Research Article

Bei Cui[#], Dinh Duy Vu*[#], Dinh Giap Vu[#], Thi Tuyet Xuan Bui, Siddiq Ur Rahman, Mai Phuong Pham, Minh Tam Nguyen, Van Sinh Nguyen, Syed Noor Muhammad Shah, Viet Ha Tran[#]

Genetic diversity and population structure of *Cinnamomum balansae* Lecomte inferred by microsatellites

Abstract: *Cinnamomum balansae* Lecomte (Lauraceae), an economically important forest tree, is distributed in the tropical forests of central and northern Vietnam, which has been threatened in recent decades due to the destruction of its habitat and over-exploitation. The genetic diversity and population structure of the species have not been fully evaluated. We used a set of 15 microsatellites to analyze 161 adult trees from 9 different populations, representing the geographical distribution of *C. balansae*. Ninety-two different alleles were identified. Here our results showed a low genetic diversity level with an average $H_o = 0.246$ and $H_e = 0.262$, and a high level of genetic differentiation ($F_{ST} = 0.601$). The bottleneck tests indicated evidence of a reduction in the population size of the two populations (TC and CP). Additionally, all three clustering methods (Bayesian analysis, principal coordinate analysis, and Neighbor-joining tree) were identified in the two genetic groups. The Mantel test showed a significant positive correlation between genetic distance and geographic distance ($R^2 = 0.7331$). This study will provide a platform for the conservation of *C. balansae* both in *ex-situ* and *in-situ* plans.

Keywords: bottleneck, *C. balansae*, conservation, genetic variability, fragmentation

# These authors contributed equally to this work and should be regarded as co-first authors.

* **Corresponding author: Dinh Duy Vu,** Graduate University of Science and Technology (GUST), Vietnam Academy of Science and Technology (VAST), 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam; Institute of Tropical Ecology, Vietnam – Russia Tropical Centre, 63 Nguyen Van Huyen, Cau Giay, Hanoi, Vietnam, tel: +84-977365689, e-mail: duydinhvu87@gmail.com, duyvu@vnmn.vast.vn
Bei Cui: Key Comprehensive Laboratory of Forestry, College of Forestry, Northwest A & F University, Yang Ling 712100, China
Dinh Giap Vu: Department of Chemical Technology – Environment, HaUI Institute of Technology, Hanoi University of Industry (HaUI), 298 Cau Dien, Bac Tu Liem, Hanoi, Vietnam
Thi Tuyet Xuan Bui: Department of Plant Ecology, Institute of Ecology and Biological Resource, VAST, 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam; Graduate University of Science and Technology, VAST, 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam
Siddiq Ur Rahman: Department of Computer Science & Bioinformatics, Khushal Khan Khattak University, Karak, Khyber Pakhtunkhwa, 27200, Pakistan
Mai Phuong Pham: Institute of Tropical Ecology, Vietnam – Russia Tropical Centre, 63 Nguyen Van Huyen, Nghia Do, Cau Giay, Hanoi, Vietnam
Minh Tam Nguyen: Department of Experimental Taxonomy and Genetic Diversity, Vietnam National Museum of Nature, VAST, 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam
Van Sinh Nguyen: Department of Plant Ecology, Institute of Ecology and Biological Resource, VAST, 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam; Graduate University of Science and Technology, VAST, 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam
Syed Noor Muhammad Shah: Department of Horticulture, Faculty of Agriculture, Gomal University, Dera Ismail Khan, 29220, Pakistan
Viet Ha Tran: Faculty of Silviculture, Vietnam National University of Forestry, Xuan Mai, Chuong My, Hanoi, Vietnam

the tropical forests of central and northern Vietnam, which has been threatened in recent decades due to the destruction of its habitat and over-exploitation. The genetic diversity and population structure of the species have not been fully evaluated. We used a set of 15 microsatellites to analyze 161 adult trees from 9 different populations, representing the geographical distribution of *C. balansae*. Ninety-two different alleles were identified. Here our results showed a low genetic diversity level with an average $H_o = 0.246$ and $H_e = 0.262$, and a high level of genetic differentiation ($F_{ST} = 0.601$). The bottleneck tests indicated evidence of a reduction in the population size of the two populations (TC and CP). Additionally, all three clustering methods (Bayesian analysis, principal coordinate analysis, and Neighbor-joining tree) were identified in the two genetic groups. The Mantel test showed a significant positive correlation between genetic distance and geographic distance ($R^2 = 0.7331$). This study will provide a platform for the conservation of *C. balansae* both in *ex-situ* and *in-situ* plans.

Keywords: bottleneck, *C. balansae*, conservation, genetic variability, fragmentation

1 Introduction

Cinnamomum balansae Lecomte, a large and evergreen tree of the Lauraceae family, is distributed in the tropical forests of central and northern Vietnam [1–4]. It is found at an elevation above 200 m in secondary forests on the ancient alluvial rocks and granite that have low relief and gentle slopes. The species prefers precipitation of 800–2,500 mm and an annual mean temperature of 20–22°C [2,4]. *C. balansae* can reach a height of 30 m with a diameter of 85–90 cm at breast height. Flowers are small, greenish-yellow, and appear between March and May. Fruiting appears from April to July, with fruit

and maturation occurring in October [2]. Its bark has oil and camphor. The fruit oil and fat are used to make soap. Wood is used for furniture. In the past few decades, this species has been over-exploited by local people and forestry enterprises [1]. Its habitats have been destroyed and degraded [5,6]. For that reason, *C. balansae* is listed as an endangered species in the IUCN Red List [7] and the Vietnam Red Data Book [1].

The adaptive and evolutionary potential of a species depends on its genetic diversity [8]. Conservation and management of a species require knowledge of its ecological characteristics and genetic variation within and between populations [9–11]. To obtain such information, particularly a better understanding of genetics, powerful and biological techniques are required [12,13]. Molecular markers such as allozyme [14], randomly amplified polymorphic DNA (RAPD) [15,16], and chloroplastic DNA [17] have been used to investigate the genetic diversity of species in the Lauraceae family. Microsatellites (SSRs) are widely distributed throughout the nuclear genome, displaying high polymorphism and reproducibility [18–20]. These markers have been widely used to study the Lauraceae species [21–29]. In the current study, microsatellites were used to assess the genetic diversity and population structure of *C. balansae* and provide a platform for effective conservation strategies of this species in Vietnam.

2 Materials and methods

2.1 Plant materials

With Illumina sequencing for the development of microsatellites, we collected three plant tissues, including roots, leaves, and stems of *C. balansae* (one individual) from a wild population (Ba Vi National Park, Hanoi city). All sample tissues were immediately frozen in liquid nitrogen and stored at −80°C until RNA extraction.

To assess the genetic variation in and population structure of *C. balansae*, the study area for this species was carried out in nine sites, one each in Hanoi, Phu Tho, Tuyen Quang, Yen Bai, Son La, Hoa Binh, Ninh Binh, Thanh Hoa, and Nghe An, representing its natural range in Vietnam (Table 1, Figure 1). Sampling locations were recorded using a global positioning system (GPS). The inner bark from 161 adult trees was randomly sampled for 9 known populations in central and northern Vietnam. Samples were placed into plastic bags containing silica gel in the field and then transferred to the Laboratory of Molecular Biology of Vietnam-Russia Tropical Centre and stored at −30°C until DNA extraction.

2.2 DNA/RNA extraction, cDNA sequencing, and microsatellite loci identification

For Illumina sequencing, total RNA from each sample was isolated using the OmniPlant RNA Kit (DNase I) (www.cwbiotech.com). Both quantity and quality of total RNA were checked through Nanodrop ND-2000 spectrophotometer (Thermo Electron Corporation, USA) and Agilent 2100 Bioanalyzer. An equal amount of total RNA from each sample was pooled together and sent to Breeding Biotechnologies Co. Ltd for cDNA library construction and transcriptome sequencing using Illumina HiSeq™4000. Cleaned mRNA was used for cDNA library construction extracted from 200 μg total RNA using Oligo (dT). The cDNA first strand was prepared from random

Table 1: Geographical characteristics of the nine *C. balansae* populations used in this study

| Populations	Location	Sample size	Latitude (N)	Longitude (E)
BV	Ba Vi National Park, Ha Noi City (Northern Vietnam)	10	21°06′36″	105°33′16″
DH	Chan Mong commune, Doan Hung district, Phu Tho province (Northern Vietnam)	15	21°33′45″	105°09′49″
SD	Dong Bua commune, Son Duong district, Tuyen Quang province (Northern Vietnam)	19	22°10′15″	105°20′10″
YB	Van Lang Commune, Yen Binh district, Yen Bai province (Northern Vietnam)	12	21°41′52″	104°59′39″
TC	Chieng Bom Commune, Thuan Chau district, Son La province (Northern Vietnam)	16	21°38′87″	103°53′52″
MC	Tan Mai commune, Mai Chau district, Hoa Binh province (Northern Vietnam)	20	20°44′46″	104°55′32″
CP	Cuc Phuong National Park, Ninh Binh province (Central Vietnam)	21	20°21′06″	105°35′57″
BE	Ben En National Park, Thanh Hoa province (Central Vietnam)	21	19°58′08″	105°45′57″
QH	Tam Hop Commune, Quy Hop district, Nghe An province (Central Vietnam)	27	19°21′21″	105°17′05″
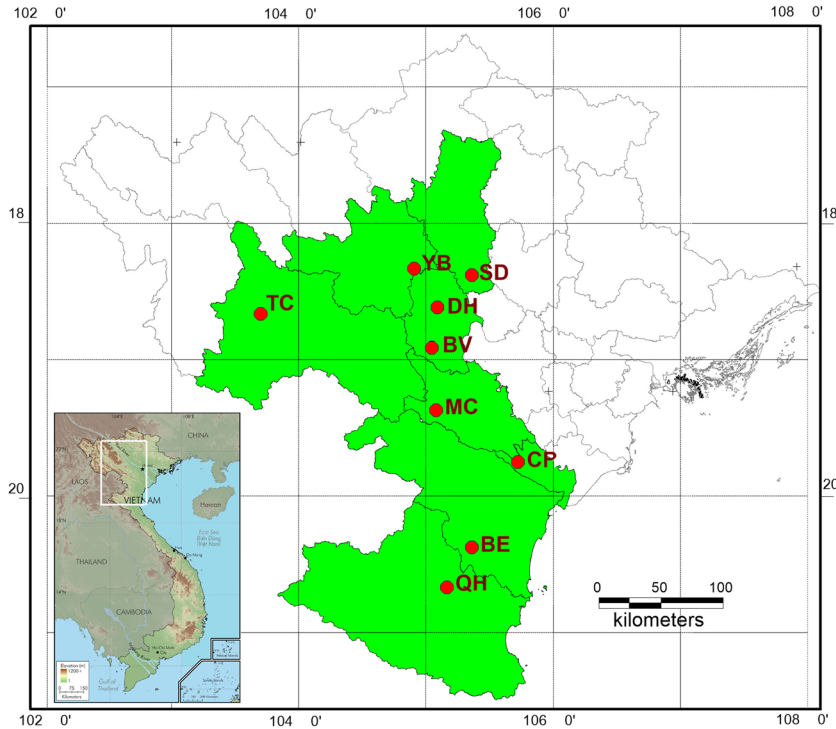
Figure 1: Map showing the collection sites of nine populations of *C. balansae* in Vietnam.

hexamers using mRNA as a template, and the other strands were from the buffer, dNTPs, RNase H, and DNA polymerase I, and then cleaned with AMPure XP beads. The cleaned double-stranded cDNA was subjected to terminal repair, the sequencing linker was ligated, and then the fragment size was selected with AMPure XP beads. The cDNA library was acquired by PCR enrichment. After validating the library on an Agilent 2100 Bioanalyzer, Breeding Biotechnologies Co., Ltd sequenced the cDNA libraries on an Illumina HiSeq™ 4000. The Trimmomatic v3.0 [30] was used for raw reads filtration. The reads showing adaptor contamination, length < 36 bp, and low-quality value (quality <20) higher than 15% were eliminated. Trinity [31] with default parameters were used for de novo assembly of the cleaned reads. Then, the TIGR Gene Indices clustering tool (TGICL) v2.1 [32] was used to cluster and eradicate redundant transcripts and identified unigenes for further analysis.

SSRs were detected in the Simple Sequence Repeat Identification Tool program (SSRIT, http://www.gramene.org/db/markers/ssrtool). Only unigenes longer than 1 kb were included in the EST-SSR detection. The parameters were set to detect perfect, di-, tri-, tetra-, penta-, and hexa-nucleotide motifs with a minimum of 6, 5, 5, 5, and 5 repeats, respectively. Identification of microsatellite loci

showed simple sequence repeats and was deposited in GenBank.

Total genomic DNA was extracted from the samples using the Plant Genome DNA Extraction Kit (BioTeke, Beijing, China). The quality of total DNA was verified by 1% agarose gel electrophoresis and NanoDrop2000 (BioTeke Instruments, Winooski, VT, USA). The concentration was then diluted to 20 ng/μL and stored at −20°C.

2.3 Microsatellite amplification

Polymerase chain reaction (PCR) was performed in a 25 μL of reaction mixture containing 2.5 μL of template DNA, 12.5 μL of 2X Taq Master mix, 1 μL of each primer, and 8 μL of deionized water. All SSR primers were designed using Primer 5.0 software [33]. The main stated selection criteria for the screening of primers were as follows GC content of 45–55 with 55% as the optimum; primer length of 18–24 bp with 20 bp as the optimum length; annealing temperature from 55 to 65°C, with 60°C as the optimum temperature; and PCR product size ranging from 100 to 300 bp. Fifteen SSR polymorphic markers were used in the present study (Table 2). The PCR

amplification was carried out on a Mastercycler X50s Eppendorf, under the following thermal cycler: initial denaturation at 94°C for 2 min, followed by 40 cycles consisting of 1 min at 94°C for denaturation, 30 s alignment at the annealing temperature for each primer pair at 55°C and 1 min at 72°C for extension, and 10 min at 72°C for final cycle to complete the extension of any remaining products before holding the samples at 4°C until they were analyzed. The amplification products were separated on 8% polyacrylamide gels in 1 × TAE buffer using the Cleaver Scientific electrophoresis system (UK) and visualized with GelRed™ 10000X. The sizes of the PCR products were detected and analyzed using GelAnalyzer software with a 20 bp DNA ladder (Invitrogen).

2.4 Data analysis

Genetic diversity was estimated using GenAlEx [34] based on SSR allele frequencies, including the number of alleles per locus (N_a), effective alleles (N_e), private alleles (N_p), the proportion of polymorphic loci ($P\%$), the coefficient of total inbreeding (F_{IT}), the genetic differentiation (F_{ST}), Nei's genetic distances among the populations, the observed heterozygosities (H_o), and the expected heterozygosities (H_e). Gene flow between populations (N_m) was calculated using the F_{ST} value: $N_m = (1/F_{ST} - 1)/4$. The inbreeding coefficients (F_{IS}) value was also calculated for each population by Fstat [35]. Tests for deviation from the Hardy-Weinberg equilibrium at each locus and

Table 2: Characteristics and genetic estimates for 15 SSR loci of *C. balansae*

Locus	Primer sequence (5′–3′)	Motif	Size (bp)	N_a	N_p	F_{IS}	F_{IT}	F_{ST}	N_m	GenBank code
V03	F: CCACCCCAATTCGATTATTG R: CCAAGCAGCTTTACTCCACC	$(AGCAGG)_5(AGC)_6$	219	5	2	0.335	0.645	0.467	0.285	MN486375
V07	F: GAGAGCTTCCAAAGCCACTG R: AAGAAGCCGACAAGCAAAAA	$(CAC)_5(CAT)_5$	234	6	3	0.635	0.864	0.626	0.149	MN486379
V08	F: AGGGATTTTGGAGATTTTGC R: GGACGCCAAAACCAAATGTA	$(CTT)_5(CCT)_7$	150	9	4	−0.130	0.239	0.327	0.515	MN486380
V11	F: TTTGAGGGAAATGGTTCCTG R: TTGGGCCAAAACAAGAAGAC	$(TAG)_5(TAA)_6$	194	6	2	−0.096	0.453	0.501	0.249	MN486383
V25	F: ACGAGTTTCCACCGATTACG R: ACTCCTTTCAGCACCGATTG	$(ATC)_8$	268	5	2	−0.024	0.992	0.992	0.002	MN486397
V30	F: TTGTTAAAAACACCAACCCCA R: CAGTGGGCCAAGTGTATCCT	$(TCA)_8$	200	5	2	−0.113	0.524	0.572	0.187	MN486402
V31	F: ACGTGAATGTGAATGGGGTT R: TAGGCAAAGACTCCGAAGGA	$(AAC)_9$	189	5	1	0.219	0.851	0.809	0.059	MN486403
V33	F: ATTCGGGTCTCTCTCCCTGT R: CTCTCTCGCTGTCTCTGCCT	$(ACCA)_6$	249	8	2	0.220	0.620	0.513	0.237	MN486405
V35	F: TGGAGAACAACTTTGGGAGG R: TGTTCCATGTTACAGATACAG	$(GATG)_6$	119	7	2	−0.258	0.518	0.617	0.155	MN486407
V41	F: ATATGGTCCCAACTCCCTCC R: ACCGTCACCAGATCATCCAT	$(ACCCA)_5$	203	8	4	0.289	0.808	0.729	0.093	MN486413
V51	F: TCCCAACTGGACGAAGTTCT R: TTTGCTCGCTGTTATGATGC	$(ATGACC)_7$	237	4	1	0.505	0.834	0.665	0.126	MN486423
V75	F: ATCCTCCCAAGGACGCTTAT R: CCTTCAAGGAAAGAAGGGCT	$(CCT)_7$	133	7	3	0.287	0.625	0.473	0.278	MN486447
V81	F: CACCACCTTCTCCTTCCAAA R: TTGTTGGGGTCTCCAAACTC	$(CTC)_7$	234	4	2	−0.351	0.740	0.808	0.060	MN486453
V93	F: CGGACAATTTTCAGGGATGA R: CCTGCATTCTAGATGCCTCC	$(GTC)_7$	166	6	2	−0.096	0.594	0.630	0.147	MN486465
V108	F: GAAAGAAGTTGGGGGAGGAG R: GAATCGCGGAGATACGGATA	$(TGA)_7$	209	6	3	−0.315	0.060	0.285	0.627	MN486480
Mean				6.1		0.074	0.624	0.601	0.211	

N_a: number of alleles; N_p: private alleles; F_{IS}: inbreeding coefficient; F_{IT}: coefficient of total inbreeding; F_{ST}: genetic differentiation; N_m: gene flow.

the linkage disequilibrium for each locus pairwise combination in each population were also performed by Genpop v.4.6 [36]. Arlequin v.3.0 [37] was used to perform an analysis of molecular variance (AMOVA) for the estimation of genetic divergence as PhiPT, and its associated significance based on 10,000 permutations. Testing for recent bottleneck events for each population using Bottleneck v.1.2 [38] was applied with the stepwise mutation model (SMM) and two-phase model (TPM), and their significance was tested. The proportion of the SMM was set to 70% under default settings. The significance of these tests was evaluated by the one-tailed Wilcoxon signed-rank test.

Principal coordinate analysis (PCoA) was implemented based on G'_{ST} values [39] using GenAlEx. Neighbor-joining (NJ) tree was constructed genetic relationships among populations based on a matrix of F_{ST} values using Poptree2 [40]. Bayesian analysis of population structure was performed with Structure v.2.3.4 [41] to determine the optimal value of the genetic cluster (K). Setting the admixture model with correlated allele frequency and the ancestry models, 10 separate runs of the number of clusters (K) in the dataset were performed from 1 to 10 for each K value at 500,000 Markov Chain Monte Carlo (MCMC) repetitions and a 100,000 burn-in period. The optimal value of K was detected using Structure Harvester [42] based on the ΔK value by Evanno et al. [43].

A Mantel test [34] was applied to determine the relationships between genetic and geographic distances among the *C. balansae* populations based on 999 permutations.

3 Results

3.1 Genetic diversity

A total of 92 different alleles were detected from 15 SSR markers in *C. balansae* in the tropical forests of Vietnam. Allele numbers ranged from 4 (VD51 and VD81) to 9 (VD08), with an average of 6.1 (Table 2). Among these, 35 alleles were private. Loci VD08 and VD49 had the greatest number of private alleles (4). One private allele at 2 loci (VD31 and VD51), 3 private alleles at 3 loci (VD07, VD75, and VD108), and 2 private alleles at 8 remaining loci were recorded.

Genetic diversities for *C. balansae* populations are presented in Table 3. The average percentage of polymorphic loci was 68.15%, ranging from 60% in the 4 populations of BV, SL, TQ, and NA to 86.67% in the TH population. The average value of alleles (N_a) per population was 1.9, ranging from 1.6 (SL) to 2.2 (YB and TH).

The highest number of private alleles (N_p = 8) was detected in the CP population, followed by three populations of YB, BE, and QH with N_p = 6, two populations of MC and DH with N_p = 4, the BV population (N_p = 2), and two populations (TC and SD) with N_p = 1. The N_e value varied from 1.4 in 5 different populations (MC, DH, SD, BE, and QH) to 1.7 (YB population), with an average of 1.5 alleles. The observed heterozygosity (H_o) and expected heterozygosity (H_e) averaged 0.246 (0.143–0.339) and 0.262 (0.198–0.338), respectively. The QH population

Table 3: Genetic diversity estimates nine populations of *C. balansae*

Population	N	N_a	N_p	N_e	$P\%$	H_o	H_e	F_{IS}	p value of bottleneck	
									TPM	SMM
BV	10	1.7	2	1.5	60	0.267	0.275	0.003	0.049	ns
MC	20	1.8	4	1.4	73.33	0.260	0.253	−0.011	ns	ns
DH	15	1.7	4	1.4	66.67	0.219	0.213	−0.028**	ns	ns
TC	16	1.6	1	1.5	60	0.250	0.267	0.073	0.014	0.014
SD	19	1.7	1	1.4	60	0.193	0.228	0.101	ns	ns
YB	12	2.2	6	1.7	73.33	0.339	0.338	−0.051	0.042	ns
BE	21	2.2	6	1.4	86.67	0.270	0.253	−0.076	ns	ns
CP	21	1.9	8	1.7	73.33	0.270	0.334	0.212*	0.0003	0.0002
QH	27	1.9	6	1.4	60	0.143	0.198	0.179	ns	ns
Mean		1.9		1.5	68.15	0.246	0.262	0.025		

N: sample size; p: percentage of polymorphic loci; N_a: alleles per locus; N_p: total number of private alleles; N_e: effective alleles; H_o and H_e: observed and expected heterozygosities; F_{IS}: inbreeding coefficient; *p < 0.05, **p < 0.01.

had the lowest genetic diversity (H_o = 0.143 and H_e = 0.198) and the YB population had the highest values (H_o = 0.339 and H_e = 0.338). Five populations of BV, SD, TC, CP, and QH had positive F_{IS} values, indicating an excess of homozygosity and inbreeding in these populations. The highest F_{IS} value (0.212) was found in the CP population and was significant ($p < 0.05$). A significant deficit of heterozygosity was detected in two populations (TC and CP), based on the Bottleneck analysis under the TPM and SMM ($p < 0.05$), and suggested evidence of a bottleneck.

3.2 Population structure and Mantel test

The average F_{IT} value was 0.624, ranging from 0.06 (VD108) to 0.992 (VD25), indicating a deficiency of heterozygosity in the populations (Table 2). The genetic differentiation among populations (F_{ST}) varied from 0.285 (VD108) to 0.992 (VD25) with an average of 0.601, indicating high differentiation among all the populations and reflecting low gene flow (N_m = 0.211). Our results of the AMOVA showed that the proportion of genetic variation among the populations was higher (74%) than within populations (Table 4). The PhiPT value and gene flow among populations were 0.744 and 0.086, respectively.

The PhiPT value for the variation among populations was significant ($p < 0.001$).

The Mantel test revealed a significantly high correlation between genetic distance and geographical distance (R^2 = 0.7331) (Figure 2), indicating high genetic differentiation among populations.

Genetic clusters were distinguished through the NJ analysis by using Poptree2 and the PCoA based on G'_{ST} values (Figure 3). Two different clusters were generated. The first cluster included three populations (CP, BE, and QH) in the central region, while the second included the remaining six populations (DH, SD, BV, MC, YB, and TC) in the northern region. An admixture model was performed to evaluate the population structure of the 161 individuals of *C. balansae* based on the Bayesian analysis using the Structure program. The results showed that the optimum number of genetic groups (K) was two, with the highest value of delta K (1,145.4) obtained from Structure Harvester might an optimum number of genetic clusters (Figure 4). Each individual in each population was assigned to one of the two genetic clusters. The ancestry coefficients of each individual and each population are shown in Figure 4. The red cluster represented all individuals from the populations of BV, MC, DH, TC, SD, and YB, while the green cluster represented CP and QH populations. All the individuals in the BE population were separated and closer to the red cluster than the green cluster.

Table 4: Analysis of molecular variance from nine populations of *C. balansae*

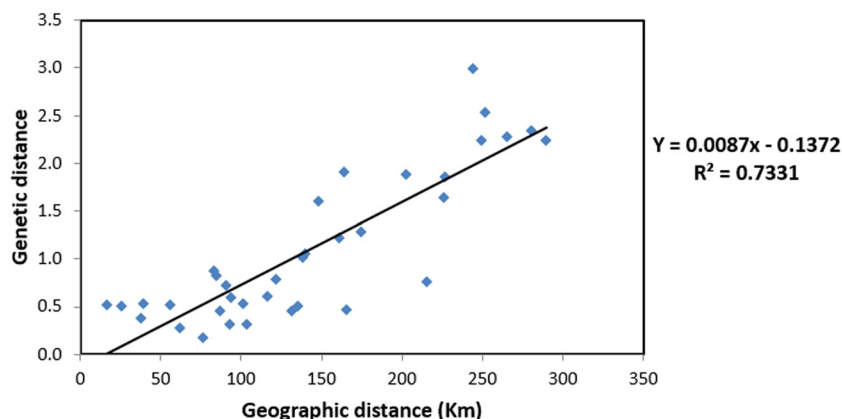|	df*	Sum of squares	Variance components	Proportion of variation (%)	PhiPT value	N_m
Among populations	8	1307.8	13.592	74	0.744**	0.086
Within populations	102	477.5	4.681	26		
Total	110	1785.3	18.273			

df* degree of freedom; **$p < 0.001$.

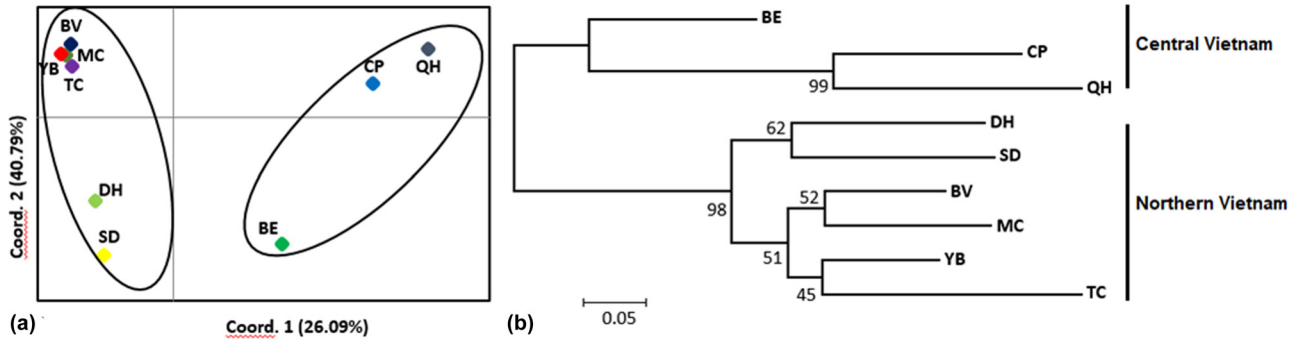**Figure 2:** Mantel test of genetic distance and geographical distance of nine *C. balansae* populations.

Figure 3: Principal Coordinates (PCoA) (a) and NJ tree for nine populations of *C. balansae* with bootstrap values above branches (b).
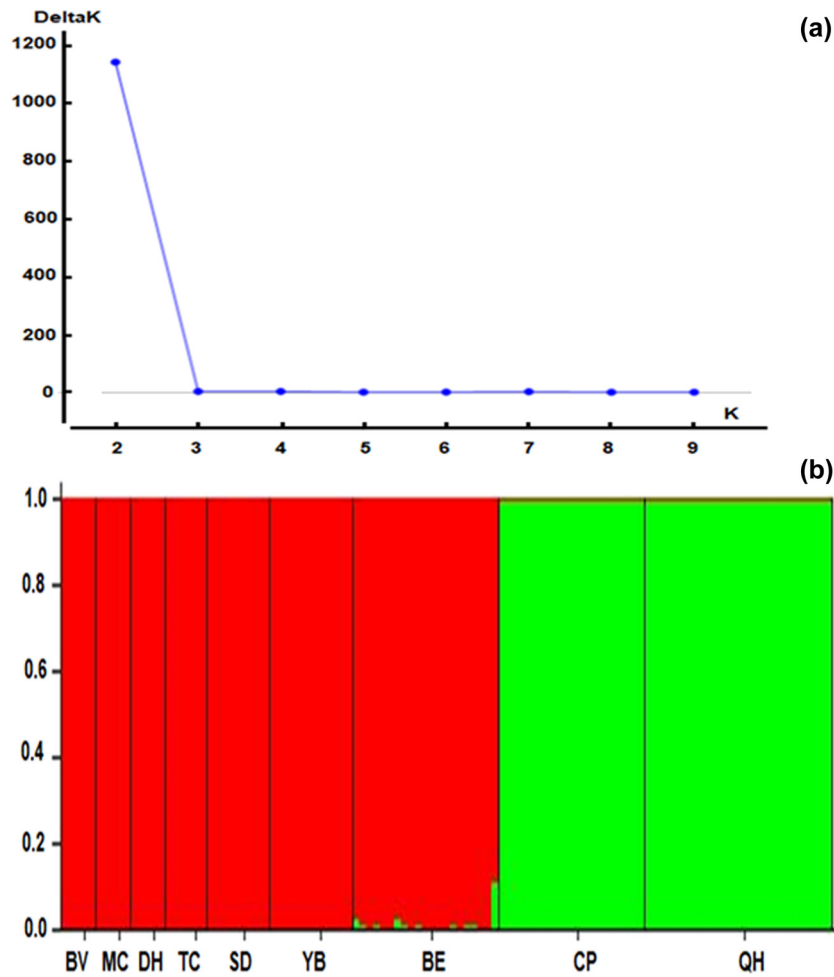

Figure 4: The Delta *K* distribution graph (a) and bar plot of admixture assignment for nine *C. balansae* populations to cluster ($K = 2$ and highest ΔK value = 1145.4) based on Bayesian analysis (b).

4 Discussion

The genetic diversity of a species is often correlated with its history and ecology [44–46]. The genetic diversity is positively correlated with geographic distribution range and population size [47]. Species with widespread distribution, large population size, longevity, and outcrosses often maintain high genetic diversity compared to species with a narrow distribution [48]. Here our results indicated that *C. balansae* has low genetic diversity

(H_o = 0.246 and H_e = 0.262) and high genetic differentiation among populations (F_{ST} = 0.601), compared to other widespread species, such as *C. camphora* in South China (H_o = 0.45 and H_e = 0.44) [29], *Albanian olive* (H_o = 0.75 and H_e = 0.60) [49 and E*ugenia dysenterica* (H_o = 0.545 and H_e = 0.62) [50]. Although this species has a widespread distribution and long life, the decline in genetic diversity may be related to anthropogenic disturbances, such as deforestation, habitat degradation, and over-exploitation of *C. balansae* [51]. Furthermore, in the current study, positive F_{IS} values were observed in the five populations of BV, TC, SD, CP, and QH, while the negative F_{IS} values were found in the remaining four populations of MC, DH, YB, and BE. A heterozygosity deficit was determined by the Bottleneck analysis, which detected a reduction in the population size of *C. balansae*. Small populations might be the result of inbreeding, which can decrease genetic variation. However, the bottlenecks also show a rapid reduction in the population sizes of TC and CP in recent decades. A high level of genetic differentiation (F_{ST} = 0.601) was determined, and reflected low gene flow (N_m = 0.211) among *C. balansae* populations. This is in contrast to previous reports on widespread species, such as *Dalbergia cochinchinensis* (F_{ST} = 0.23) [52] and *Erythrophleum fordii* Oliv (F_{ST} = 0.18) [53]. The dispersal of pollen and seeds plays an important role in maintaining high genetic diversity within and low genetic differentiation among populations [54,55]. The large geographic distance between *C. balansae* populations may be responsible for high genetic differentiation. Moreover, habitat fragmentation can increase the isolation between populations. Thus, these suggested the geographic distance and habitat fragmentation as a barrier to restrict gene flow between populations. Population structure and genetic relationships are important for establishing the appropriate scale and subunits for species conservation and management [56]. The study also identified two genetic groups using NJ, Structure, and PCoA: one in northern Vietnam (DH, SD, BV, MC, YB, and TC populations), and the other in central Vietnam (BE, CP, and QH populations). The Mantel test shows that there are almost obvious geographical boundaries between the studied populations, and it indicated the effects of the geographic distance on the number of migrants per generation inferred from the studied loci. This suggests that past human activities can have a significant impact on the present population structure and genetic diversity of *C. balansae* in Vietnam. Thus, the best conservation strategy should be to establish an *ex-situ* conservation site for this species through collecting seeds, germinating, and cultivating seedlings. Collecting seeds from individuals of all the populations should be conducted so that random loss of genetic variability due to loss of individual ecotypes should be avoided and move towards new self-sustaining populations. The establishment of a new large population of *C. balansae* should provide new alleles, which might improve its fitness under different environmental stresses to prevent the gradual reduction in the population of *C. balansae*.

5 Conclusion

These results confirmed that *C. balansae* had low genetic diversity and high genetic differentiation among populations in the evergreen tropical forests. The nine populations of *C. balansae* were divided into two genetic groups related to the geographic distance. Our study will contribute to understanding the population history of this species and the genetic relationships among the *C. balansae* populations.

Author contributions: Cui B, Vu DD, Vu DG and Tran VH designed the study. Vu DD, Vu DG, Tran VH, Bui TTX, Pham MP, Nguyen MT and Nguyen VS collected samples in field works. Vu DD, Vu DG, Tran VH, Bui TTX, Pham MP, Nguyen MT implemented the molecular experiments. Cui B, Vu DD, Vu DG, Tran VH, Nguyen MT, Rahman SU and Shah SNM analyzed data and wrote the original draft of manuscript. All authors have read and approved the final manuscript.

Conflict of interest: Authors state no conflict of interest.

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

References

[1] MOST, VAST. Vietnam red data book, part II. Plants. Pub Sci Tech; 2007. p. 289–90.

[2]   Nguyen KD. Flora of Vietnam – Lauraceae. Sci Technics Publ House. 2010;20:22–234.

[3]   Tai VA, Thin NN. Survey and inventory of threatened plant species in Ha Giang province, Vietnam. J Biol. 2014;36(3):323–9.

[4]   Hai TN, Nghi DH, Phuong LD, Hoang TV. Study on the sylviculture characters of Cinnamomum balansae Lecomte at Ben En National Park. J Forestry Sci Technol. 2016;6:176–85.

[5]   Amos W, Harwood J. Factors affecting levels of genetic diversity in natural populations. Phil Trans R Soc Land B. 1998;353:177–86.

[6]   Keller LF, Waller DM. Inbreeding effects in wild population. Treds Ecol Evol. 2002;11:424–9.

[7]   IUCN. The IUCN red list of threatened species. In: World Conservation Monitoring Centre. *Cinnamomum balansae*. The IUCN Red List of Threatened Species; 1998. e.T35937A9969078. (accessed 8 May 2020).

[8]   Lande R. Extinction ricks from anthropogenic, ecological and genetic factors. In: Landweber LA, Dobson AP, editors. Genetic and extinction of species. Princeton, NJ: Princeton University; 1998. p. 1–22

[9]   Vu DD, Bui BTT, Nguyen MT, Vu DG, Nguyen MD, Bui VT, et al. Genetic diversity in two threatened species in Vietnam: *Taxus chinensis* and *Taxus wallichiana*. J Res. 2017;28(2):265–72.

[10] Vu DD, Bui TTX, Nguyen MD, Shah ANM, Vu DG, Zhang Y, et al. Genetic diversity and conservation of two threatened dipterocarps (Dipterocarpaceae) in southeast Vietnam. J Res. 2019;30(5):1821–31.

[11] Nguyen MT, Vu DD, Dang PH, Bui TTX, Nguyen HPL, Nguyen MD. Population genetic structure and demographic history of the dipterocarp species *Anisoptera costata* Korth revealed by microsatellite analysis. Planta. 2021;253:66.

[12] Petit RJ, Mousadik EA, Pons O. Identifying populations for conservation on the basis of genetic markers. Conserv Biol. 1998;12:844–55.

[13] Frankham R, Ballou JD, Briscoe DA. Introduction to conservation genetics. Vol. 190. Cambridge, UK V: Cambridge University Press; 2004. p. 385–6

[14] Lin TP, Cheng YP, Huang SG. Allozyme variation in four geographic areas of *Cinnamomum kanehirae*. J Heredity. 1997;88:433–8.

[15] Joy P, Maridass M. Inter species relationship of *Cinnamomum* species using RAPD marker analysis. Ethnobot Leafl. 2008;12:476–80.

[16] Abeysinghe PD, Samarajeewa NGCD, Li, G, Wijesinghe KGG. Preliminary investigation for the identification of Sri Lanka *Cinnamomum* species using randomly amplified polymorphic DNA (RAPD) and sequence related amplified polymorphic (SRAP). J Natn Sci Found Sri Lanka. 2014;42(3):201–8.

[17] Kuo DC, Lin CC, Ho KC, Cheng YB, Hwang SY, Lin TP. Two genetic divergence centers revealed by chloroplastic DNA variation in populations of *Cinnamomum kanehirae* Hay. Conserv Genet. 2010;11:803–12.

[18] Varshney RK, Graner A, Sprrell ME. Genic microsatellite markers in plants: features and applications. Trends Biotechnol. 2005;23:48–55.

[19] Guo R, Hou XY, Fang CL, An JY, Wu J, Qiu YX. Characterization and cross-species transferability of EST-SSR markers developed from the transcriptome of *Dysosma versipellis*

(Berberidaceae) and their application to population genetic stidues. Mol Breed. 2014;34:1733–46.

[20] Tam NM, Duy VD, Duc NM, Thanh TTV, Hien PD, Trang NTP, et al. Genetic variation and outcrossing rates of the endangered tropical species *Dipterocarpus dyeri*. J Tropical Sci. 2019;31(2):259–67.

[21] Tsuneki S, Mori K, Kaneko S, Isagi Y. Identification and characterization of eight microsatellite loci in *Machilus pseudokobu* (Lauraceae), an endemic species of the Bonin Islands. Conserv Genet. 2009;10(6):2009–11.

[22] Kameyama Y. Development of microsatellite markers for *Cinnamomum camphora* (Lauraceae). Am J Bot. 2012;99:e1–3.

[23] Chen LY, Cao YN, Yuan N, Nakamura K, Wang GM, Qiu YX. Characterization of transcriptome and development of novel EST-SSR markers based on next-generation sequencing technology in *Neolitsea sericea* (Lauraceae) endemic to East Asian land-bridge islands. Mol Breed. 2015;35:187.

[24] Martins EM, Lamont RW, Martinelli G, Medeiros FL, Quinet A, Shapcott A. Genetic diversity and population genetic structure in three threatened *Ocotea* species (Lauraceae) from Brazil's atlantic rainforest and implications for their conservation. Conserv Genet. 2015;16:1–14.

[25] Zhu S, Ding Y, Yap Z, Qiu Y. De novo assembly and characterization of the floral transcriptome of an economically important tree species, *Lindera glauca* (Lauraceae), including the development of EST-SSR markers for population genetics. Mol Biol Rep. 2016;43:1243–50.

[26] Chen CH, Zheng YJ, Liu SA, Zhong YD, Wu Y, Li J, et al. The complete chloroplast genome of *Cinnamomum camphora* and its comparison with related Lauraceae species. Peer-Rev J. 2017;5:e3820.

[27] Kameyama Y, Furumichi J, Li J, Tseng YH. Natural genetic differentiation and human-mediated gene flow: the spatiotemporal tendency observed in a long-lived *Cinnamomum camphora* (Lauraceae) tree. Tree Genet Genomes. 2017;13:38.

[28] Li Z, Zhong Y, Yu F, Xu M. Novel SSR marker development and genetic diversity analysis of *Cinnamomum camphora* based on transcriptome sequencing. Plant Genet Resour. 2018;16(6):568–71.

[29] Zhong Y, Yang A, Li Z, Zhang H, Liu L, Wu Z, et al. Genetic diversity and population genetic structure of *Cinnamomum camphora* in South China revealed by EST-SSR Markers. Forests. 2019;10(11):1019.

[30] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

[31] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644.

[32] Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, et al. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics. 2003;19(5):651–2.

[33] Clarke KR, Gorley RN. Primer (Plymouth Routines in Multivariate Ecological Research) v5: User Manual. Tutorial. PRIMER-E Ltd Plymouth. 2001. p. 91.

[34] Peakall R, Smouse PE. GenAlEx 6.5: genetic analysis in excel. Population genetic software for teaching and research an update. Bioinformatics. 2012;28:2537–9.

[35] Goudet J. FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3); 2001. http://www.unil.ch/izea/softwares/fstat.html.

[36] Rousset F. Genepop'007: a complete re-implementation genepop Softw Windows Linux. Mol Ecol Resour. 2008;8:103–6. doi: 10.1111/j.1471-8286.2007.01931.x

[37] Excoffier L, Laval G, Schneider S. Arlequin v.3.0. An integrated software package for population genetics data analysis. Evolut Bioinforma Online. 2005;1:47–50.

[38] Piry S, Luikart G, Cornnet JM. Bottleneck: a computer program for detecting recent reductions in the effective population size frequency data. J Heredity. 1999;90:502–3.

[39] Hedrick PW. A standardized genetic differentiation measure. Evolution. 2005;59(8):1633–8.

[40] Takezaki N, Nei M, Tamura K. Software for constructing population trees from allele frequency data and computing other population statistics with Windows interface. Mol Evolution. 2010;27:747–52.

[41] Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59.

[42] Earl DA, Von-Honldt BM. Structure Harvester: a website and program for visualizing structure output and implementing the Evanno method. Conserv Genet Resour. 2012;4:359–61.

[43] Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software structure: a simulaton study. Mol Ecol. 2005;14:2611–20.

[44] Hamrick JL, Godt MJW. Effects of history traits on genetic diversity in plant species. Phil Trans Roy Soc Lond B Biol Sci. 1996;351:1683–5.

[45] Lammi A, Siikamaki P, Mustajarvi K. Genetic diversity, population size, and fitness in central and peripheral populations of a rare plant Lychnis viscaria. Conserv Biol. 1999;13:1069–78.

[46] White TL, Adams WT, Neale DB. Forest genetics. Boston, MA, USA: CABI Publshing; 2007. p. 149–86.

[47] Ferrer M, Eguiarte L, Montana C. Genetic structure and outcrossing rates in *Flourensia cernua* (Asteraceae) growing at different densities in the south-western Chihuahuan Desert. Ann Bot. 2004;94:419–26.

[48] Dick C, Hardy O, Jones FA, Petit R. Spatial scales of pollen and seed-mediated gene flow in tropical rain forest trees. Trop Plant Biol. 2008;1:20–33.

[49] Dervishi A, Jakše J, Ismaili H, Javornik B, Štajner N. Comparative assessment of genetic diversity in Albanian olive (*Olea europaea* L.) using SSRs from anonymous and transcribed genomic regions. Tree Genet Genomes. 2018;14:53.

[50] Boaventura-Novaes CRD, Novaes E, Mota EES, Telles MPC, Coelho ASG, Chaves LJ. Genetic drift and uniform selection shape evolution of most traits in *Eugenia dysenterica* DC. (Myrtaceae). Tree Genet Genomes. 2018;14:76.

[51] Vu DD, Bui TTX, Nguyen VS, Pham MP, Ta TTH, Nguyen V, et al. Evaluation of genetic diversity and a bottleneck of *Cinnamomum balansae* H. Lecomte populations in Phu Tho province by microsatelite (SSR) markers. J Forestry Sci Technol. 2021;2021(1):3–9. (Vietnamese, abstract by English).

[52] Hartvig I, So T, Changtragoon S, Tran HT, Bouamanivong S, Theilade I, et al. Population genetic structure of the endemic rosewoods *Dalbergia cochinchinensis* and *D. oliveri* at a regional scale reflects the Indochinese landscape and life-history traits. Ecol Evol. 2018;8:530–45.

[53] Tan J, Zhao ZG, Guo JJ, Wang CS, Zeng J. Genetic diversity and population genetic structure of *Erythrophleum fordii* Oliv., an endangered rosewood species in South China. Forests. 2018;9:636.

[54] Fukue Y, Kado T, Lee SL, Ng KK, Muhammad N, Tsumura Y. Effects of flowering tree density on the mating system and gene flow in *Shorea leprosula* (Dipterocarpaceae) in Peninsular Malaysia. J Plant Res. 2007;120:413–20.

[55] Chan HT. Reproductive biology of some Malaysian dipterocarps II. Fruiting biology and seedling studies. Malay Forester. 1980;48:438–51.

[56] Ferriere R, Dieckmann U, Couvet D. Evolutionary conservation biology. Cambridge, United Kingdom: Press Syndicate of the University of Cambridge; 2004. p. 56.