GMR The Original

Transcriptome-Derived SSR Marker Development for Genetic Insights into Dipterocarpus alatus Roxb. ex G.Don, a Vulnerable and Ecologically Significant Species

Preeya P. Wangsomnuk[1], Xin Chun Mo[2], Manop Poopath[3]

[1]Department of Biology, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand, 40002. prepua@kku.ac.th
[2]Yunnan Economics Trade and Foreign Affairs College, Kunming 651700, China. ljnumxc@163.com
[3]Forest Herbarium, Department of National Parks, Wildlife and Plant Conservation, Chatuchak, Bangkok, 10900, Thailand. poomanop@hotmail.com

Corresponding author: P.P. Wangsomnuk; E-mail: prepua@kku.ac.th

ABSTRACT

Dipterocarpus alatus Roxb. ex G.Don is a tropical tree species of significant ecological and economic importance that is increasingly threatened by habitat fragmentation and overexploitation. However, there are no efficient tools to estimate the population structure and genetic diversity of D. alatus. Here, we employed a simple sequence repeat (SSR) marker development method derived from the leaf transcriptome sequences and validated for the analysis of genetic diversity and population structure. Over 48 million clean reads were generated from the transcriptome of the young leaf tissue. A total of 12,514 SSR loci were identified from the assembled transcripts. Among them, mononucleotide repeats were most abundant (75.32%), followed by trinucleotide (15.76%), dinucleotide (7.89%), tetranucleotide (0.81%), hexanucleotide (0.13%), and pentanucleotide (0.09%) motifs. Although mononucleotide repeats were predominant, only di-, tri, and tetranucleotide motifs were selected for marker development due to their higher suitability. 61 loci were chosen to screen the polymorphic information, and 26 of them yielded reproducible amplification across 30 individuals from five D. alatus populations. These markers exhibited polymorphism information content (PIC) values ranging from 0.315 to 0.410, with an average discriminatory power (Dp = 0.385), indicating that these markers are informative and suitable for assessing genetic variation among individuals. Trinucleotide repeats, often located in coding regions, demonstrated high success rates in primer design. Genetic diversity analysis revealed observed heterozygosity (Ho) ranging from 0.000 to 0.823 and expected heterozygosity (He) from 0.036 to 0.770. Population structure analysis revealed three genetically distinct clusters that largely corresponded to geographic origins. The presence of admixed individuals and within-cluster substructuring suggests ongoing gene flow and complex genetic relationships among populations. This study presents the first set of transcriptome-based genic SSR markers for D. alatus, offering valuable molecular tools for conservation genetics and sustainable forest resource management.

Keywords: *Dipterocarpus alatus; genic SSR markers, Transcriptome sequencing; Genetic diversity; Population structure; Vulnerable species; Conservation genetics*

INTRODUCTION

Dipterocarpus alatus Roxb. ex G.Don, which is mainly found near rivers in densely populated forests, is a member of the Dipterocarpaceae family with a chromosome number of 2n = 22 (Roy and Jha, 1965). This tree can reach up to 60 meters and is well known in countries such as Thailand, Myanmar, Vietnam, India, and Sri Lanka, due to its multi-purpose uses, especially in timber production (Boontawee, 2001; Ly et al., 2023). It is also a source of oleoresin, which has traditionally been used in various local applications (Dyrmose et al., 2017).

Better plantation management is not very effective, as *D. alatus* is a slow-growing species, which necessitates a significant amount of time, effort, and financial investment for the establishment of managed plantations. Although *D. alatus* is valued for its ecological functions and economic uses, its populations have suffered noticeable declines. The species is currently classified as Vulnerable on the IUCN Red List under criteria A2cd, a status that highlights its susceptibility to habitat degradation and overexploitation by humans (Khoo et al., 2022; Ly et al., 2023). In contrast to rapidly growing species like *Acacia* and *Eucalyptus*, establishing dipterocarp plantations is a more complex process. Wetland deforestation related to agricultural expansion has a direct, rapid, and devastating impact on species by removing the natural habitats necessary for their survival. This situation is exacerbated by the expansion of resin-tapping procedures, which include tree trunk burning, logging, and forest fires.

All kinds of species are preconditioned for evolutionary alteration due to their genetic diversity; if this genetic diversity is missing within the population, they will be unable to respond positively to novel environmental changes, such as global warming. Correspondingly, the adverse effects of population inbreeding can be diminished if genetic diversity levels are high, providing nature a route to adapt to environmental change (Hoban et al., 2021). Thailand has followed other nations in experiencing a decline in *D. alatus* genetic diversity, which is attributed to several factors. One of these factors is the negative impact of global warming on the natural habitat, affecting the development patterns, germination, and seed dissemination of *D. alatus*, as well as other tree species. Other factors include deforestation that supports the construction and agricultural industries, as well as population growth, which stimulates urban development. The consequence is a significant reduction in the *D. alatus* population size, and more importantly, the tree's ability to connect within its natural habitat has been put at risk, thus challenging the tree's ability to flourish through natural genetic diversity, potentially putting at risk the mature reproductive capability of the species (IPBES, 2019).

Due to their codominant inheritance and high levels of polymorphism, SSRs are considered highly efficient genetic markers (Luo et al., 2024). Only a few studies have been conducted on dipterocarp species. For example, using a limited genomic library constructed with the M13 phage vector, dinucleotide microsatellites containing GT, CA, and CT repeat motifs were isolated from *Dryobalanops lanceolata*. Among the six microsatellite loci sequenced, only one locus, DL(GT)202, was deemed suitable for further analysis (Terauchi, 1994). Eight vectorette PCR reactions and one colony hybridization method collectively generated nine polymorphic SSR markers for *Shorea curtisii* (Ujino et al., 1998). Including repeats such as $(CT)_8$, $(CT)_{12}$, and $(CT)_{16}$, the markers primarily consisted of simple dinucleotide motifs. By utilizing genomic DNA fragments, ten dinucleotide SSR markers targeting (GA)n and (CA)n repeats were developed for *Dipterocarpus tempehes* (Isagi et al., 2002). When tested on 34 adult *D. tempehes* trees, these markers displayed 7 to 14 alleles per locus, with an average of 10.3 alleles, thereby demonstrating a high level of polymorphism. More recently, twelve polymorphic genomic SSR markers were successfully developed

using Illumina Paired-end sequencing technology for *Hopea hainanensis* (Wang et al., 2020). Of these three, seven primers can be transferred to *Hopea chinensis* Hand. -Mazz. and *Hopea reticulata* Tardieu. Although EST-SSR markers have proven effective for estimating the genetic diversity for particular endangered plant species, including 193 plant genera registered in the IUCN Red List using dbEST data (NCBI) (Lopez et al., 2015), suitable loci for *D. alatus* remain limited due to inadequate access to transcriptome sequences. To overcome and enhance their application in studying the genetic diversity and population structure of *D. alatus* and related species, EST-SSR markers must be developed. The advent of high-throughput sequencing technologies, particularly RNA-Seq using next-generation sequencing (NGS), revolutionized the identification and utilization of SSR markers by enabling the rapid acquisition of extensive expressed sequence data. Transcriptome sequencing not only facilitated the classification of genes of interest and molecular functional markers, including SSRs, but also significantly improved the efficiency and accuracy of SSR marker identification, especially for non-model organisms with limited genomic resources. Previous research has established the transmissibility of EST-SSR markers to associated species (Xu et al., 2024), emphasizing their prospective uses in conservation genetics and population structure.

In this study, we employed de novo transcriptome sequencing of the young leaf tissue of *D. alatus* and screened for SSR loci within the assembled transcripts. The entire set of selected SSR loci was further identified within five *D. alatus* populations collected from Thailand to evaluate the efficiency of the developed SSR markers. The findings highlight the efficacy of transcriptomic-derived SSR markers in capturing genetic diversity within the Dipterocarpaceae family. Nevertheless, the limited number of markers utilized in earlier studies underscores the need for further comprehensive research to improve the conservation of *D. alatus* and other dipterocarp species.

MATERIAL AND METHODS

Plant Material

Inner bark or leaf tissues from 30 individuals of *D. alatus* were collected for DNA extraction from five distinct populations. The first to the forth populations collected from naturally growing trees in Pho Tak District, Nong Khai Province (PT; samples S1 to S6; 17°51'06.2"N 102°27'30.1"E), Mueang Nakhon Sawan District, Nakhon Sawan Province (MNS; samples S7 to S12; 15°43'49.2"N 100°10'48.0" E), Phen District, Udon Thani Province (PH; samples S13 to S16; 17°34'51.3"N 103°00'49.7" E), In Buri District, Sing Buri Province (IB; samples S17 to S20; 14°59'45.3"N 100°16'37.2" E), and the last population consisted of 10 individuals (Samples S21 to S30) randomly selected from approximately 200–300 accessions cultivated under the ex-situ conservation project at Khon Kaen University, Mueang Khon Kaen District, Khon Kaen Province (MKK). This project spans approximately 9,600 m² and is located within the *D. alatus* Genetic Conservation Area. Leaf samples were randomly collected from trees spaced 200–500 meters apart within the geographical range of 16°27'32"N 102°48'56"E.

For RNA extraction, young leaves of *D. alatus* were collected from the Genetic Conservation Area at Khon Kaen University, immediately cryopreserved in liquid nitrogen, and stored at -80°C until use. For DNA extraction and validation of expressed sequence tag-simple sequence repeat (eSSR) markers, inner bark or leaf tissues were preserved in plastic bags containing silica gel in the field. Then they were transported to the Laboratory of the Department of Biology at Khon Kaen University and stored at -20°C until further processing.

Genomic DNA Isolation

Genomic DNA was extracted from 0.10 g of dried leaf or bark tissue per individual, following the extraction protocol described by Wangsomnuk et al (2014). DNA pellets were air-dried at 40°C for 20 minutes and then resuspended in 100 µL of TE buffer. The concentration and purity of the DNA were measured using a

NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific Inc., MA, USA). DNA integrity was verified through electrophoresis on a 1% LE agarose gel (SBIO, Product No. SB102-100), stained with RedSafe (20,000× nucleic acid staining solution, Cat. No. 21141, iNtRON Biotechnology), and visualized under UV light.

RNA Extraction

Total RNA was extracted from young leaves of *Dipterocarpus alatus* collected from several individuals. Equal amounts of RNA from each tree were pooled to obtain bulk RNA for transcriptome sequencing. Extraction was performed using the NucleoSpin® RNA kit (MACHEREY-NAGEL) following the manufacturer's instructions, with an additional treatment of RNase-free DNase I at 37 °C for 30 min to remove genomic DNA contamination. RNA quality was initially assessed by agarose gel electrophoresis, and concentration was determined using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA). Integrity was further assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), with RIN values above 7.0 and A260/280 ratios of 1.9–2.1 confirming suitability for downstream analysis. The bulk RNA sample was then sent to Macrogen Inc. (Seoul, South Korea) for cDNA library preparation and Illumina sequencing.

cDNA Library Construction and Sequencing

RNA library preparation was conducted using the TruSeq RNA Sample Preparation Kit v2 (Illumina), following the manufacturer's protocol. Fragmented polyadenylated RNA was reverse-transcribed into first-strand cDNA using random hexamer primers and reverse transcriptase. Subsequently, second-strand cDNA synthesis was carried out to obtain double-stranded cDNA. The resulting cDNA fragments were purified using AMPure XP beads (Beckman Coulter).

End repair was performed to generate blunt-ended fragments, adding a single 'A' nucleotide at the 3' ends to facilitate adapter ligation. Indexing adapters were ligated to the cDNA fragments, and the adapter-ligated library was enriched through PCR amplification with a minimal number of cycles to maintain transcriptome complexity. The final libraries were assessed for quality and concentration using appropriate validation tools before sequencing. The sequencing process was carried out on the HiSeq 2500 System (Illumina) using paired-end sequencing to generate reads of 101 base pairs. The system employed the TruSeq rapid SBS kit or TruSeq SBS Kit v4, with protocols outlined in the HiSeq 2500 System User Guide (Part #15011190 Rev. V) and sequencing control software HCS v2.2.70.

***De novo* Assembly**

The quality of the raw sequencing data was evaluated using FastQC version 0.11.9 http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Following this, low-quality reads and adapter sequences were removed using Trimmomatic (version 0.32; Bolger et al., 2014). Processed high-quality reads were subjected to transcriptome assembly using Trinity (r20140717; Haas et al., 2013) to facilitate efficient *de novo* assembly of RNA-Seq data into transcript contigs without a reference genome. BLASTx version 2.13.0 (https://blast.ncbi.nlm.nih.gov/Blast.cgi) against a protein sequence database (go_v20150407) was used to annotate the assembled transcripts (Altschul et al., 1990). Subsequently, the transcripts were further categorized by Blast2GO software (version 2.5; Conesa et al., 2005) with default parameters for functional annotation (http://www.geneontology.org/).

SSRs Identification

Simple sequence repeats (SSRs) were identified using the MISA software (MIcroSAtellite, version 1.0; http://pgrc.ipk-gatersleben.de/misa) following the default settings described by Thiel et al. (2003). Repeat thresholds were defined as follows: mononucleotide motifs ≥10, dinucleotide motifs ≥6, and trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide motifs ≥5.

Evaluation of the Efficiency of Developed SSR Markers in the *D. alatus Population*

Primers of selected SSR loci were generated by using Premier 3.0 software (Untergasser et al., 2012). The design criteria ensured that primers were long, ranging from 20 to 24 base pairs, with an optimal GC content between 40% and 60%. The melting temperature (T_m) was kept in the range of 54°C to 60°C, and the predicted amplicon range was planned to be between 100 and 300 base pairs.

The efficiency of primers was assessed through PCR amplification. The PCR process was executed using a 10 µL reaction volume containing 50 ng of diluted template DNA, 1 µL of 10X PCR buffer (comprising 750 mM Tris-HCl, 200 mM $(NH_4)_2SO_4$, and 0.1% Tween20), 1 µL of 2 mM dNTPs, 3.32 µL of sterile distilled water, 2 µL of 5 µM primers, and 0.08 µL of Taq DNA polymerase (5 units/µL, Vivantis). A thermal cycler was used for amplification with the following settings: an initial denaturation step was conducted at 94°C for 1 minute, followed by 35 cycles, each including a denaturation step at 94°C for 1 minute, with annealing at 50–60°C for 40 seconds and an extension phase at 72°C for 1 minute and 30 seconds, finalizing with an extension at 72°C for 5 minutes.

The resulting PCR fragments from 61 primer pairs across five randomly selected genotypes representing five populations were analyzed using gel electrophoresis on a 4% high-resolution Smart Agarose gel (SBIO, Smart Science Co., Ltd.), prepared with TBE buffer (0.5–1X) and stained with RedSafe (20,000x nucleic acid staining solution, Cat. No: 21141, iNtRON Biotechnology). Smart Agarose is renowned for its high resolution in detecting small allelic variations, making it an ideal choice for separating polymorphic alleles. Band patterns were visualized under UV light using a UVITEC gel documentation system (UVITEC Cambridge).) To validate the PCR product against the anticipated sizes based on primer design, the products were evaluated using a DNA marker (VC 100 bp DNA Ladder Plus, Vivantis).

Primers that produced the anticipated results were used to amplify the DNA of 30 D. alatus individuals across five populations. The ensuing PCR products were analyzed to determine allele sizes, numbers, and polymorphisms, as indicated by their ability to amplify varying-sized alleles across all genotypes. Each SSR marker-produced band profile was scored by hand to generate two dataset types: binary data and allelic data. Binary data, in which "1" denotes the presence of a DNA band and "0" denotes the absence, were utilized to determine the polymorphism information content (PIC), and discrimination power (Dp) using the iMEC online marker efficiency calculator (Amiryousefi et al., 2018).

Genetic Diversity Analysis

Allelic data were generated by assigning estimated allele sizes to each DNA band based on a DNA ladder, further refined according to the repeat motifs of the primers. These allele sizes were analyzed in GenAlEx v6.5 software (Peakall and Smouse, 2006, 2012) to determine key genetic diversity parameters, including the number of different alleles (Na), effective alleles (Ne), observed heterozygosity (Ho), and expected heterozygosity (He). A dissimilarity matrix was computed from the binary data, and the cophenetic correlation coefficient (CCC) was employed to select the most suitable dissimilarity matrix. Subsequently, the heat map from the Jaccard Dissimilarity matrix (Jaccard, 1901) was generated using the R software environment (R Core Team, 2024) to visualize the genetic relationships among the genotypes. The same matrix was imported into MEGA X (Kumar et al., 2018) to construct a neighbor-joining (NJ) dendrogram.

Furthermore, an examination of the population composition and dynamics was conducted to understand the population structure. A Bayesian clustering methodology was employed utilizing STRUCTURE v2.3.4 (Pritchard et al., 2000). The analysis was performed with a burn-in period of 100,000 iterations followed by 100,000 MCMC (Markov Chain Monte Carlo) replications, assuming an admixture model and correlated

allele frequencies. The number of clusters (K) was tested from K = 1 to K = 10, with 15 independent runs performed at each K value. The optimal K was determined using the ΔK method (Evanno et al., 2005), as implemented in StructureSelector (Li and Liu, 2018). The clustering results were visualized using STRUCTURE v2.3.4 (Pritchard et al., 2000) to generate a bar plot representing individual genetic ancestry.

RESULTS

De novo assembly of the transcriptome

We generated a total of 48,901,612 reads (~4.939 Gb) from the leaf transcriptome of *D. alatus* with a GC content of 48.31%. After quality filtering and adapter trimming, 48,593,374 high-quality reads (~4.887 Gb) were obtained, with improved Q20 and Q30 values of 98.08% and 96.52%, respectively (Table 1). These high-quality reads were then used for *de novo* assembly.

Table 1. Overview of sequencing data statistics derived from the leaf transcriptome of *D. alatus*.

Data Type	Total Bases	Read Count	GC (%)	Q20 (%)	Q30 (%)
Raw Data	4,939,062,812	48,901,612	48.31	97.73	96.09
Trimmed Data	4,887,318,889	48,593,374	48.28	98.08	96.52

A total of 50,422 Trinity genes and 66,281 transcripts were generated by de novo assembly using the Trinity software, with a GC content of 42.69%. The N50 value was 1,397 bp, and the total assembled bases amounted to 55,837,442 bp. When only the longest isoform per gene was retained, the number of transcripts decreased to 50,422, with a slight increase in GC content to 42.91% and a decrease in N50 value to 1,289 bp (Table 2). These assembled transcripts formed the foundation for identifying genic SSR loci for marker development.

Table 2. Summary of assembly statistics derived from the leaf transcriptome of *D. alatus*.

Metric	ALL transcript contigs	Only the longest isoform per 'GENE'
Total trinity 'genes'	50,422	50,422
Total Trinity transcripts	66281	50,422
GC contents (%)	42.69	42.91
N50 (bp)	1,397	1,289
Average contig length (bp)	842	750
Total assembled bases	55,837,442	37,820,793

Functional Annotation

Functional annotation and GO analysis were performed to facilitate the functional classification of the transcripts. Over one-third of the transcripts (33.5%) showed no hits in the database, suggesting that these transcripts may represent novel or poorly characterized genes. Among the annotated sequences, the biological process category was the most predominant (26.5%). This category includes genes involved in metabolism, signaling, and various cellular activities. Additionally, the cellular component category represented 21.3% of the annotated sequences, which describes the physical locations of gene products, such as organelles or membranes. The molecular function category accounted for 18.7% of the sequences, encompassing specific biochemical activities such as enzyme binding or catalytic functions (Figure 1).

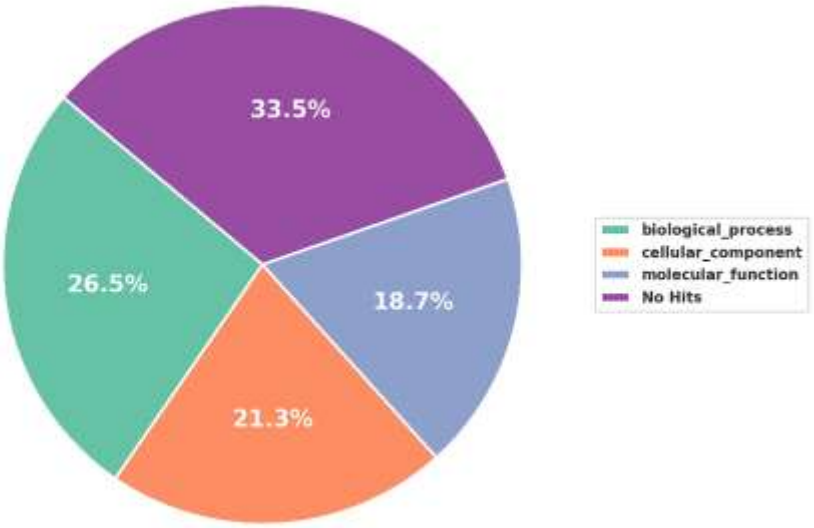


Figure 1. GO functional classifications of *D. alatus* transcripts. Different colors were used to indicate each GO category: biological_process (green), cellular_component (orange), molecular_function (blue), and **No Hits (purple).**

In total, 22,660 unigenes were classified into 20 subcategories within the biological process domain (Figure 2). The most abundant unigenes were grouped into metabolic processes (4,286 genes, 18.9%), followed by unclassified biological processes (2,666 genes, 11.8%) and biological regulation (2,449 genes, 10.8%). Some moderate amount of unigenes were categoried into these functional groups included response to stimulus (2,286 genes, 10.1%), cellular processes (1,502 genes, 6.6%), single-organism processes (1,498 genes, 6.6%), localization (1,091 genes, 4.8%), developmental processes (642 genes, 2.8%), and cellular component organization or biogenesis (560 genes, 2.5%). A very small number of unigenes were grouped into these categories, such as reproductive processes (169 genes, 0.75%) and hormone secretion (0.03 genes, 0.0001%), suggesting that reproductive and endocrine-related transcripts might be expressed at lower levels in the leaf tissue or that transcriptomic coverage of these functions is inherently limited.

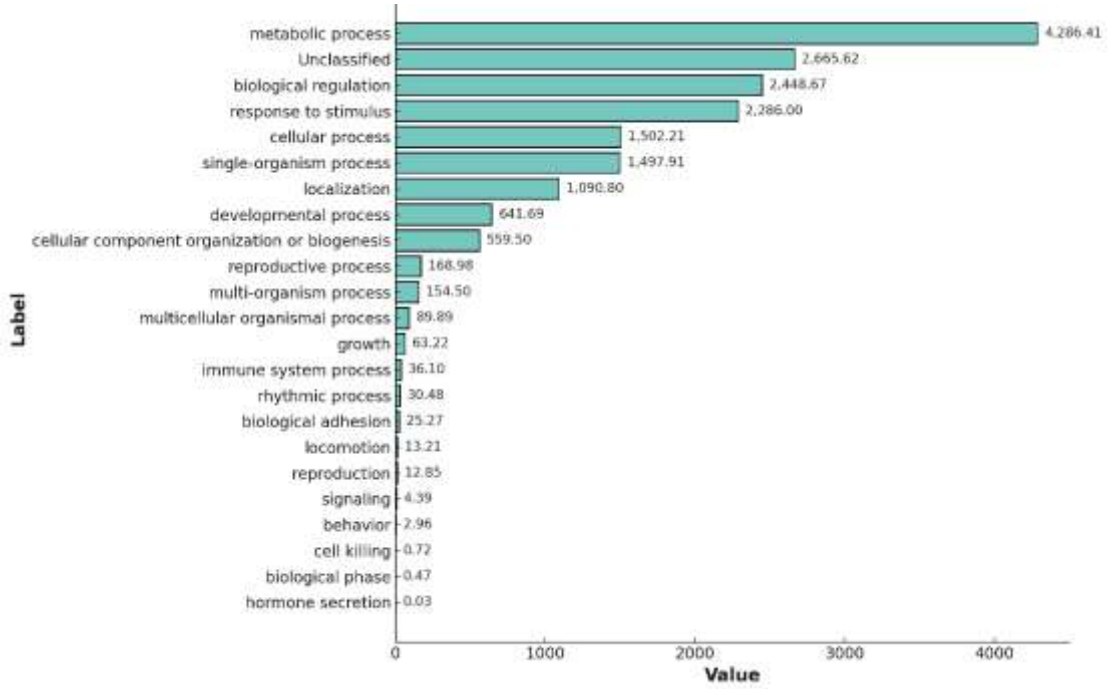


Figure 2. Classification of unigenes within the biological process subcategories. The horizontal axis represented the value (number of genes), while the vertical axis listed the biological process categories.

Additionally, 15,948 unigenes were grouped into 18 cellular component (CC) groups, providing insight into the structural and functional organization within the cell (**Figure 3)**. In this category, the most predominant subcategory was "cell part," with 6,180 unigenes (38.75%), followed by 3,542 unigenes (22.21%) in "organelle" and 1,029 unigenes (6.45%) in "membrane". Other notable subcategories included "membrane part" (885 unigenes, 5.55%) and "extracellular region" (554 unigenes, 3.47%), highlighting key structural components. Smaller but biologically significant subcategories, such as "organelle part" (469 unigenes, 2.94%) and "cell junction" (361 unigenes, 2.26%), played an essential role in cellular connectivity and organization. Similarly, categories such as "macromolecular complex" (310 unigenes, 1.94%) and "cell" (237 unigenes, 1.48%) illustrated the structural complexity of the cellular environment. Less common subcategories, including "collagen trimer" (1 unigene, 0.01%), might represent specialized or niche functions within the transcriptome.

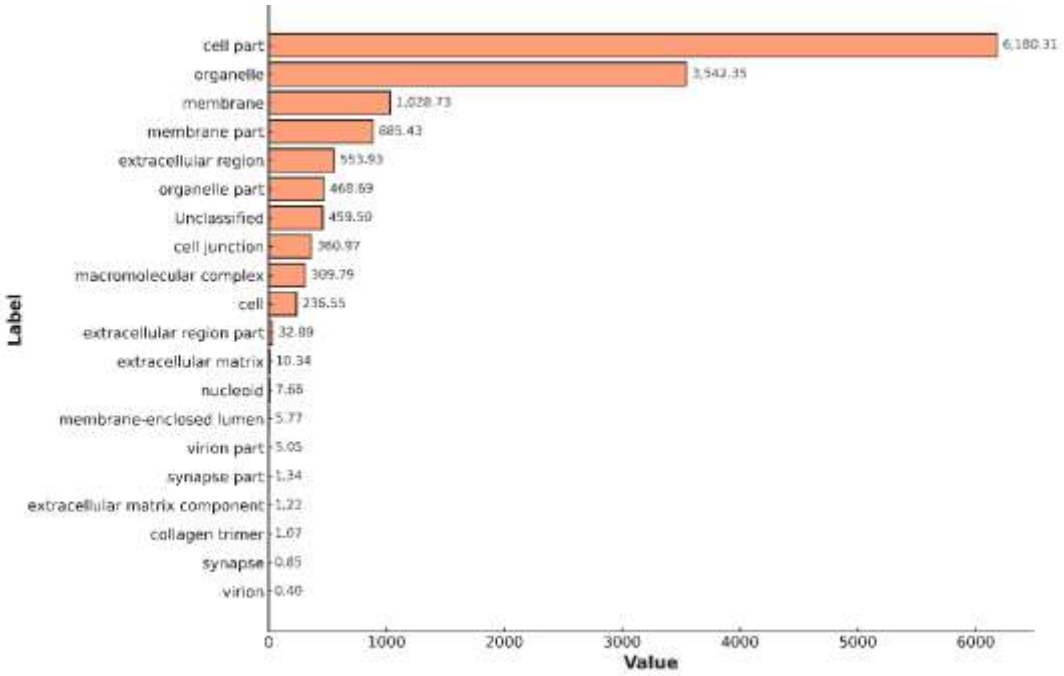


Figure 3. Classification of identified unigenes into cellular component (CC) subcategories. The horizontal axis represented the number of genes, while the vertical axis listed the cellular component categories.

In the subcategory molecular function (Figure 4), the predominant group was the binding group with 4,844 unigenes involved, followed by catalytic activity (4,016 unigenes) and the unclassified category (2,192 unigenes). Three groups, such as transporter activity (499 unigenes involved), nucleic acid binding transcription factor activity (284 unigenes involved), and molecular transducer activity (237 unigenes involved), contributed to the overall molecular function composition, followed by the structural molecule activity (139 unigenes involved) and molecular function regulator (37 unigenes involved). Low-representation categories, such as enzyme regulator activity (30 unigenes), protein binding transcription factor activity (26 unigenes), and antioxidant activity (12 unigenes), indicated specialized molecular functions were found in the leaf. The least represented functions found in the transcriptome include translation regulator activity (represented by only one unigene), metallochaperone activity, and other niche molecular functions.

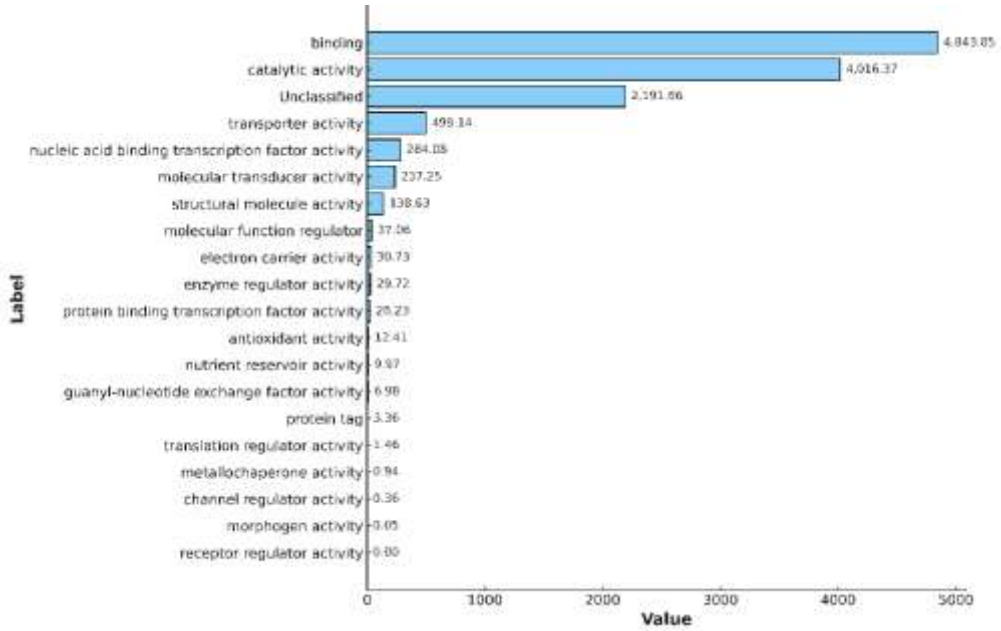


Figure 4. Classification of identified unigenes into molecular function (MF) subcategories. The horizontal axis represented the number of genes, while the vertical axis listed the molecular function categories.

SSR identification in the leaf transcriptome of *D. alatus*

A total of 12,514 SSRs were identified from the de novo assembled transcriptome of *D. alatus* leaf tissue, highlighting the substantial prevalence of simple sequence repeats (Table 3). Among the analyzed sequences, 10,726 sequences were found to possess SSRs, accounting for approximately 16.18% of all sequences. Additionally, 1,518 sequences were found to possess more than one SSR, representing 14.15% of the SSR-containing sequences. Among the estimated SSRs, 678 SSRs were found to have a compound formation, constituting about 5.42% of the total SSRs. Further analysis of SSR motifs was conducted. Results showed that the distribution of SSR motifs in the leaf transcriptome followed this order: mononucleotide (75.32%) > trinucleotide (15.76%) > dinucleotide (7.89%) > tetranucleotide (0.81%) > hexanucleotide (0.13%) > pentanucleotide (0.09%) (Table 3 and Supplementary Table S1).

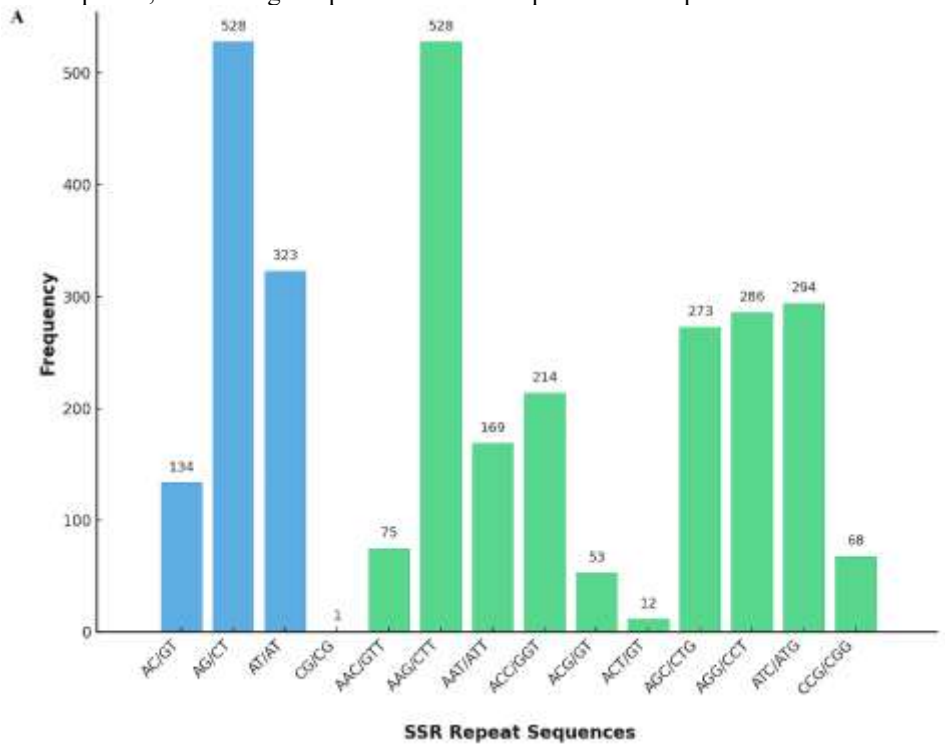
Table 3. Characteristics of SSRs identified from the leaf transcriptome of *D. alatus*.

SSR Characteristics	Count
Total number of sequences examined	66,281
Total size of examined sequences (bp)	55,837,442
Total number of identified SSRs	12,514
Number of SSR-containing sequences	10,726 (16.18%)
The number of sequences containing more than 1 SSR	1,518 (14.15%)
Number of SSRs present in the compound formation	678 (5.42% of total SSRs)
Mononucleotide repeat	9427 (75.32%)
Dinucleotide repeat	986 (7.89%)
Trinucleotide repeat	1972 (15.76%)
Tetranucleotide repeat	102 (0.81%)

Pentanucleotide repeat	11 (0.09%)
Hexanucleotide repeat	16 (0.13%)

Characteristics of SSRs, Motifs, and Frequency Distribution

To further calculate the mononucleotide repeats within identified SSR motifs, the nucleotide repeat frequencies of estimated SSRs were analyzed, and four types of repeats, such as (A)n, (C)n, (G)n, and (T)n, were calculated. Results showed that adenine (A) and thymine (T) repeats were the predominant mononucleotides, occurring at least 10 times the frequency. The adenine mononucleotide repeat ranged from 10 to 23 iterations, with 4,846 occurrences, making it the most frequent repeat type. In single conversion, A/T mononucleotide repeats were found to appear 9,289 times. Regarding other types of SSR motifs (Figure 5), the most frequent dinucleotide motif was AG/CT, with 528 occurrences, followed by AT/TA (323 occurrences), while CG/GC was the rarest, appearing only once. Among trinucleotide motifs, AAG/CTT was the most abundant, with 528 occurrences, followed by ACC/GGT with 294 occurrences and ATC/ATG with 286 occurrences. The most frequent tetranucleotide motif was AGAT/CTAT, with 33 occurrences, while other tetranucleotide motifs ranged from 1 to 10 occurrences. Pentanucleotide and hexanucleotide motifs were the least, occurring one to three times. The distribution of repeat numbers for di- and trinucleotides is shown in Figure 5 (see Supplementary Table S1 for details). The decline in frequency at higher repeat numbers likely reflects lower stability and/or selection against long repeats in leaf transcripts. Six-repeat dinucleotide SSRs were common, indicating relative stability and less susceptibility to mutations or deletions. In contrast, the sharp drop in frequency of longer repeats suggested greater instability or selective pressures that may affect transcript levels in the leaf of *D. alatus*. Similarly, the predominance of five-repeat trinucleotide SSRs pointed to evolutionary stability in the *D. alatus* transcriptome, where longer repeats were less frequent due to replication constraints.



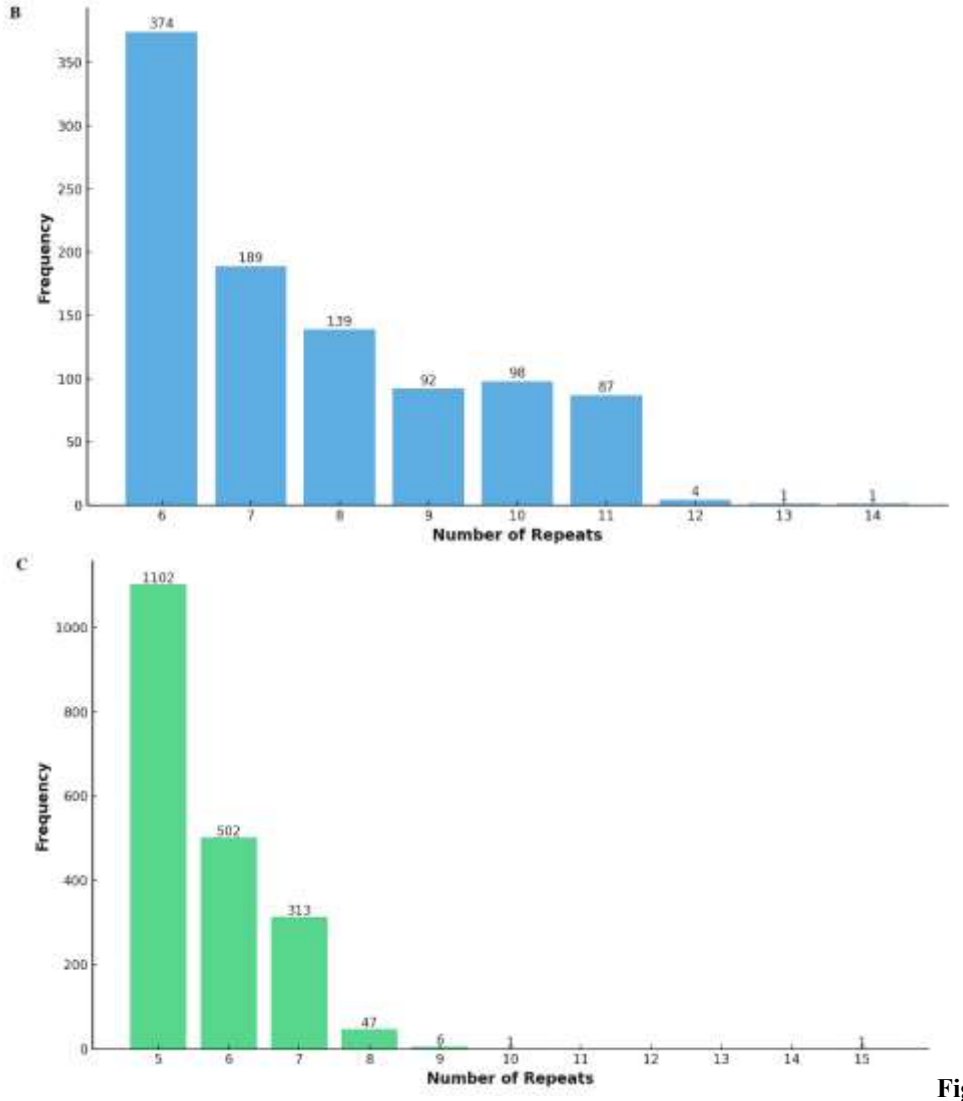


Figure 5.
(A,B,C) Frequency distribution of repeat numbers for SSR motifs. A. Distribution of di-, and trinucleotide repeats among the identified SSRs; B–C. Distribution of dinucleotide and trinucleotide repeats, respectively.

Among the 12,514 identified SSRs, primer pairs were designed for 6,115 loci, achieving a success rate of 48.87%. Mononucleotide repeats were the most abundant, with 7,481 occurrences, of which primers were designed for 4,005 (53.54%). Dinucleotide repeats were less frequent, comprising 797 entries, with 537 (67.38%) successfully designed primers. Trinucleotide repeats, commonly found in coding regions, were identified in 1,619 occurrences, with 1,193 (73.69%) primers designed. Tetranucleotide repeats were the least frequent, with only 84 occurrences, and primers were successfully designed for 41 (48.81%). Complex or compound repeats accounted for 579 occurrences, with 339 (58.55%) primer pairs successfully designed.

The reliability of the developed EST-SSR primers was evaluated using 30 plant accessions from five populations. Of the tested primers, 26 pairs showed clear polymorphisms and were consistently

reproducible, making them suitable for genetic studies (Table 4). The allele profiles of all 30 individuals across the 26 loci are summarized in Supplementary Table S2, which provides the allele sizes detected in each sample (S1–S30). Representative amplification patterns from two primer pairs, Da6 and Da11, tested across 15 accessions of *D. alatus*, are shown in Figure 6.

Across the 26 loci, the number of alleles detected ranged from 2 to 9, with fragment sizes spanning 107–312 bp. The PIC values varied between 0.315 (Da1 and Da18) and 0.410 (Da8), while discrimination power (Dp) ranged from 0.296 (Da8) to 0.484 (Da23), with mean values of 0.381 and 0.385, respectively. These results indicate that the developed markers are sufficiently informative and suitable for use in population genetic studies of *D. alatus*.

Table 4. Summary of polymorphic characteristics of EST-SSR loci derived from the transcriptome of *D. alatus*.

Primer ID	Motif	Primer 5' - 3'	Expected size (bp)	Observed range (bp)	Number of alleles	PIC value	Dp value
Da1	$(TG)_{10}$	F: TGAGAAGAGAAGCGAGGCTG	231	231 – 241	2	0.315	0.331
		R: CCTCTTGGATCCTGCGACAT					
Da2	$(TC)_{10}$	F: CAAGGGTGGCCTTGTGATCT	124	114 – 148	5	0.393	0.368
		R: TGAGGGGAGTGAAGGAGGAG					
Da3	$(TC)_{10}$	F: ACAGATTAGGCTTGGCTGCA	144	136 – 154	4	0.350	0.413
		R: AAATGACGGGAGGCCATTGA					
Da4	$(GA)_{10}$	F: ACTCCGCAGAAACACCCTTT	183	167 – 205	8	0.398	0.344
		R: CGCCGTATTGGAGTAGAGGG					
Da5	$(TA)_{10}$	F: TCTTTTGGCTGCTGAGGTGT	187	187 – 235	9	0.399	0.316
		R: GCTTCAAGCTACTGAAAATGGC					
Da6	$(CTC)_7$	F: GGCGCTTTCACCCAAAAGAA	175	175 – 205	6	0.395	0.390
		R: GGGTTCCCATTCCCAGATCC					
Da7	$(TTG)_7$	F: GAATCCTGATGATGCTGCAGC	260	260 – 287	4	0.367	0.467
		R: CAACGCACCGGAGTATCAGA					
Da8	$(AGA)_7$	F: CTCTTTGGGGGTCCAGCAAT	216	201 – 243	6	0.410	0.296
		R: TCTCATGGCTACTCCTTGCA					
Da9	$(GGT)_7$	F: CGCCGTATTGGAGTAGAGGG	121	121 – 151	4	0.377	0.459
		R: CAACACCGAAGCCTCCTACA					
Da10	$(AGG)_7$	F: AGGAGGAAAGGAGAGCCAGT	203	197 – 239	6	0.398	0.377
		R: GGACGCATCTTTCCTGTCCT					
Da11	$(GCG)_7$	F: AGCGGAGAGAGAAAGAGGGT	276	267 – 312	6	0.385	0.377
		R: CATACCGGATCCGATCGGTT					
Da12	$(AAG)_7$	F: CGTAAGAGATGTCGGACCCG	240	240 – 255	4	0.370	0.340
		R: CGACGGGAAAGAGATCCTCG					
Da13	$(TAT)_7$	F: AACCACTAACACGAGGGTCC	211	211 – 235	4	0.381	0.382
		R: TGACACGCAAGCTTCTCCTT					
Da14	$(GAA)_7$	F: GGCGTAATACGGAATTAACGGG	244	235 – 274	7	0.404	0.349
		R: TCCGACCAGTACTGTACGGT					
Da15	$(TCT)_7$	F: ACCTTCTCCGCGTCCATTTT	176	164 – 200	6	0.403	0.320
		R: GAATCCGACCAGACCAGTCC					
Da16	$(TGT)_7$	F: AGCAAAGGAGGGAACACGAG	250	250 – 286	6	0.392	0.373
		R: GGCGTACGAAACTCGAAAGG					
Da17	$(CAA)_7$	F: TCCAGCTAGCCTTGAATGCA	263	248 – 285	6	0.385	0.405
		R: GAGAGCATGGAGGATCTGGC					
Da18	$(GCA)_7$	F: AAGCAGCCCATCTTGAGGTT	276	251 – 275	2	0.315	0.370
		R: TCGAGGCATTTGCTCCACTT					

Da19	$(ACG)_{15}$	F: TGGTTTTGTGCTTCTTCGCG R: GACTCCAACCTCACTCCGTC	233	220 – 247	7	0.395	0.309
Da20	(CAT)8	F: AAACAGAGGCTTCCATCGCA R: ATGGATAGGCTGTGGCTTGG	136	111 – 150	7	0.395	0.359
Da21	(CCA)8	F: AAAGCTTCCCCACAGGTCAG R: GGCTCTGGACCTCATCATCG	225	225 – 252	4	0.377	0.474
Da22	$(TTC)_7$	F: GAGTTGCCACCGGAAAAACC R: TTGCTTCTCCCCAGAAGCAG	107	107 – 131	5	0.398	0.443
Da23	$(GAG)_7$	F: GGTAAGGTCTGCGGTGACAA R: GACAAAGGGAGAGATGGCCA	202	208 – 238	4	0.370	0.484
Da24	$(AAT)_7$	F: CGCTAGGATTTCGCAAACCG R: TTCTGGCTGCCTTCCGATTT	270	264 – 273	3	0.342	0.431
Da25	$(TTC)_7$	F: CTTGTGAACCCCACCTCTCC R: TGCTGAGGCAAATGGTCCAT	208	208 – 232	6	0.385	0.375
Da26	$(GAG)_7$	F: CGCAGGGTAAGGATGGCTAG R: AATTACCCCCACAACCCACC	123	123 – 150	5	0.396	0.447
				Average	5.23	0.381	0.385

Note: PIC, polymorphism information content; Dp, discrimination power value.

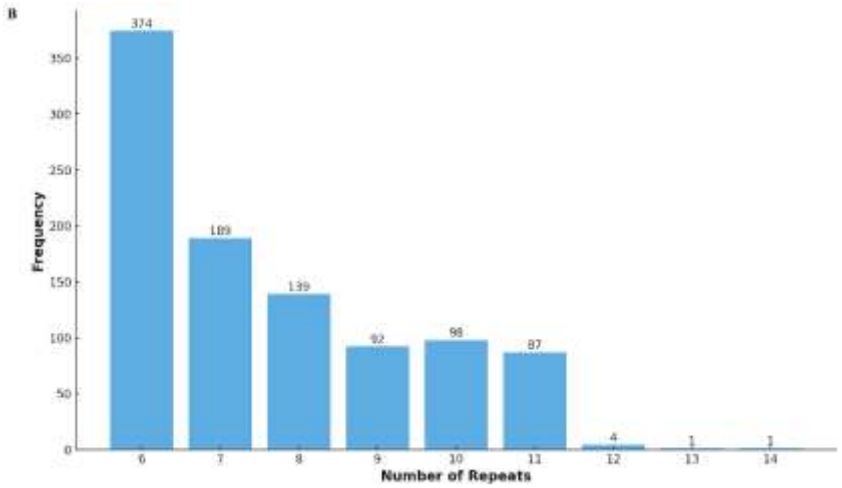


Figure 6. Electrophoresis of PCR products from 15 accessions of *D. alatus* using two SSR loci. A, amplification profile of locus Da6; B, amplification profile of locus Da11. M = VC 100 bp DNA ladder (Vivantis); lanes 1–15 = PCR products from different accessions.

To further verify the performance of 26 SSR loci, the genetic diversity was evaluated across five populations of *D. alatus* (Table 5). Results showed that the observed heterozygosity (H_o) values ranged from 0.000 (locus Da1, Da12, Da18, and Da24, respectively) to 0.823, where the locus Da5 possessed the highest H_o value among all loci. Expected heterozygosity (H_e) analysis revealed that the H_e values varied from 0.036 (locus Da18) to 0.770 (Da5). Among the screened EST-SSR loci, locus Da5 exhibited high H_e values (0.770) and H_o values (0.823), indicating its potential utility in population genetic studies. Conversely, locus Da18 displayed low expected heterozygosity (0.036) with a homozygosity value of 0.000, suggesting that it possessed low genetic diversity, which may contribute to the occurrence of inbreeding within the *D. alatus* population.

Table 5. Genetic diversity parameters of 26 EST-SSR loci across five populations of *D. alatus*.

Locus	Na	Ne	I	Ho	He
Da1	1.400 ± 0.245	1.197 ± 0.125	0.203 ± 0.125	0.000 ± 0.000	0.131 ± 0.081
Da2	3.200 ± 0.583	2.620 ± 0.616	0.941 ± 0.223	0.543 ± 0.151	0.528 ± 0.107
Da3	3.000 ± 0.447	2.351 ± 0.356	0.901 ± 0.164	0.677 ± 0.134	0.526 ± 0.084
Da4	4.200 ± 0.663	3.382 ± 0.483	1.262 ± 0.169	0.720 ± 0.140	0.673 ± 0.056
Da5	5.600 ± 0.812	4.838 ± 0.655	1.591 ± 0.165	0.823 ± 0.051	0.770 ± 0.045
Da6	4.000 ± 0.707	2.973 ± 0.448	1.153 ± 0.152	0.613 ± 0.102	0.630 ± 0.057
Da7	3.000 ± 0.316	2.410 ± 0.343	0.910 ± 0.164	0.623 ± 0.139	0.531 ± 0.097
Da8	3.000 ± 0.632	2.521 ± 0.564	0.916 ± 0.200	0.450 ± 0.200	0.535 ± 0.078
Da9	3.000 ± 0.316	2.369 ± 0.196	0.945 ± 0.098	0.467 ± 0.133	0.562 ± 0.047
Da10	4.200 ± 0.490	2.854 ± 0.220	1.187 ± 0.073	0.560 ± 0.090	0.640 ± 0.031
Da11	4.200 ± 0.374	3.353 ± 0.280	1.281 ± 0.095	0.687 ± 0.060	0.691 ± 0.032
Da12	2.000 ± 0.316	1.632 ± 0.283	0.510 ± 0.167	0.000 ± 0.000	0.320 ± 0.102
Da13	2.200 ± 0.200	1.792 ± 0.180	0.625 ± 0.099	0.067 ± 0.041	0.414 ± 0.071
Da14	4.000 ± 0.775	3.313 ± 0.590	1.224 ± 0.165	0.557 ± 0.135	0.664 ± 0.050
Da15	3.400 ± 0.812	2.671 ± 0.531	0.959 ± 0.278	0.470 ± 0.126	0.521 ± 0.139
Da16	4.000 ± 0.316	3.041 ± 0.374	1.199 ± 0.105	0.630 ± 0.080	0.648 ± 0.048
Da17	4.400 ± 0.245	3.733 ± 0.339	1.371 ± 0.079	0.687 ± 0.060	0.722 ± 0.028
Da18	1.200 ± 0.200	1.044 ± 0.044	0.065 ± 0.065	0.000 ± 0.000	0.036 ± 0.036
Da19	3.800 ± 0.663	2.948 ± 0.367	1.135 ± 0.144	0.670 ± 0.126	0.636 ± 0.050
Da20	4.400 ± 0.678	3.352 ± 0.598	1.238 ± 0.223	0.670 ± 0.126	0.631 ± 0.105
Da21	2.800 ± 0.374	2.030 ± 0.232	0.798 ± 0.132	0.467 ± 0.141	0.476 ± 0.070
Da22	3.600 ± 0.245	2.642 ± 0.308	1.071 ± 0.100	0.377 ± 0.139	0.597 ± 0.054
Da23	3.400 ± 0.245	2.647 ± 0.163	1.065 ± 0.044	0.593 ± 0.105	0.616 ± 0.025
Da24	2.400 ± 0.245	1.875 ± 0.117	0.717 ± 0.072	0.000 ± 0.000	0.458 ± 0.035
Da25	4.200 ± 0.490	3.211 ± 0.424	1.232 ± 0.138	0.640 ± 0.101	0.656 ± 0.064
Da26	3.800 ± 0.374	2.821 ± 0.354	1.131 ± 0.114	0.473 ± 0.109	0.622 ± 0.049
Average	3.400± 0.123	2.678± 0.098	0.986± 0.039	0.479± 0.029	0.548± 0.019

Note: Na, mean number of alleles per locus across populations; values are shown as mean ± standard error; Ne, Effective number of alleles; H_o, Observed heterozygosity; H_e, Expected heterozygosity

Genetic diversity and structure analysis

The pairwise Jaccard dissimilarity coefficients among the 30 *D. alatus* accessions revealed substantial variation in genetic diversity, as shown in Figure 7A, suggesting substantial genetic similarity within populations. The matrix of pairwise Jaccard dissimilarity values is provided in Supplementary Table S3. The mean Jaccard dissimilarity (0.758) indicated a high level of genetic differentiation among individuals of *D. alatus*. Accessions S26 and S27 exhibited a minimal dissimilarity value (0.317), suggesting a close genetic affinity between these two accessions. The maximum value (0.913) was observed between S14 and S24, which signifies pronounced genetic divergence between individuals from separate populations. These patterns were consistent with clustering obtained from the NJ analysis (Figure 7B).

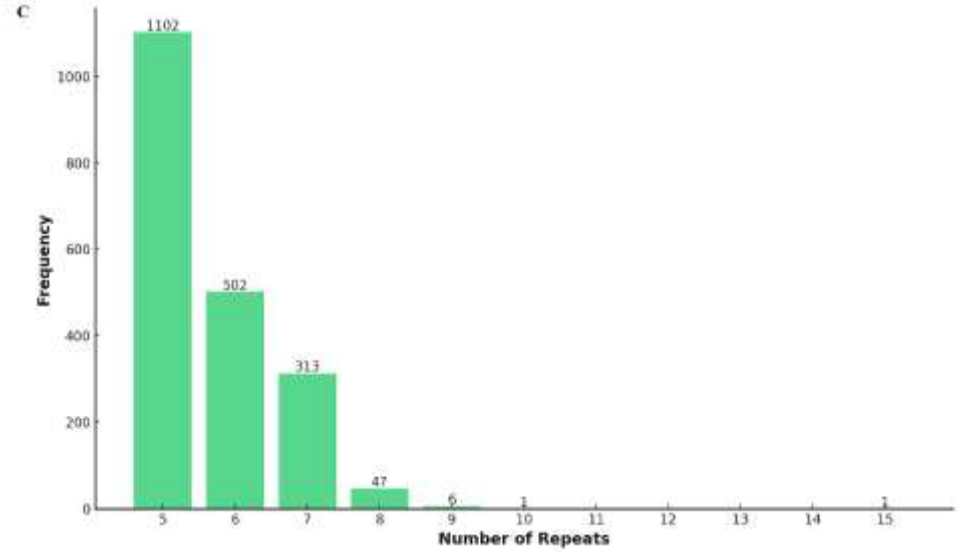


Figure 7. Heatmap (A) and circular NJ dendrogram (B) of 30 D. alatus individuals based on 26 SSR loci. Colored dots indicate five populations: PT (sky blue), MNS (orange), PH (green), IB (pink), and MKK (purple).

Bayesian clustering classified the 30 genotypes of *D. alatus* into three distinct clusters (Figure 8) as shown in the bar plot and supported by ΔK analysis. Individuals S7–S11, S1–S6, and S30, and S26–S29 were consistently assigned to clusters 1, 2, and 3, respectively. S6, S13, and S28 individuals exhibited admixed ancestry. The NJ tree based on Jaccard dissimilarity coefficients showed clustering patterns largely consistent with STRUCTURE assignments, grouping individuals from the same cluster together. Subclusters were also observed within the dendrogram, reflecting finer genetic relationships among genotypes. The NJ tree also showed clustering patterns that largely corresponded to the geographic origins of individuals.

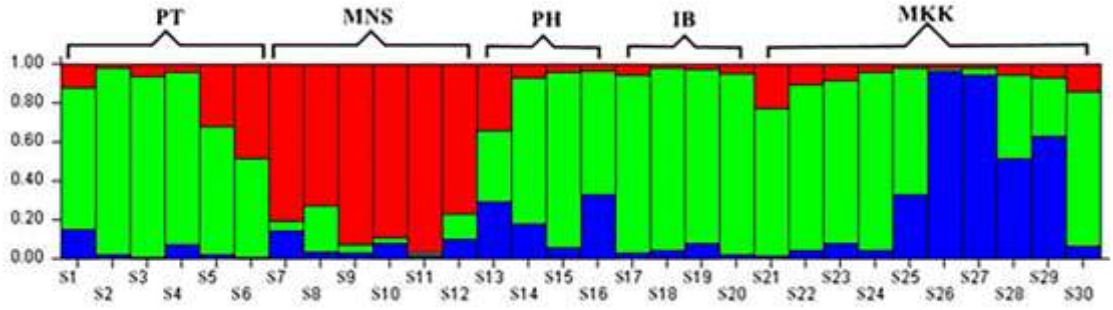


Figure 8. STRUCTURE bar plot of 30 D. alatus individuals from five populations based on 26 SSR loci. Each bar corresponds to one individual and shows the proportion of genetic ancestry assigned to three inferred clusters (K = 3). Individuals (S1–S30) are arranged by population of origin.

DISCUSSION

The transcriptome of *D. alatus* leaf tissue clean reads in this study totaled 4.939 GB. A total of 66,281 unigenes were generated by *de novo assembly*, and 66.5% of them were successfully annotated. Among them, 10,726 sequences were identified to contain SSRs, with a genic SSR frequency of 6.2%. The SSR density was estimated at 224.1 loci per megabase, which is higher than that reported in *Paeonia* spp. (64.31 loci/Mb; Luo et al., 2021). However, several monocot species, such as *Zea mays*, *Sorghum bicolor*, and *Oryza sativa*, have shown even higher SSR densities, ranging from 450 to 580 loci/Mb (Jayashree et al., 2006), reflecting taxonomic variability in SSR distribution.

Compared to other sequenced species, *D. alatus* exhibited a similar SSR motif distribution pattern, with trinucleotide repeats being the most abundant, observed in *Chimonanthus praecox* (58.90%; Liu et al., 2024), and *Carthamus tinctorius* (35.7%; Ahmadi and Ahmadikhah, 2022). The dominance of trinucleotide repeats may be attributed to evolutionary constraints to preserve the reading frame in coding regions, as indels that are multiples of three do not disrupt codon structure (Metzgar et al., 2000). Such constraints are particularly relevant to SSRs located within ESTs, which are typically found in expressed regions and thus under stronger purifying selection.

Moreover, in other plant species, the distribution of SSR motifs varies; in *Bergenia ciliata,* dinucleotide repeats were more prevalent (Singh et al., 2024), and a similar pattern was also observed in *Paeonia* spp. (Luo et al., 2021). These variations may be due to differences in SSR mining criteria, database size, genome composition, or bioinformatic algorithms. Despite these variations, the uniform presence of trinucleotide motifs across species denotes their evolutionary permanence and appropriateness as gene-based molecular markers. The most dominant dinucleotide motif in *D. alatus* was AG/TC, which is similar to patterns found in dicotyledonous species, such as *Corchorus capsularis* (Saha et al., 2017). Still, the CG/GC motif occurred only once. AAG/CTT was the most prolific (528 occurrences) among the trinucleotide motifs, followed by ATC/ATG and ACC/GGT. *C. capsularis* also showed a similar trend. These findings suggest that the *D. alatus* SSR motif distribution is comparable to that of other dicot species.

To evaluate *D. alatus'* genetic diversity, 26 SSR loci containing di- and trinucleotide motifs obtained from the transcriptome were selected. The results showed moderate to high genetic diversity, with mean observed heterozygosity (Ho) of 0.479, mean expected heterozygosity (He) of 0.548, mean effective number of alleles (Ne) of 2.678, and a mean polymorphism information content (PIC) of 0.320 which are higher values than those reported from isozyme markers (Ho = 0.088, He = 0.092) (Changtragoon, 2001) and fall within or above the range reported from SSRs from other species and applied to *D. alatus* in Thailand (Ho = 0.237,

He = 0.733) (Chokthaweepanich et al., 2022) and Vietnam (Ho = 0.209, He = 0.239) (Tam et al., 2014). Although previous studies had computed the genetic diversity of *D. alatus* at the population level, the species-specific EST-SSRs derived from the leaf transcriptome of *D. alatus* exhibited more efficiency in presenting averages across loci in detecting allelic variation.

Furthermore, the moderate PIC values of gene-derived SSRs suggested that they tended to be more conserved than genomic SSRs, likely due to their location in functional regions. The transcriptome-based SSR markers developed in this study present a valuable molecular resource for genetic investigations of *D. alatus* and potentially other species within the Dipterocarpaceae family. Due to their reproducibility and association with expressed genomic regions, these markers were well-suited for applications in genetic diversity assessment, forest resource monitoring, and conservation genetics. These findings align with the IPBES (2019) and Hoban et al. (2021) recommendations, which support the integration of genetic indicators into national biodiversity strategies to enhance the future of natural and historic ecosystems. Earlier dipterocarp species studies primarily used genomic SSRs from non-coding regions (Terauchi, 1994; Ujino et al., 1998; Isagi et al., 2002; Wang et al., 2020); however, our study concentrated on SSRs identified from transcribed regions (ESTs). Genomic SSRs often present high polymorphism but can be deficient in functional relevance. On the other hand, EST-derived SSRs were, in general, better-maintained and, in theory, transferable across related taxa, having a source in coding regions under discerning limitations. These qualities highlight the uniqueness and rational use of the markers developed herein, especially for use in gene-based diversity studies, conservation planning, and future marker-assisted selection in non-model forest tree species.

The genetic structure inferred from both Bayesian clustering and the NJ tree highlights a strong geographic component, suggesting that spatial separation contributes significantly to population differentiation in *D. alatus*. The general agreement between STRUCTURE and the dendrogram underscores the reliability of the observed patterns. Nonetheless, signs of admixture in specific individuals imply that gene flow among populations persists, possibly through pollen or seed dispersal. Additionally, the emergence of subclusters within major groups may reflect localized differentiation, perhaps influenced by microhabitat variation or historical demographic dynamics. These findings point to a complex interplay between isolation and connectivity in shaping the genetic landscape of this species. Nonetheless, it should be acknowledged that the sample sizes for specific natural populations in this study were relatively small (4–6 individuals). Such limited sampling may constrain the accuracy of estimates for heterozygosity, allele frequencies, and Hardy–Weinberg equilibrium, and may reduce the statistical power in detecting population structure. To partially offset this limitation, one ex situ population comprising approximately 200–300 accessions was included, thereby broadening the allelic representation. While the present results provide meaningful insights into the genetic diversity and structure of *D. alatus*, further studies with larger sample sizes and broader population coverage will be necessary to validate and refine these findings.

CONCLUSION

This study identified 12,514 SSR loci from the transcriptome dataset of D. alatus, based on 66,281 assembled sequences. A total of 26 polymorphic EST-SSR markers were developed and applied to assess the genetic diversity and population structure of 30 individuals. The results revealed a moderate to high level of genetic diversity, enabling clear differentiation of population groups according to their geographic origins. The transcriptome-derived SSR markers demonstrated high accuracy, discriminatory power, and reproducibility, making them suitable for studies in conservation genetics, forest restoration, and sustainable resource management in D. alatus and other related species in the Dipterocarpaceae family. Moreover, transcriptomic technology plays a vital role in the development of molecular markers for non-model plants, especially local forest species with limited genomic information, which are crucial for future policy development and biodiversity conservation strategies.

ACKNOWLEDGMENTS
We extend our sincere gratitude to the journal reviewers for their insightful and constructive comments, which significantly improved the quality of this manuscript. This research was generously supported by Khon Kaen University and the Plant Genetic Conservation Project under the Royal Initiative of Her Royal Highness Princess Maha Chakri Sirindhorn.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

DATA AVAILABILITY
The raw sequencing data and the assembled transcriptome have not been deposited in a public repository but are available from the corresponding author upon reasonable request.

REFERENCES
Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. Journal of Molecular Biology 215: 403-410.

Ahmadi AJ, Ahmadikhah A (2022). Occurrence of simple sequence repeats in cDNA sequences of safflower (Carthamus tinctorius) reveals the importance of SSR-containing genes for cell biology and dynamic response to environmental cues. Frontiers in Plant Science 13: 991107.

Amiryousefi A, Hyvönen J, Poczai P (2018). iMEC: online marker efficiency calculator. Applications in Plant Sciences 6: e1159.
Boontawee, B. (2001). The status of in situ and ex situ conservation of Dipterocarpus alatus Roxb. in Thailand. In Proceedings of the International Conference on Ex situ and In situ Conservation of Commercial Tropical Trees (pp. 171–181). Yogyakarta, Indonesia.

Bolger AM, Lohse M, Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.

Changtragoon, S. (2001). Evaluation of Genetic Diversity of Dipterocarpus alatus Genetic Resources in Thailand using Isozyme Gene Markers. Ex Situ and In Situ Conservation of Commercial Tropical Trees: the Contribution of Genetic Resource Conservation to Tree Breeding, Biotechnology, and Future Commercial Plantation Programs. 11-13 June 2001, Yogyakarta, Indonesia. 13 p.

Chokthaweepanich H, Butdavapee S, Phongkaew P (2022). Genetic diversity and relationships of Dipterocarpus alatus Roxb. (Dipterocarpaceae) on island in freshwater lake and nearby mainland. Agriculture and Natural Resources 56: 177-186.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, et al. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.

Dyrmose AMH, Turreira-García N, Theilade I, Meilby H (2017). Economic importance of oleoresin (Dipterocarpus alatus) to forest-adjacent households in Cambodia. Natural History Bulletin of the Siam Society 62: 67-84.

Evanno G, Regnaut S, Goudet J (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14: 2611-2620.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols 8: 1494-1512.

Hoban S, Campbell CD, da Silva JM, Ekblom R, Funk WC, et al. (2021). Genetic diversity is considered important but interpreted narrowly in country reports to the Convention on Biological Diversity: current actions and indicators are insufficient. Biological Conservation 261: 109233.

IPBES (2019). Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Díaz S, Settele J, Brondízio ES, Ngo HT, Guèze M, et al. (eds.). IPBES Secretariat, Bonn, Germany, 56 pp.

Isagi Y, Kenta T, Nakashizuka T (2002). Microsatellite loci for a tropical emergent tree, Dipterocarpus tempehes V. Sl. (Dipterocarpaceae). Molecular Ecology Notes 2: 12-13.

Jaccard P (1901). Etude de la distribution florale dans une portion des Alpes et du Jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37: 547-579.

Jayashree B, Punna R, Prasad P, Bantte K, et al. (2006). A database of simple sequence repeats from cereal and legume expressed sequence tags mined in silico: survey and evaluation. In Silico Biology 6: 607-620.

Khoo E, Barstow M, Maycock C, Bodos V, et al. (2022). Red List of Dipterocarpaceae. Botanic Gardens Conservation International, Richmond, UK.

Kumar S, Stecher G, Li M, Knyaz C, et al. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution 35: 1547-1549.

Li YL, Liu JX (2018). StructureSelector: A web-based software to select and visualize the optimal number of clusters using multiple methods. Molecular Ecology Resources 18: 176-177.

Liu B, Wu HF, Cao YZ, Yang XM, et al. (2024). Establishment of novel simple sequence repeat (SSR) markers from Chimonanthus praecox transcriptome data and their application in the identification of varieties. Plants 13: 2131.
Lopez L, Barreiro R, Fischer M, Koch MA (2015). Mining microsatellite markers from public expressed sequence tags databases for the study of threatened plants. BMC Genomics 16: 781.

Luo LM, Yang YY, Zhao HW, Leng PS, et al. (2021). Development of EST-SSR markers and association analysis of floral scent in tree peony. Scientia Horticulturae 289: 110409.

Luo LM, Ye P, Lin QQ, Liu M, et al. (2024). From sequences to sustainability: Exploring dipterocarp genomes for oleoresin production, timber quality, and conservation. Plant Science 346: 112139.

Ly V, Nanthavong K, Pooma R, Barstow M, Luu HT, et al. (2023). Dipterocarpus alatus (amended version of 2017 assessment). The IUCN Red List of Threatened Species 2023: e.T33007A231194805. Available at: https://dx.doi.org/10.2305/IUCN.UK.2023-1.RLTS.T33007A231194805.en. Accessed 3 Jun 2025.

Metzgar D, Bytof J, Wills C (2000). Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Research 10: 72-80.

Peakall R, Smouse PE (2006). GENEALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes 6: 288-295.

Peakall R, Smouse PE (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. Bioinformatics 28: 2537-2539.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.r-project.org

Roy RP, Jha RP (1965). Cytological studies in Dipterocarpaceae I. Journal of the Indian Botanical Society. 44: 387-397.

Saha D, Rana RS, Chakraborty S, Datta S, Kumar AA, Chakraborty AK, Karmakar PG (2017). Development of a set of SSR markers for genetic polymorphism detection and interspecific hybrid jute breeding. The Crop Journal. 5: 416-429.

Singh HC, Tiwari V, Tiwari A, Rana TS (2024). Development of EST-SSR markers in Bergenia ciliata using de novo transcriptome sequencing. Genome. 67: 119-124.

Tam NM, Duy VD, Duc NM, Giap VD, et al. (2014). Genetic variation in and spatial structure of natural populations of Dipterocarpus alatus (Dipterocarpaceae) determined using single sequence repeat markers. Genetics and Molecular Research. 13: 5378-5386.

Thiel T, Michalek W, Varshney RK, Graner A (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theoretical and Applied Genetics. 106: 411-422.

Terauchi R (1994). A polymorphic microsatellite marker from the tropical tree Dryobalanops lanceolata (Dipterocarpaceae). Japanese Journal of Genetics. 69: 567-576.

Ujino T, Kawahara T, Tsumura Y, Nagamitsu T, et al. (1998). Development and polymorphism of simple sequence repeat DNA markers for Shorea curtisii and other Dipterocarpaceae species. Heredity. 81: 422-428.

Untergasser A, Cutcutache I, Koressaar T, Ye J, et al. (2012). Primer3—new capabilities and interfaces. Nucleic Acids Research. 40(15): e115.

Wangsomnuk PP, Ritthitum W, Ruttawat B, Wangsomnuk P (2014). Comparative analysis of DNA extracted from mature leaves of rubber tree and application for seventeen tropical plant species for PCR amplification. AAB Bioflux. 6: 44-56.

Wang C, Ma X, Tang L (2020). Isolation and characterization of twelve polymorphic microsatellite markers in the endangered Hopea hainanensis (Dipterocarpaceae). Ecology and Evolution. 11: 4-10.

Xu J, Wang Y, Wu K, Chen J (2024). Identification and characterization of functionally relevant SSR markers in natural Dalbergia odorifera populations. BMC Plant Biology. 24: 315.