Research Square

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

Genetic Diversity and Population Structure of *Madhuca longifolia* in Tamil Nadu, India Revealed by SSR Markers

D. Thangamani
Institute of Forest Genetics and Tree Breeding
Raju Kannan
Krishi Vigyan Kendra (KVK)
Fatima Shirin
Tropical Forest Research Institute
R. Manikandan
Botanical Survey of India
Gayathri Prabhakaran
Tamil Nadu Agricultural University
Venkatesan Sathiya Aravindan
sathiyajournal@gmail.com

Annamalai University

Research Article

Keywords: Conservation, Mahua, Polymorphism, Precision breeding, Simple Sequence Repeat

Posted Date: September 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10713205/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Madhuca longifolia a multipurpose tree species of significant ecological and socioeconomic importance, remains underexplored at the molecular level despite its wide distribution and utility. The present study aimed to assess the genetic diversity and population structure of *M. longifolia* accessions collected from Tamil Nadu, India and to evaluate the effectiveness of simple sequence repeat (SSR) markers for molecular characterization and precision breeding applications. A total of 56 accessions representing three geographic populations (South, Central and North) were genotyped using 15 polymorphic SSR markers developed through transcriptome analysis. The markers exhibited high polymorphism, with polymorphic information content (PIC) values ranging from 0.608 to 0.666, indicating strong discriminatory power. Genetic diversity parameters revealed substantial allelic richness across populations, with mean expected heterozygosity (He = 0.657) and Shannon's information index (I = 1.080). Population-level analysis indicated moderate genetic differentiation, with Nei's genetic distance ranging from 0.553 to 1.319 and the highest divergence observed between Central and North populations. Analysis of molecular variance (AMOVA) demonstrated that 75% of the total genetic variation resided within individuals, while only 18% was attributed to differences among populations, suggesting extensive gene flow and outcrossing behavior. Cluster analysis and dendrogram construction grouped the accessions into distinct genetic clusters corresponding to geographic origin, with certain clusters exhibiting high genetic divergence, highlighting their potential for heterosis breeding. The identification of highly informative loci such as MLO24 and MLO33 further strengthens their applicability in molecular fingerprinting and germplasm management. Overall, the study confirms that *M. longifolia* populations in Tamil Nadu possess high genetic diversity and a structured yet interconnected genetic architecture. These findings provide a robust foundation for conservation strategies, marker-assisted selection and the development of precision breeding programs aimed at improving economically important traits.

INTRODUCTION

Mahua is a hardy deciduous forest tree thriving in dry, rocky, saline soils across much of India, Nepal and Sri Lanka, providing food, fodder, fuel and key livelihood support for forest-dwelling communities. Flowers are eaten, fermented into liquor and incorporated into sweets and baked products; fruits are used as vegetables; seeds yield "mowrah butter" used as cooking oil, lamp fuel and raw material for soaps, detergents and lubricants [1]. *In vitro* and *in vivo* work reports antioxidant, anti-inflammatory, analgesic, antipyretic, antiulcer, hepatoprotective, antihyperglycemic/antidiabetic, anticancer/antitumor, antibacterial, anthelmintic, neuropharmacological and wound-healing activities from different parts of the plant. Flowers, seeds, bark and leaves contain sapogenins, triterpenoids, steroids, saponins, flavonoids, glycosides, alkaloids, tannins and phenolics. Seed fat "mowrah butter" is rich in oleic acid and other unsaturated fatty acids; seeds and seed cake provide minerals such as calcium, phosphorus, magnesium and iron. Flowers supply sugars, vitamins (A, C, B-complex) and minerals, supporting both nutritional and medicinal uses [2]. Phenotypes in long-lived trees are strongly influenced by environment and age, so selection on observable traits alone can be slow and may not reflect true genetic variation.

Genetic diversity is vital for conserving and improving plant resources, as it enables adaptation and supports the selection of superior genotypes. In long-lived tree species like *M. longifolia*, conventional phenotypic selection is slow, environment-dependent and inadequate for capturing true genetic variation [3]. Molecular marker technologies offer a reliable and environment-independent approach for assessing genetic variation. Among these, simple sequence repeat (SSR) markers are widely preferred due to their high polymorphism, codominant inheritance, reproducibility, and genome-wide distribution [4]. Despite the ecological and socioeconomic importance of *M. longifolia* comprehensive molecular characterization of its genetic resources remains limited, particularly at a regional scale in Southern India. Existing studies have largely focused on phenotypic evaluation or fragmented molecular analyses, with minimal integration of population-level diversity, genetic structure and breeding relevance. In addition, the use of simple sequence repeat (SSR) markers in this species remains underexplored, limiting the availability of robust tools for high-resolution genotyping. The present study lies in the development and application of highly informative SSR markers to assess genetic diversity across accessions collected from diverse agro-climatic regions of Tamil Nadu, India. Among the available marker systems, simple sequence repeat (SSR) markers are particularly advantageous for this species due to their high polymorphism, codominant inheritance, reproducibility and wide genome coverage, which collectively allow accurate detection of heterozygosity, allelic variation and population structure. In contrast to dominant markers such as RAPD and AFLP, SSRs enable discrimination between homozygous and heterozygous loci, while offering greater locus specificity and reproducibility than random marker systems and being more cost-effective and technically accessible than high-throughput SNP genotyping for non-model tree species with limited genomic resources. These attributes make SSRs especially suitable for population-level studies, germplasm characterization and the development of reliable molecular fingerprints in underutilized forest trees like *M. longifolia.* By integrating population genetic parameters, molecular variance (AMOVA) and cluster-based analyses, this study provides a comprehensive understanding of genetic structure and diversity distribution. Accordingly, the study aims to evaluate genetic diversity using polymorphic SSR markers, elucidate population structure and genetic relationships among geographically distinct accessions, quantify gene flow and partition genetic variation and identify highly informative loci for molecular fingerprinting, conservation planning, and marker-assisted breeding.

MATERIALS AND METHODS

Plant Material and Study Area

A total of 56 *Madhuca longifolia* (J. Koenig ex L.) J.F. Macbr. accessions, comprising both varieties *M. longifolia* var. *longifolia* and *M. longifolia* var. *latifolia* were collected from three geographically distinct populations across Tamil Nadu, India, representing the South (19 accessions; MLS1–19), Central (19 accessions; MLC20–38) and North (18 accessions; MLN39–56) regions (Table 1 & Table 2). Sampling encompassed a broad spectrum of agro-climatic zones to maximise geographic and ecological representation. Young, healthy leaf tissue was harvested from each accession, immediately snap-frozen

in liquid nitrogen in the field and transported to the laboratory under cryogenic conditions. Voucher specimens were deposited at the Institute of Forest Genetics and Tree Breeding (IFGTB), Coimbatore, India. All field collections were made during 2024–2025 in compliance with institutional guidelines for the collection of forest genetic resources [4].

Table 1
Populations and their geographic locations

Population	Locations
South	Theni
	Virudhunagar
	Sathur
	Gudalur
	Nagercoil
	Sundarapuri
	T. Kunnathur
	Ottanchatthiram
	Sivakiripatti
	Madathukulam
	Thalayam–Sithiraikulam
Central	Palani
	Thiruchuli
	Aruppukottai
	Palavanatham
	Thanakankulam
	S. Puliyankulam
	Iluppayur
	Paaluthu
	Kanavaypatti
	Rajagopalanpatti
	Arumalaikathaarppanar Kovil
	Angammal Kovil
	Peraiyur road–Usilampatti
	Kariyanampatti
North	Palamalai
	Kulathupalayam
	Alanthurai

Population	Locations
	Velliyangiri Temple
	Karunya
	Sundarapuri
	T. Kunnathur
	Thirumangalam
	Poyyapatti
	S. Agaram
	Thurinchapuram
	Valathiyamaram
	K.K. Pudur
	Suppeygoundampudur
	Periyapothu
	Ottanchatthiram

Table 2
Sample distribution used for molecular analysis

Sl. No.	Population	Code	Individuals
1	South	MLS1–19	19
2	Central	MLC20–38	19
3	North	MLN39–56	18
	Total		56

Field collections of *Madhuca longifolia* were conducted during 2024–2025 under institutional guidelines for forest genetic resource collection. Necessary permissions were obtained from the Principal Chief Conservator of Forests, Government of Tamil Nadu, vide Ref. No. WL5(A)579/2021, Permit No. 28/2021, valid until June 2026, covering Anamalai Tiger Reserve and other designated forest divisions. The plant specimens were taxonomically identified by plant taxonomist. Voucher specimens were deposited in the Fischer Herbarium (FRC), Institute of Forest Genetics and Tree Breeding, Coimbatore under Accession No. 123656. The voucher specimen of *Madhuca longifolia* collected from Anamalai Tiger Reserve is publicly available at IFGTB Herbarium for reference.

Genomic DNA Isolation

Genomic DNA was extracted from ~ 100 mg of fresh leaf tissue per accession using a modified CTAB (cetyltrimethylammonium bromide) protocol. Leaf tissue was ground to a fine powder under liquid nitrogen using a pre-chilled mortar and pestle and transferred to a 2 ml microcentrifuge tube. 1 ml of preheated (65°C) CTAB extraction buffer [2% CTAB, 100 mM Tris-HCl (pH 8.0), 1.4 M NaCl, 20 mM EDTA, 1% polyvinylpyrrolidone (PVP-40), 0.2% β-mercaptoethanol] was added and the mixture was incubated at 65°C for 60 min with periodic inversion. The lysate was extracted twice with an equal volume of chloroform: isoamyl alcohol (24:1, v/v) and the aqueous phase recovered by centrifugation at 12,000 × g for 10 min. DNA was precipitated with 0.6 volumes of ice-cold isopropanol, pelleted, washed twice with 70% ethanol, air-dried, and resuspended in 100 µl TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0). DNA quantity and purity were assessed using a NanoDrop ND-1000 spectrophotometer (Thermo Scientific, USA) at A260/A280 and A260/A230 ratios; DNA integrity was confirmed by electrophoresis on a 0.8% agarose gel stained with ethidium bromide. Working stocks were diluted to 50 ng/µl in TE buffer and stored at − 20°C until use.

SSR Marker Development and PCR Amplification

A total of 15 polymorphic SSR markers comprising 11 EST-SSR loci (MLO series) derived from expressed sequence data and 4 genomic SSR loci (MLA series) were developed from transcriptome sequence data of *M. longifolia* using MISA (MIcroSAtellite identification tool) for repeat identification and Primer3 (v.2.6.0) for primer design, targeting di- and trinucleotide repeat motifs with a minimum of six repeat units. Primer sequences, melting temperatures (Tm) and expected amplicon sizes are listed in (Table 3). Prior to full-panel genotyping, all 15 markers were validated on a representative subset of accessions to confirm consistent amplification and polymorphism. PCR amplifications were conducted in a total volume of 10 µl containing: 50 ng template DNA, 1× PCR Master Mix (KAPA2G Robust, Sigma-Aldrich) and 0.5 µM each of forward and reverse primer. Thermal cycling was performed on a Bio-Rad T100 cycler using the following programme: initial denaturation at 95°C for 5 min; 35 cycles of denaturation at 94°C for 45 s, locus-specific annealing (52–58°C, (Table 3) for 45 s, and extension at 72°C for 45 s; followed by a final extension at 72°C for 10 min and a 4°C hold. Annealing temperatures were individually optimised per primer pair using a gradient PCR approach. Amplicons were resolved on 3% MetaPhor agarose gels (Lonza, USA) in 0.5× TBE at 6 V/cm for 2–3 h, stained with ethidium bromide (0.5 µg/ml), and visualised under UV transillumination using a Bio-Rad Gel Doc XR+ system. Allele sizes were estimated against a 100 bp DNA ladder (New England Biolabs) and scored independently by two researchers to minimise scoring errors (Fig. 3 & Fig. 4).

Table 3
SSR primers used for amplification

Primer	Tm (°C)	Amplicon size (bp)
MLO5	54	203–220
MLO10	54	209–230
MLO12	54	181–230
MLO16	56	205–233
MLO19	56	193–230
MLO24	56	191–202
MLO33	56	208–210
MLO38	56	200–241
MLO50	54	189–193
MLO63	55	200
MLO82	58	200–225
MLA2	52	208–215
MLA11	55	204–291
MLA35	53	184–239
MLA70	55	194–207

Data Analysis

SSR alleles were scored as a binary matrix (presence = 1, absence = 0) for all 56 accessions across 15 loci. Population genetic parameters including number of alleles per locus (Na), effective number of alleles (Ne), observed heterozygosity (Ho), expected heterozygosity (He), unbiased expected heterozygosity (uHe), Shannon's Information Index (I) and the fixation index (F) were computed at both locus and population levels using GenAlEx v.6.5 [5]. Gene flow (Nm) was estimated as

Nm = *0.25(1 − FST)/FST.*

Where

Nm = Number of migrants per generation (gene flow)

FST = Fixation index (a measure of genetic differentiation among populations)

Polymorphism information content (PIC) was calculated using the formula by [6]

PIC = $1 - \Sigma pi^2$,

where pi is the frequency of the ith allele at each locus. Pairwise Nei's unbiased genetic distance (D) and genetic identity (I) between populations were estimated in POPGENE v.1.32 [7, 8]. Analysis of Molecular Variance (AMOVA) was performed to partition total genetic variance into three hierarchical components among populations, among individuals within populations, and within individuals, implemented in GenAlEx v.6.5 [9]. The significance of each variance component was evaluated by 999 random permutations. A Neighbour-Joining (NJ) dendrogram was constructed from the matrix of pairwise Nei's genetic distances using MEGA v.11.0 [10, 11], with bootstrap confidence values estimated from 1,000 replicates. Principal Coordinate Analysis (PCoA) and cluster assignment plots were generated in GenAlEx v.6.5 to visualise genetic relationships and population structure among the 56 accessions. (Fig. 2)

RESULTS AND DISCUSSION

A total of 56 *Madhuca longifolia* accessions were systematically collected from three geographically distinct populations spanning the South, Central and North regions of Tamil Nadu, and this broad sampling across diverse agro-climatic conditions and multiple districts within each region minimized sampling bias and strengthened the reliability of the downstream population genetic analyses; the sampled germplasm was structured into three populations using distinct coding systems (MLS, MLC, MLN), comprising 19 accessions each from the South and Central regions and 18 from the North, a near-balanced design that ensured comparability across populations, while the inclusion of both *M. longifolia* var. *longifolia* and var. *latifolia* further enhanced the genetic breadth captured in the dataset. Against this sampling framework, the 15 SSR markers employed comprising both EST-derived (MLO series) and genomic (MLA series) loci exhibited consistent and robust amplification across all accessions, with the observed amplicon size range (181–291 bp) closely matching expected product sizes and confirming primer specificity, while the absence of null alleles or amplification failure across loci underscored the reliability of the marker set, and the diversity in annealing temperatures and fragment lengths reflected locus-specific variability contributing to the overall discriminatory power of the system.

Building on this robust marker performance, genetic diversity estimates revealed consistently high levels of variation across all populations with only subtle regional differentiation: Shannon’s Information Index (I) values were uniformly high, ranging from 1.050 in the South to 1.096 in the Central and North populations, indicating substantial allelic richness, while expected heterozygosity (He) showed minimal variation (0.641–0.665), suggesting a relatively stable genetic structure across regions; observed heterozygosity (Ho), however, displayed greater variability, particularly in the South population (0.579), indicating reduced heterozygote frequency relative to expectations a pattern reflected in the fixation index (F), where a positive value in the South (0.096) suggested mild inbreeding or non-random mating, whereas negative values in the Central (–0.136) and North (–0.102) populations indicated heterozygote excess likely driven by outcrossing and gene flow, collectively pointing to a predominantly outcrossing system with localized deviations in mating dynamics (Table 4). At the locus level, this same pattern of

high polymorphism held: gene diversity values ranged from 0.952 (MLO63) to 1.097 (MLO33 and MLA11), suggesting that most loci contribute substantially to overall genetic variation, and Polymorphism Information Content (PIC) values were consistently high (0.608–0.666), classifying all markers as highly informative, while gene flow (Nm) estimates varied considerably among loci, from 0.757 (MLO12) to 2.801 (MLO24), with an overall mean of 1.344, indicating moderate to high levels of allelic exchange; loci such as MLO24 (high Nm) and MLO33 (high PIC and diversity) emerged as particularly informative for assessing population connectivity, whereas MLO63 exhibited comparatively lower diversity, suggesting a more conserved genomic region (Table 5).

Table 4
Population-wise genetic diversity parameters

Population	**I**	**Ho**	**He**	**uHe**	**F**
South	1.050	0.579	0.641	0.660	0.096
Central	1.096	0.755	0.665	0.684	-0.136
North	1.096	0.732	0.665	0.685	-0.102
Mean	1.080	0.689	0.657	0.676	-0.047

Population-wise genetic diversity indices. I = Shannon's Index; Ho = observed heterozygosity; He = expected heterozygosity; uHe = unbiased expected heterozygosity; F = fixation index.

Table 5
Locus-wise genetic diversity parameters

Locus	Nm	Gene Diversity	Heterozygosity	PIC
MLO5	1.354	1.070	0.647	0.647
MLO10	1.375	1.078	0.652	0.652
MLO12	0.757	1.078	0.653	0.653
MLO16	1.308	1.096	0.665	0.665
MLO19	1.295	1.094	0.664	0.664
MLO24	2.801	1.092	0.662	0.662
MLO33	1.347	1.097	0.666	0.666
MLO38	0.978	1.095	0.664	0.664
MLO50	1.643	1.093	0.663	0.663
MLO63	0.789	0.952	0.608	0.608
MLO82	1.572	1.092	0.662	0.662
MLA2	1.702	1.093	0.660	0.660
MLA11	0.921	1.097	0.662	0.662
MLA35	1.147	1.093	0.664	0.664
MLA70	1.171	1.087	0.659	0.659

Turning from marker- to population-level relationships, pairwise Nei's genetic distance analysis revealed clear differentiation among populations, with values ranging from 0.553 to 1.319, the lowest distance between the South and Central populations indicating closer genetic affinity and the highest divergence between the Central and North populations suggesting restricted gene flow and possible geographic or ecological isolation (Table 6), a pattern mirrored in genetic identity, which ranged from 0.268 to 0.575, with the highest identity between the South and Central populations indicating shared genetic background and the low identity between the Central and North populations confirming substantial divergence (Table 7). Finally, the Analysis of Molecular Variance (AMOVA) situated these pairwise patterns within a broader hierarchical context, revealing that the majority of genetic variation was partitioned within individuals (75%), indicating high heterozygosity at the individual level, while variation among populations accounted for 18%, suggesting moderate genetic structuring, and only 6% of variation was attributed to differences among individuals within populations a hierarchical distribution that highlights the dominance of intrapopulation genetic variability and supports the presence of substantial gene flow and outcrossing behaviour (Table 8 & Fig. 1)

Table 6
Nei's genetic distance matrix

Population	South	Central	North
South	0.000		
Central	0.553	0.000	
North	0.782	1.319	0.000

Table 7
Nei's genetic identity matrix

Population	South	Central	North
South	1.000		
Central	0.575	1.000	
North	0.458	0.268	1.000

Table 8
Analysis of Variance Partitioning in the Study

Source	df	SS	MS	Est. Var.	%
Among populations	**2**	**100.699**	**50.349**	**1.195**	**18%**
Among individuals	**53**	**304.453**	**5.744**	**0.421**	**6%**
Within individuals	**56**	**274.500**	**4.902**	**4.902**	**75%**
Total	**111**	**679.652**		**6.518**	**100%**

df (Degrees of Freedom): Number of independent values used in variance estimation.

SS (Sum of Squares): Total variation attributed to each source.

MS (Mean Square): Average variation (SS ÷ df).

Est. Var. (Estimated Variance): Contribution of each source to overall variance.

% (Percentage): Proportion of total variance explained by each source.

Taken together, these results are consistent with, and are further illuminated by, comparisons with other tropical and forest tree species. The 15 SSR markers, comprising 11 EST-derived (MLO series) and 4 genomic (MLA series) loci, exhibited uniformly high polymorphism across all 56 accessions, as evidenced by PIC values ranging from 0.608 (MLO63) to 0.666 (MLO33); according to the widely adopted classification of [5], PIC values exceeding 0.50 are considered highly informative, and all 15 loci surpassed this threshold, confirming the suitability of the marker panel for high-resolution population

differentiation and germplasm characterization. These values are notably high compared with those reported for several other multipurpose forest tree species for example, [20] reported a PIC range of 0.15–0.81 (mean 0.56) using 16 SSR markers in *Quercus griffithii*, while [33] recorded a mean PIC of 0.6125 with 20 EST-SSR markers in *Castanopsis carlesii*, values marginally lower than those obtained here and reinforcing the discriminatory power of the developed MLO and MLA loci [12]. The superior PIC performance of loci such as MLO33 (PIC = 0.666) and MLO24 (PIC = 0.662, Nm = 2.801) suggests that transcriptome-derived EST-SSRs are particularly effective in capturing functional genomic variation in *M. longifolia*, consistent with [13], who found that EST-SSR markers developed from transcriptome sequences of *Larix principis-rupprechtii* achieved a mean PIC of 0.424, and with [14], who reported a mean PIC of 0.52 in *Cinnamomum parthenoxylon*, both substantially lower than the present study; the comparatively elevated PIC values in *M. longifolia* may reflect the species' inherently high outcrossing rate, which promotes allelic diversity at functional loci by preventing fixation of common alleles, while MLO63, showing the lowest gene diversity (0.952) and PIC (0.608), may reside within a more conserved genomic region subject to stabilizing selection, a pattern previously observed in EST-SSR loci located within or near coding sequences under purifying selection [15].

This same signature of high diversity extends to the population level: the overall genetic diversity observed in *M. longifolia* populations across Tamil Nadu was high, with a mean expected heterozygosity (He = 0.657) and Shannon's information index (I = 1.080), indicative of substantial allelic richness maintained across geographically distinct agro-climatic zones values considerably higher than those recorded for ecologically comparable tree species [16] reported a mean He of 0.680 for *Ailanthus altissima* using EST-SSRs across 34 populations, comparable to the present findings, while [17] documented He = 0.670 in *Juglans mandshurica* across 15 populations, with Shannon indices (I = 1.455) markedly higher, likely reflecting a greater marker set (15 EST-SSRs) and wider geographic sampling. The He value for *M. longifolia* (0.657) is also comparable to that reported for the endangered tree *Boswellia papyrifera* (He = 0.664–0.724), a species known for its predominantly outcrossing mating system [18], further supporting the notion that *M. longifolia* is predominantly allogamous. High within-population genetic diversity is, in turn, an ecologically and evolutionarily favourable condition for long-lived, outcrossing forest trees, as it confers greater adaptive potential under shifting environmental conditions [19], and the consistently high He values recorded across all three populations (South: 0.641; Central: 0.665; North: 0.665) suggest that no significant erosion of genetic variation has occurred at the population level a pattern attributable to the historically broad distribution of *M. longifolia* across Tamil Nadu, its cross-pollinating reproductive strategy, and the wide diversity of agro-climatic environments sampled here, in contrast to the significantly lower He values reported for fragmented or range-restricted populations of other forest tree species under heavy anthropogenic pressure [20], which underscores the relatively favourable conservation status of *M. longifolia* in this region.

The hierarchical structuring revealed by AMOVA reinforces this picture: 75% of the total genetic variation was partitioned within individuals, 6% among individuals within populations and 18% among populations, a pattern characteristic of predominantly outcrossing tree species and widely reported in comparable studies of [21] found that 77% of genetic variance in oil-tea camellia species resided within populations

using SSR markers, [22] documented 93% within-population variation in *Robinia pseudoacacia*, and [23] reported 92.54% within-individual variation in *Guadua angustifolia* using RAD-seq, consistent with strong outcrossing tendencies. The 18% among-population component recorded here indicates moderate genetic structuring, notably higher than the 5% reported in some conifers [24] but lower than the 41% documented in *Coffea canephora*, where gene flow is more restricted [25], suggesting that some degree of geographic and ecological isolation exists among the South, Central and North populations of Tamil Nadu, tempered by significant gene flow consistent with the mean Nm = 1.344 estimated in this study. This dominant within-individual variation signal is, moreover, consistent with the codominant SSR marker system used, which is sensitive to heterozygosity levels at individual loci, and the finding that individual-level heterozygosity is the primary repository of genetic variation in this species parallels results from other tropical outcrossing species such as *Parkia multijuga* [26] and *Acrocomia aculeata* [27], where individual-level outcrossing rates approached or exceeded 0.98 together with the high mean He, these AMOVA results confirm that *M. longifolia* maintains its genetic diversity primarily through active cross-pollination rather than vegetative propagation or selfing, a finding with direct relevance to seed-source strategies for conservation and afforestation programmes.

The gene flow estimates lend further support to this interpretation of the species as predominantly outcrossing. The mean gene flow estimate (Nm = 1.344) indicates moderate allelic exchange among the three populations, and values of Nm > 1 are generally considered sufficient to counteract the effects of genetic drift and prevent significant differentiation; the locus-specific Nm ranged from 0.757 (MLO12) to 2.801 (MLO24), indicating heterogeneity in gene flow across the genome, with MLO24 recording the highest Nm value (2.801), suggesting that the genomic region it represents may be under positive selection or located within a region of elevated recombination that facilitates allele exchange, a pattern consistent with reports from functionally annotated EST-SSR loci in tropical species [28], while loci such as MLO12 (Nm = 0.757) and MLO63 (Nm = 0.789) reflect localized differentiation, possibly driven by local adaptation or selective sweeps at those genomic positions. The moderate overall Nm observed is ecologically consistent with the species' biology: *M. longifolia* is pollinated by nocturnal insects and bats, and its seeds are dispersed by a variety of frugivorous animals and gravity across relatively long distances, facilitating inter-population gene flow [4] nonetheless, the geographic barriers between agro-climatic zones, particularly between the Central and North populations (lowest genetic identity: 0.268; highest Nei's distance: 1.319), may have partially restricted allelic exchange between these two groups, a pattern that parallels the findings of [10], who demonstrated in a meta-analysis of seasonally dry tropical forest species that Nm and spatial genetic structure are strongly mediated by pollination syndrome, successional stage and geographic distance, while the comparatively higher gene flow between the South and Central populations (genetic identity = 0.575; Nei's distance = 0.553) may be attributable to their greater geographic proximity, shared soil and rainfall profiles across southern Tamil Nadu, and overlapping pollinator and disperser communities.

Set against this backdrop of gene flow, the population-level fixation indices (F) revealed contrasting mating dynamics across the three geographic populations: the South population exhibited a positive F value (+ 0.096), indicative of a mild deficiency of heterozygotes relative to Hardy–Weinberg expectations

and consistent with localised inbreeding or a Wahlund effect arising from cryptic population substructure, whereas the Central (F = − 0.136) and North (F = − 0.102) populations showed heterozygote excess, reflecting predominantly outcrossing mating behaviour characteristic of many tropical tree species [15, 26]; the negative fixation indices in the Central and North populations may further indicate selection against homozygotes during early life stages, a phenomenon documented in *Swietenia macrophylla* [29] and *Campomanesia xanthocarpa* [30], where post-zygotic selection systematically eliminates inbred individuals during seedling establishment. The mild positive F observed in the South population may, in turn, reflect a combination of reduced effective population size and historically restricted pollen dispersal in that agro-climatic zone, consistent with the lower observed heterozygosity (Ho = 0.579) and comparatively lower Shannon index (I = 1.050) recorded there; nonetheless, the mean fixation index across all populations was negative (F = − 0.047), collectively confirming that *M. longifolia* in Tamil Nadu is predominantly allogamous, congruent with the species' protandrous floral development, its reliance on cross-pollination by insects and bats, and the self-incompatibility mechanisms reported in related members of the Sapotaceae family mixed mating signals that warrant consideration in designing seed orchard management strategies, particularly for the South population, where recruitment from a broader pool of pollen donors would be advisable to prevent genetic erosion.

These fixation and gene-flow patterns are further reflected in the pairwise genetic-distance data: values ranging from 0.553 (South vs. Central) to 1.319 (Central vs. North) reveal a clear geographic gradient of genetic divergence, with proximity in agro-climatic zone broadly paralleling genetic affinity, and the high divergence between the Central and North populations (genetic identity = 0.268) is particularly noteworthy, as it suggests significant differentiation despite apparent geographic continuity, possibly reflecting past historical barriers, differential land-use intensities, or distinct ecological selective pressures operating within those zones across generations. Comparable population-level divergences have been reported in *Quercus griffithii* [20], where FST = 0.06 and Nm = 3.91, indicating lower differentiation likely as a consequence of greater connectivity in its northeastern Indian range, and the genetic-distance values observed here are broadly consistent with those recorded in *Juglans mandshurica* (0.190–0.357, Fst = 0.109), a species of comparable ecological niche [22]. These distance and identity values carry direct implications for precision breeding: the high Nei's genetic distance between the Central and North populations (D = 1.319) indicates substantial heterotic potential if accessions from these two groups are used as parents in controlled hybridization programmes and [31], in a comprehensive review of the molecular mechanisms of heterosis in tree species, demonstrated that genetic divergence between parental lines, combined with marker-assisted selection of divergent genotypes, consistently predicts greater hybrid vigour for growth-related traits, while [32] established in *Fraxinus* spp. that Nei's genetic distance showed significant linear correlations with better-parent heterosis for volume index ($r^2 > 0.80$), providing a quantifiable genetic-based parent-selection method; the identification of the Central and North populations as genetically most divergent therefore provides a strong foundation for structured crossing designs in future *M. longifolia* hybrid-breeding programmes targeting improved seed-oil yield, fruit characteristics and drought tolerance.

Consistent with this pattern of geographic structuring, the Neighbour-Joining (NJ) dendrogram constructed from Nei's genetic distances and the PCoA ordination collectively demonstrated geographic correspondence with genetic clustering, wherein accessions from the same regional population largely grouped together while certain outlier accessions bridged cluster boundaries, indicating genetic admixture and historical connectivity a pattern consistent with the findings of [20] in *Castanopsis hystrix*, where 96% of total variation resided within populations, yet three genetically distinct sub-populations were identified. The geographic alignment of genetic clusters in *M. longifolia* supports the hypothesis that natural distribution and agro-climatic gradients in Tamil Nadu, from the southern coastal plains to the northern transition zones bordering the Coimbatore plateau, serve as primary drivers of genetic structure. Taken as a whole, this comprehensive characterization of genetic diversity and population structure provides a robust, evidence-based framework for the conservation and genetic improvement of *M. longifolia* germplasm in Tamil Nadu: given the high within-individual heterozygosity and the moderate interpopulation differentiation documented, *ex situ* germplasm conservation efforts should priorities the representation of all three geographic populations to capture the full allelic breadth identified in this study, with sampling strategies including a minimum of 25–30 individuals per population and particularly targeting genetically divergent accessions identified as outliers in the NJ dendrogram, to ensure adequate coverage of rare alleles and population-specific diversity; *in situ* conservation units should likewise be established in districts representing the agro-climatic extremes of each population (for example, Nagercoil for the South, Palani for the Central, and Palamalai for the North), complemented by targeted replanting using representative accessions to reinforce natural regeneration under field conditions. Finally, for precision-breeding applications, the two loci identified as most informative MLO24 (Nm = 2.801, PIC = 0.662) and MLO33 (PIC = 0.666, gene diversity = 1.097) are recommended for inclusion in any reduced marker panel for molecular fingerprinting, variety identification and parentage assignment, since highly informative SSR loci have been shown to serve effectively as anchors in marker-assisted selection (MAS) programmes in non-model forest tree species [13, 22] crucially, the combination of high PIC and differential gene flow at MLO24 makes it a particularly powerful locus for detecting recent migration events and assigning parentage in open-pollinated families, applications directly relevant to seed-orchard management and provenance trials, while the validated 15-locus SSR panel constitutes a portable, cost-effective genotyping platform suitable for ongoing germplasm monitoring, genetic certification of elite clones and quality control in vegetative propagation programmes across southern India.

CONCLUSION

Looking ahead, integrating the population genetic baseline generated herein with next-generation sequencing approaches including genotyping-by-sequencing (GBS) or whole-genome resequencing would enable the identification of genome-wide SNP markers, facilitate QTL mapping for economically important traits (seed oil content, mowrah butter composition, flower biomass) and allow selective sweep analyses to identify loci under local adaptation. The transition from SSR-based marker-assisted selection to genomic selection (GS) has already been demonstrated in tree breeding contexts and the

present study provides the essential population structure and relatedness matrix required for training genomic prediction models in *M. longifolia*. Coupling the genetic diversity data presented here with phenotypic evaluations of yield-related and bioprospecting traits across the 56 accessions would further enable genome-wide association studies (GWAS), accelerating the development of precision-bred varieties optimized for both industrial and agroforestry deployment.

DECLARATIONS

Acknowledgement

The authors thank the Institute of Forest Genetics and Tree Breeding (IFGTB), Coimbatore, for providing laboratory facilities and infrastructural support for this study. The authors also acknowledge the Krishi Vigyan Kendra (KVK), Sirugamani, Tiruchirappalli, and the Tropical Forest Research Institute (TFRI), Jabalpur, for their assistance during field sampling across Tamil Nadu.

Author contributions

D. Thangamani: conceptualization, methodology, data collection, formal analysis, writing original draft. Raju Kannan: field sampling, data curation, writing review and editing. Fatima Shirin: methodology, molecular analysis, writing review and editing.

R. Manikandan: involved in survey and taxonomical identification and helped in getting the accession Gayathri Prabhakaran: data analysis, statistical analysis, writing review and editing. Venkatesan Sathiya Aravindan: supervision, conceptualization, project administration, writing review and editing. All authors read and approved the final manuscript.

Funding

The authors declare that they did not receive any funding from external agencies.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Permissions to Collect Plant Material

Field collections of *Madhuca longifolia* were conducted during 2024–2025 under institutional guidelines for forest genetic resource collection. Necessary permissions were obtained from the Principal Chief Conservator of Forests, Government of Tamil Nadu, vide Ref. No. WL5(A)579/2021, Permit No. 28/2021, valid until June 2026, covering Anamalai Tiger Reserve and other designated forest divisions.

Plant Ethics

Identifier of Plant Material: The plant specimens were taxonomically identified by Dr. R. Manikandan, Scientist E, Botanical Survey of India Southern circle Coimbatore

Herbarium Details: Voucher specimens were deposited in the Fischer Herbarium (FRC), Institute of Forest Genetics and Tree Breeding, Coimbatore, under Accession No. 123656.

Voucher Specimens: The voucher specimen of *Madhuca longifolia* collected from Anamalai Tiger Reserve is publicly available at IFGTB Herbarium for reference

Ethics approval and consent to participate. Plant materials were acquired in compliance with local or national guidelines, requiring no further authorization.

Consent to publication

Not applicable.

Competing interests

The authors declare no competing interests.

Clinical trial number

Not applicable.

REFERENCES

1. Ankita B, Nabin K, Sagarika P. A review on potential application of *Madhuca longifolia* (J. Konig.) J.F. Macbr. in traditional systems and industrial purpose. Plant Sci Today. 2025. https://doi.org/10.14719/pst.5962.
2. Kumar A, Singh U. Chemical constituents, traditional uses and pharmacological properties of *Madhuca longifolia*: a review. Int J Pharm Sci Med. 2024;9(5):1–15. https://doi.org/10.47760/ijpsm.2024.v09i05.001.
3. Salgotra RK, Chauhan BS. Genetic diversity, conservation, and utilization of plant genetic resources. Genes. 2023;14(1):174. https://doi.org/10.3390/genes14010174.
4. Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK. Microsatellite markers: an overview of the recent progress in plants. Euphytica. 2011;177(3):309–34. https://doi.org/10.1007/s10681-010-0286-9.
5. 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. https://doi.org/10.1093/bioinformatics/bts460.
6. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980;32(3):314–31.

7. Nei M. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics. 1978;89(3):583–90. https://doi.org/10.1093/genetics/89.3.583.
8. Yeh FC, Yang RC, Boyle T. POPGENE Version 1.32: Microsoft Windows-based freeware for population genetic analysis. University of Alberta; 1999.
9. Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics. 1992;131(2):479–91. https://doi.org/10.1093/genetics/131.2.479.
10. Tamura K, Stecher G, Kumar S. MEGA11: molecular evolutionary genetics analysis version 11. Mol Biol Evol. 2021;38(7):3022–7. https://doi.org/10.1093/molbev/msab120.
11. Mohan R, Maurya A, Nair NK, Pandey N, Rana TS. Analyses of genetic diversity and population structure in *Quercus griffithii* Hook. f. & Thomson ex Miq. using simple sequence repeat (SSR) markers. Tree Genet Genomes. 2024;20:20. https://doi.org/10.1007/s11295-024-01654-w.
12. Zhong X, Lin C, Lin X, Lin W. Development of EST-SSRs based on the transcriptome of *Castanopsis carlesii* and cross-species transferability in other *Castanopsis* species. PLoS ONE, 2023;18(5), e0285291.
13. Dong M, Zhang L, Zhang Y, Feng J, Zhang S, Wang X. Development of EST-SSR markers in *Larix principis-rupprechtii* Mayr and evaluation of their polymorphism and cross-species amplification. Trees. 2018;32(6):1559–71. https://doi.org/10.1007/s00468-018-1733-9.
14. Pham M-P, Vu DD, Bei C, Bui TTX, Vu DG, Shah SNM. Characterisation of the *Cinnamomum parthenoxylon* (Jack) Meisn (Lauraceae) transcriptome using Illumina paired-end sequencing and EST-SSR markers development for population genetics. Biodivers Data J. 2024;12:e123405. https://doi.org/10.3897/BDJ.12.e123405.
15. Ahmad A, Wang J, Pan Y, Sharif R, Gao S. Development and use of simple sequence repeat (SSR) markers for sugarcane breeding and genetic studies. Agronomy. 2018;8(11):260. https://doi.org/10.3390/agronomy8110260.
16. Saina JK, Li Z-Z, Ngarega BK, Gituru RW, Chen J-M, Liao Y-Y. Exploring the genetic diversity and population structure of *Ailanthus altissima* using chloroplast and nuclear microsatellite DNA markers across its native range. Front Plant Sci. 2023;14:1197137. https://doi.org/10.3389/fpls.2023.1197137.
17. Zhang Y, Zhang X, Zhang Y, Zhang J, Zhang Y, Zhang X, Zhang J, Zhang Y. Genetic diversity of *Juglans mandshurica* populations in Northeast China based on SSR markers. Front Plant Sci. 2022;13:931578. https://doi.org/10.3389/fpls.2022.931578.
18. Addisalem AB, Duminil J, Wouters D, Girmay G, Bongers F, Smulders MJM. Fine-scale spatial genetic structure in the frankincense tree *Boswellia papyrifera* (Del.) Hochst. and implications for conservation. *Tree Genetics & Genomes*, 2016;*12*(86). https://doi.org/10.1007/s11295-016-1039-2
19. Goncalves AL, Moraes MLT, Freitas MLM, Zanatto ACS, Sebbenn AM, Oda S. Spatial genetic structure and mating system in forest tree populations from seasonally dry tropical forests: a review. Tree Genet Genomes. 2022;18(6):49. https://doi.org/10.1007/s11295-022-01550-1.

20. Li N, Yang Y, Xu F, Chen X, Wei R, Li Z, Pan W, Zhang W. Genetic diversity and population structure analysis of *Castanopsis hystrix* and construction of a core collection using phenotypic traits and molecular markers. Genes. 2022;13(12):2383. https://doi.org/10.3390/genes13122383.
21. Yan H, Qi H, Li Y, Wu Y, Wang Y, Chen J, Yu J. Assessment of the genetic relationship and population structure in oil-tea *Camellia* species using simple equence repeat (SSR) markers. Genes. 2022;13(11):2162. https://doi.org/10.3390/genes13112162.
22. Guo Q, Li Y, Cao S, Fang Q, Yue Y, Jia L. Genetic diversity and population structure of *Robinia pseudoacacia* from six improved variety bases in China as revealed by simple sequence repeat markers. J Forestry Res. 2021;32:1547–57. https://doi.org/10.1007/s11676-021-01356-2.
23. Lozano-Puentes HS, Tarazona-Pulido L, López-Alvarez D, Ruiz-Sanchez E, Costa GM, Díaz-Ariza LA. Assessing genetic variation in *Guadua angustifolia* through RAD-seq analysis. Int J Mol Sci. 2025;26(12):5879. https://doi.org/10.3390/ijms26125879.
24. Dias A, Roque N, Araújo M, Tomas D, Dias LS, Semedo-Lemsaddek T. Portuguese *Pinus nigra* J.F. Arnold populations: genetic diversity, structure and relationships inferred by SSR markers. Ann For Sci. 2020;77:87. https://doi.org/10.1007/s13595-020-00967-9.
25. Akpertey A, Padi FK, Meinhardt L, Zhang D. Effectiveness of single nucleotide polymorphism markers in genotyping germplasm collections of *Coffea canephora* using KASP assay. Front Plant Sci. 2021;11:612593. https://doi.org/10.3389/fpls.2020.612593.
26. Oliveira AM, Ramos SLF, Ferreira MJ, Lopes R, Meneses CHSG, Valente MSF, da Silva RF, Batista JdS, Muniz AW, Lopes MTG. Mating system analysis and genetic diversity of *Parkia multijuga* Benth.: one native tree species of the Amazon. Forests. 2024;15(1):172. https://doi.org/10.3390/f15010172.
27. Lanes ÉCM, Motoike SY, Kuki KN, Nick C, Sousa VB, Bhering LL. Mating system and genetic composition of the macaw palm (*Acrocomia aculeata*): implications for breeding and genetic conservation programs. J Hered. 2016;107(3):258–66. https://europepmc.org/article/MED/27288529.
28. Coelho NHP, Carvalho CR, Bhering LL. Understanding genetic diversity, spatial genetic structure, and mating system through microsatellite markers for the conservation and sustainable use of *Acrocomia aculeata* (Jacq.) Lodd. Ex Mart. Conserv Genet. 2018;19(5):1151–62. https://doi.org/10.1007/s10592-018-1061-z.
29. Peng W, Song T, Du H, Chen H, Zeng F, Liu Y, Luo Y, Tan W. Inconsistent diversity patterns of soil fungi and woody plants among habitat types in a karst broadleaf forest. For Ecol Manag. 2023;532:120707. https://doi.org/10.1016/j.foreco.2023.120707.
30. Homczinski I, Lerner J, Peres FSB, Gasparin E, Machado SA, Tambarussi EV. Molecular and quantitative genetic analysis of the neotropical tree *Campomanesia xanthocarpa* (Mart.) O. Berg. Annals For Res. 2022;65(1):37–54. https://doi.org/10.15287/afr.2022.2223.
31. Li Z, Cheng Y, Li F, Liu J. Molecular mechanisms of heterosis and its applications in tree breeding: progress and perspectives. Int J Mol Sci. 2024;25(22):12344. https://doi.org/10.3390/ijms252212344.

32. Yan L, Gao C, Liu C, Wang Y, Liu N, Zhang X, Liu F. Predicting heterosis and selecting superior families and individuals in *Fraxinus* spp. based on growth traits and genetic distance coupling. Plants. 2025;14(16):2601. https://doi.org/10.3390/plants14162601.

Figures

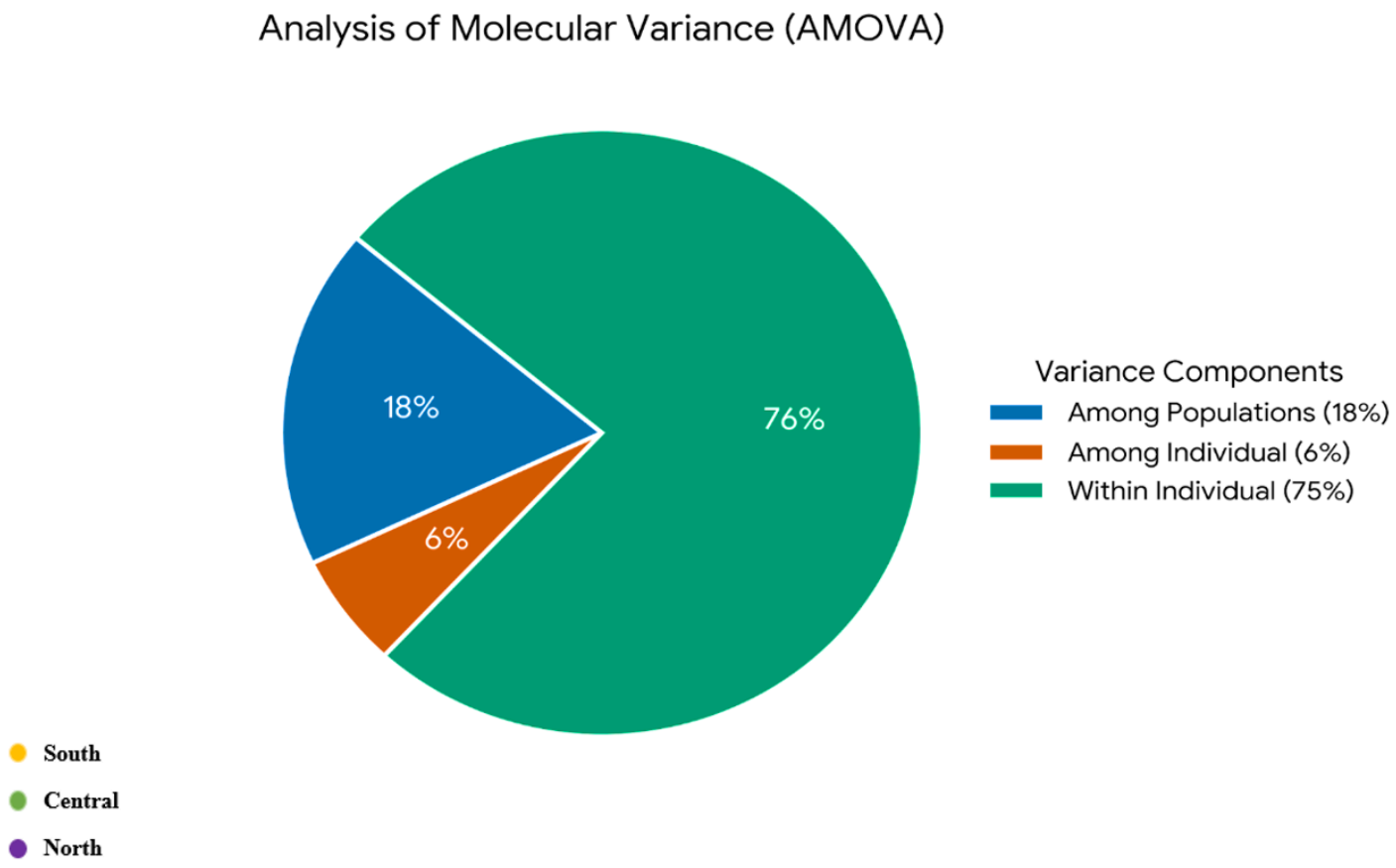


Figure 1

Percentage of Molecular Variance

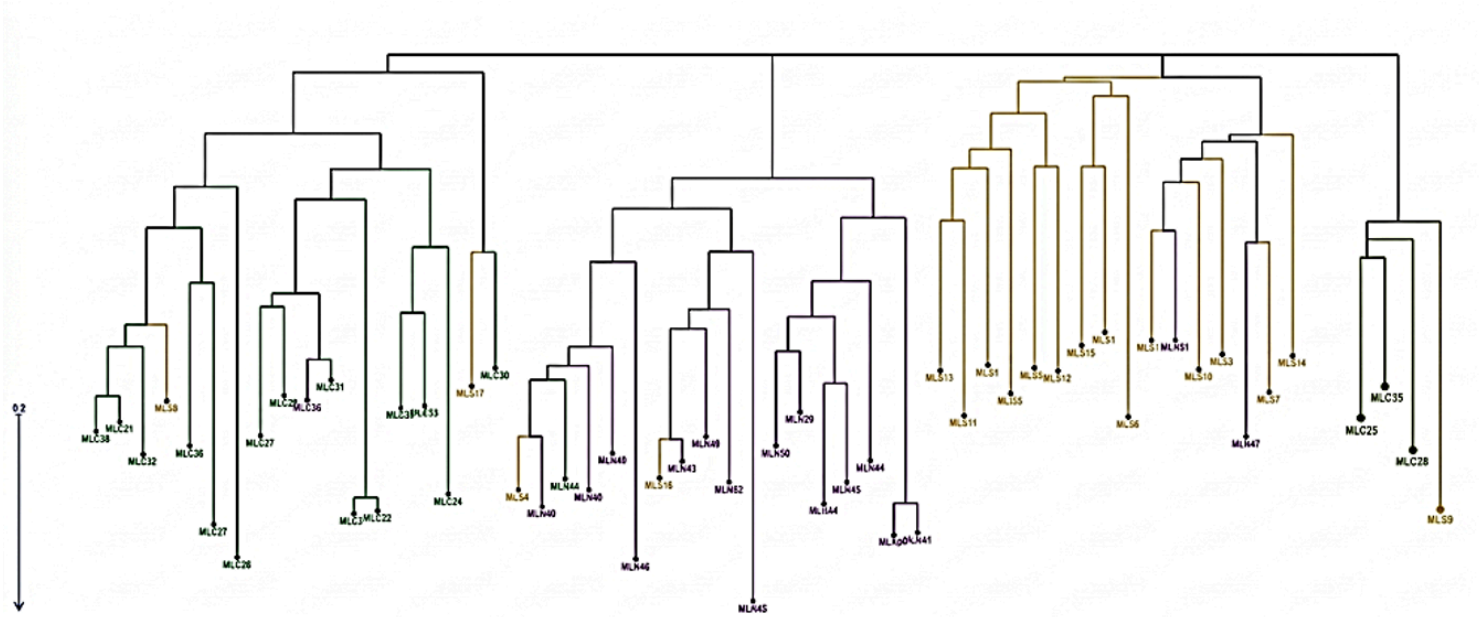


Figure 2

Neighbour joining tree of *Madhuca longifolia* genotype constructed using SSR loci

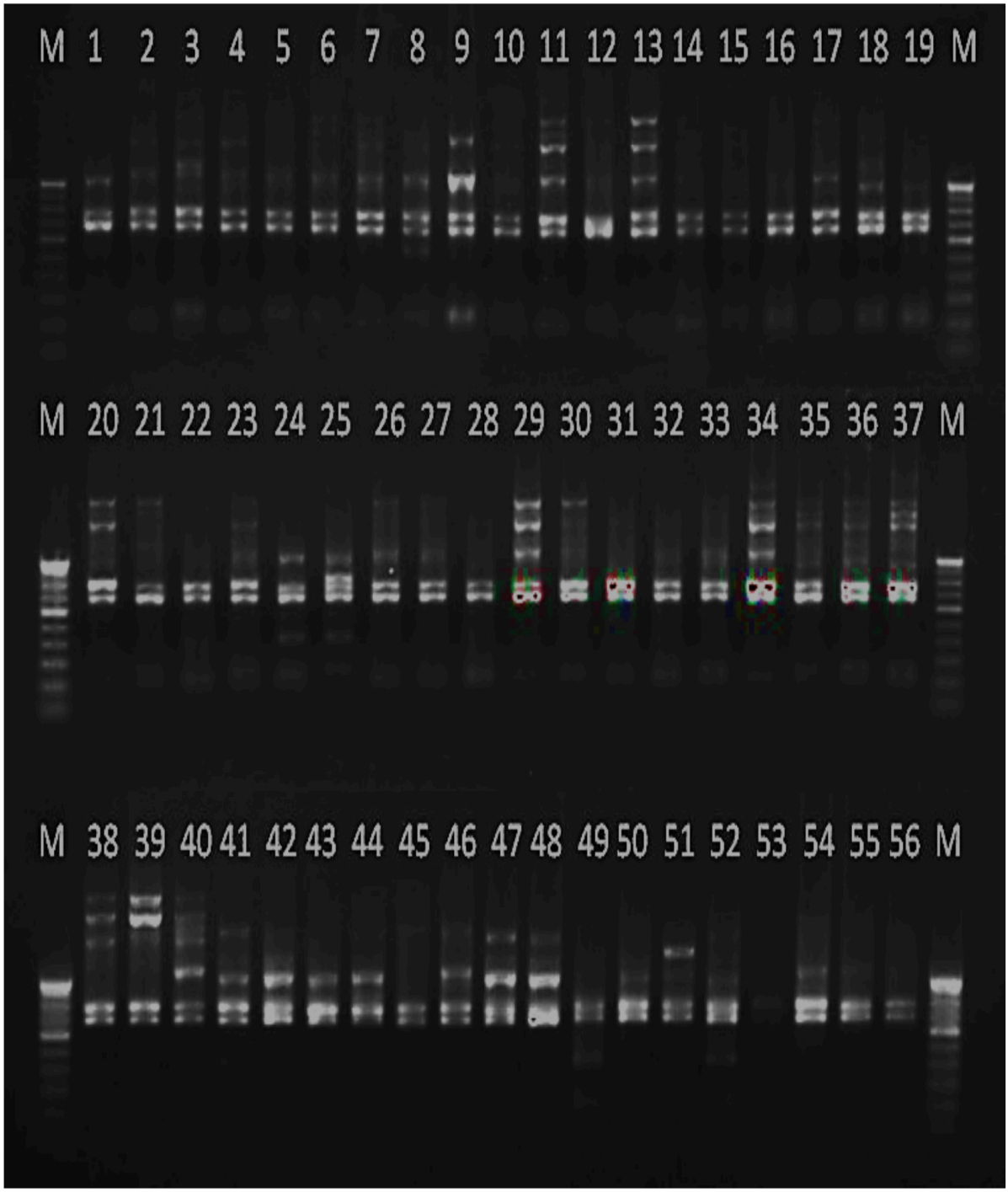


Figure 3

SSR profile of different *Madhuca longifolia* germplasms produced by SSR primer 2 Lane M = 100 bp ladders

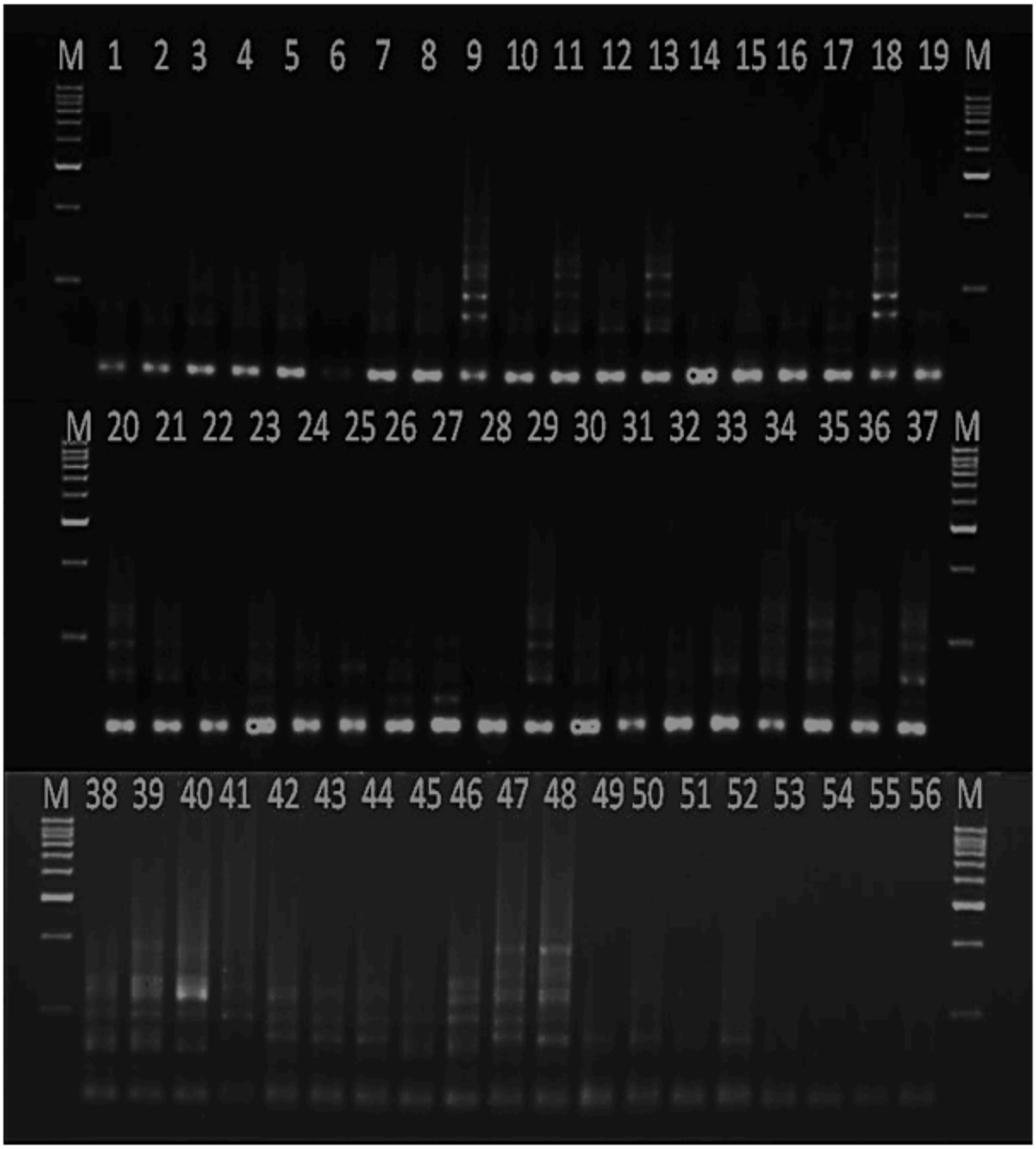


Figure 4

SSR profile of different *Madhuca longifolia* germplasms produced by SSR primer 4 Lane M = 100 bp ladders

Supplementary Files

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

- Madhucalongifoliaherbariumsheet.png