۱۷ تاریخ پذیرش: ۱۴۰۵/۰۳/۱۸

چکیده

هدف: ریحان شیرین (*Ocimum basilicum* L.) یکی از گیاهان معطر و اقتصادی مهم است که برای بهینه‌سازی برنامه‌های اصلاح نژادی و حفاظت از ذخایر ژرم‌پلاسم، نیازمند مطالعات ژنتیکی جامع می‌باشد. هدف این پژوهش، بررسی تنوع ژنتیکی و ساختار جمعیتی ریحان‌های کشت‌شده با استفاده از ترکیب روش بارکدگذاری DNA کلروپلاستی (*matK*) و نشانگرهای تک‌نوکلئوتیدی گسترده در سراسر ژنوم (SNP) بود.

مواد و روش‌ها: در این مطالعه، ۳۰ نمونه ژنتیکی (اکسشن) مختلف از *O. basilicum* مورد بررسی قرار گرفتند (۱۰ اکسشن از هر یک از سه منطقه جغرافیایی متفاوت در بغداد، عراق). ژن *matK* از تمامی نمونه‌ها تکثیر و با استفاده از روش تعیین توالی سانگر (Sanger Sequencing) توالی‌یابی شد. همچنین، نشانگرهای SNP ژنومی با استفاده از روش تعیین ژنوتیپ مبتنی بر توالی‌یابی (GBS؛ Genotyping-by-Sequencing) تولید شدند. پس از کنترل کیفیت داده‌ها، تعداد ۶۲۳۴ نشانگر SNP با کیفیت بالا برای تحلیل‌های بعدی تأیید گردید. تنوع ژنتیکی، ساختار جمعیت و میزان تمایز ژنتیکی با استفاده از تحلیل هاپلوتیپی، نرم‌افزار

STRUCTURE، تحلیل مختصات اصلی (PCA)، تحلیل واریانس مولکولی (AMOVA) و شاخص‌های Fst مورد ارزیابی قرار گرفت.

نتایج: توالی‌یابی ژن *matK* وجود تنوع ژنتیکی متوسط را در قالب هشت هاپلوتیپ نشان داد ($Hd = 0.685 \pm 0.045$؛ $\pi = 0.0076 \pm 0.0012$). در مقابل، تحلیل SNPها کمبود قابل توجه هتروزیگوت‌ها را آشکار ساخت ($Ho = 0.092 \pm 0.038$؛ $He = 0.251 \pm 0.072$؛ $Fis = 0.634 \pm 0.089$؛ $P < 0.001$). خوشه‌بندی بیزی سه جمعیت ژنتیکی متمایز ($K = 3$) را شناسایی کرد. نتایج تحلیل AMOVA نشان داد که ۲۴/۳ درصد از کل تنوع ژنتیکی به تفاوت میان جمعیت‌ها مربوط می‌شود ($\Phi st = 0.243$؛ $P < 0.001$). به‌طور کلی، نشانگرهای SNP نسبت به ژن *matK* قدرت تفکیک بیشتری داشتند و همبستگی متوسطی میان روابط ژنتیکی حاصل از این دو روش مشاهده شد (Mantel r = 0.523؛ P = 0.002).

نتیجه‌گیری: نتایج این مطالعه نشان می‌دهد که سه جمعیت ژنتیکی متمایز با جریان ژنی محدود وجود دارند که برای حفظ تنوع ژنتیکی، نیازمند برنامه‌های حفاظتی مستقل هستند. کمبود چشمگیر هتروزیگوت‌ها بیانگر آن است که تلاقی بین جمعیت‌های ژنتیکی متمایز می‌تواند موجب افزایش هتروزیس (برتری دورگه‌ای) در برنامه‌های اصلاح نباتات شود. یافته‌های این پژوهش بینش‌های جدیدی درباره ساختار ژنتیکی ریحان کشت‌شده فراهم می‌کند؛ از جمله وجود سه خوشه ژنتیکی مشخص، کمبود ساختارمند هتروزیگوت‌ها و همخوانی نسبی میان ژنوم کلروپلاستی و ژنوم هسته‌ای. این نتایج می‌توانند در درک بهتر جریان ژنی، اثرات اهلی‌سازی، تمایز جمعیت‌ها و تدوین راهبردهای آینده برای اصلاح و حفاظت ژنتیکی ریحان مؤثر باشند.

کلمات کلیدی: بارکدگذاری DNA، ژنتیک جمعیت، ژنوتیپ‌سازی مبتنی بر SNP ژنومیک حفاظتی، کمبود هتروزیگوت

نوع مقاله: پژوهشی

استناد: اُلا هاتف حازم، میثم احسان علی الحسنی، زینه امیر عبود (۱۴۰۵). شناسایی مولکولی و تحلیل تنوع ژنتیکی *Ocimum basilicum* L. با استفاده از توالی‌های کلروپلاستی *matK* و نشانگرهای SNP در سطح کل ژنوم. *مجله بیوتکنولوژی کشاورزی*، ۱۸(۴)، ۲۷۶-۲۵۹.

Publisher: Shahid Bahonar University of Kerman & Iranian Biotechnology Society.