



Research Article

Population genetics analysis of *Diospyros mun* A.Chev. ex Lecomte (Ebenaceae) based on EST-SSR markers derived from a novel transcriptome

Xuan Thi Tuyet Bui[‡], Duy Dinh Vu[§]

[‡] Institute of Ecology and Biological Resources, Vietnam Academy of Science and Technology, Hanoi, Vietnam
[§] Join Vietnam–Russia Tropical Science and Technology Research Center, Hanoi, Vietnam

Corresponding author: Duy Dinh Vu (duydinhvu87@gmail.com)

Academic editor: Yuanjun Zhu

Received: 24 Jun 2024 | Accepted: 14 Sep 2024 | Published: 18 Sep 2024

Citation: Bui XT, Vu D (2024) Population genetics analysis of *Diospyros mun* A.Chev. ex Lecomte (Ebenaceae) based on EST-SSR markers derived from a novel transcriptome. Biodiversity Data Journal 12: e130385.
<https://doi.org/10.3897/BDJ.12.e130385>

Abstract

Diospyros mun A.Chev. ex Lecomte (Ebenaceae), a native evergreen tree in Vietnam, has important economic and ecological values. The absence of effective and reliable molecular markers has hampered the study of *D. mun*'s genetic diversity and population structure, even though it is an endemic and endangered species. Therefore, significant enrichment of genomic resources is urgently needed to uncover and better understand the genetic architecture of *D. mun*. This study aims to demonstrate an efficient and reliable tool to explore the polymorphism within *D. mun* germplasm. It provides a valuable platform for the breeding and conservation of this species and other endangered species worldwide. The Illumina HiSeq™ 4000 sequencing technology was applied for the transcriptomic analysis, genetic differentiation and population structure of *D. mun* in Vietnam. In this study, the transcriptomes of *D. mun* were analysed using the Illumina HiSeq™ 4000 sequencing system and a total of 5,588,615,700 base pairs were generated. *De novo* assembly indicated that 91,134 unigenes were generated (average length = 645.55 bp, N50 = 957 bp, Q20 = 98.08% and Q30 = 94.51%). A total of 92,798 and 21,134 unigenes had significant similarities amongst Nr and Swiss-Prot, respectively. In the GO database, 19,929 unigenes were annotated and these genes were divided into three major categories and 50 subcategories. In the KOG analysis,

18,499 unigenes were annotated and divided into 25 gene function categories. In the KEGG analysis, 12,017 unigenes were annotated. According to the related pathways involved, they could be classified into 56 subclasses. In this study, we have identified a total of 9,391 EST-SSR markers. Ten microsatellite loci were employed to assess the genetic diversity and structure of 82 adult *D. mun* trees across three populations in Vietnam. The results indicated moderate levels of genetic diversity with $PIC = 0.77$, $N_A = 3.9$, $N_E = 2.8$, $H_o = 0.56$ and $H_E = 0.58$ and the fixation index value was recorded as positive for three populations (NS, NH and CP). Genetic differentiation among populations was low ($F_{ST} = 0.045$), suggesting limited gene flow ($Nm = 5.34$). This result indicates gene exchange between the populations of ancient *D. mun* from different geographical areas and regions. The analysis of molecular variance (AMOVA) showed that high genetic variation existed within individuals (91%) compared to amongst populations (4%). Genetic structure analysis, DAPC and the NJ tree indicated that the three populations were divided into three main clusters. With this study, we provide a molecular resources for the breeding and conservation of *D. mun*.

Keywords

Diospyros mun, genetic diversity, endangered species, SSR markers, transcriptome

Introduction

Diospyros mun A.Chev. ex Lecomte (Ebenaceae) is an endemic and rare species from Vietnam, this species only growing in restricted habitat at higher altitudes in Tuyen Quang, Hoa Binh, Ninh Binh, Ninh Thuan and Binh Thuan Provinces (Nguyen 2003, Ngo 2014, Vu et al. 2014). They have been extensively used for furniture, handicrafts, construction the commercial trade as well as a source of valuable medicine or ornamental plants. Globally, *D. mun* is listed as Critically Endangered under criteria A1cd (Nguyen 1998). In Vietnam, due to the reduction of habitat and over-exploitation, all these species have been listed as Endangered (EN A1c,d, B1 + 2a) (MOST and VAST 2007). Currently, questions on the genetic analyses of this species remain unanswered due to the limitation of available molecular markers. Therefore, it is necessary to conduct research into the fields of distribution, ecology, natural active components, phytochemicals and any other potential resources that *D. mun* possesses to ensure its survival. Accordingly, the data from the transcriptome sequence were utilised to enhance the effectiveness of the SSR generation process for the species.

Conservation and management of a species require information on the ecological and genetic diversity within and amongst populations (Vu et al. 2016, Liu et al. 2023, Yan et al. 2024). To gather the data, particularly to gain deeper insights into genetic processes, advanced biological techniques are necessary (Durmaz et al. 2015). Genetic diversity is reflected in the wide variation of phenotypes and genotypes in plants (Govindaraj et al. 2015). Polymorphic genes lead to the presence of heterozygous genotypes within populations (Li et al. 2023). This variation in genotypes allows populations to adapt to

environmental changes and interact with other populations. Furthermore, genetic diversity serves as the foundation for heterosis development (Salgotra and Chauhan 2023).

To some extent, if genetic diversity between parents is greater, the higher the heterosis. Populations with high genetic diversity are valuable as breeding material for genetic improvement programs (Pham et al. 2024). These are factors that help the organism's genetic lineage, resistance to epidemics and adaptation to the changes in environmental conditions. Assessment of genetic diversity may rely on morphological characteristics, agronomy and the genetic level of the plant. Methods of evaluation are based on morphological characteristics and agronomy is conducted easily, but mistakes can be made because phenotypic expressions are the interactions between the genotype and the environment or are due to subjective assessment by researchers.

The emergence of next-generation sequencing technology in recent years, particularly high throughput transcriptome sequencing, has offered extensive expressed sequence data for plants that are not models (Egan et al. 2012, Vlk and Řepkova 2017). Transcriptome sequencing is a simple and effective tool used to develop a large number of unigene-based microsatellites (Simple sequence repeats, SSR) for comprehensive analysis of genome in plants (Vu et al. 2020, Zhang et al. 2021). Transcriptome sequences generally represent the genetic information of gene expression and may be directly related to gene function (Vu et al. 2020, Li et al. 2023). Recently, the development of SSR markers has been reported in persimmon - *Diospyros kaki* Thunb. (Soriano et al. 2006, Naval et al. 2010, Guo and Luo 2011, Luo et al. 2013, Liang et al. 2015, Wang et al. 2018, Li et al. 2019, Wang et al. 2020, Samarina et al. 2021, Li et al. 2023), date-plum (*Diospyros lotus* L.), Chekiang persimmon (*Diospyros glaucifolia* Metc.) and oily persimmon (*Diospyros oleifera* Cheng.), Jinzaoshi (*Diospyros* sp.) (Yang et al. 2015), but transcriptome sequences have not been reported in the development of SSR markers in *D. mun*. Moreover, many ESTs for different plant species have been available in recent decades. However, the genetic diversity and genetic structure of *D. mun* have not been explored, which are important prerequisites for formulating and implementing conservation and restoration measures. Therefore, transcriptome sequencing needs to be conducted to understand the genetic and metabolic adaptations in its habitats. An examination of the genetic diversity present within each population and between them is necessary to enhance conservation and management efforts about this species. As a result, this will contribute to improved knowledge of ecology, evolution and conservation genetics for this species. A transcriptome is the complete set of all RNA molecules that includes mRNA, tRNA, rRNA and other non-coding RNA in one cell or a group of cells in a particular environment or a developmental stage of the species.

The Illumina HiSeq™4000 platform was used in this study to obtain characteristics of the comprehensive transcriptome of *D. mun* and to develop a large number of expressed sequence tags - SSR (EST-SSR) markers. The genetic structure of all of the individuals studied, which came from three different populations, was analysed with the help of these markers and the degree of genetic variability within each population and between the

populations was calculated. Following this, the study's findings are incorporated into this species' management decisions, conservation efforts and restoration efforts.

Materials and Methods

Sample collection and RNA extraction

The leaves and stems of *D. mun* were collected in Ngoc Son Nature Reserve, Hoa Binh Province in May 2024 (Fig. 1) and immediately placed in liquid nitrogen in the field. They were later transported to the Laboratory of Molecular Biology, Institute of Ecology and Biological Resource, where they were stored at -80°C for RNA extraction. To perform Illumina sequencing, total RNA was first extracted from each sample using a plant RNA Kit and then processed with DNase I. These procedures followed the manufacturer's instructions (Omega Biotack, Inc.). The quality of the processed RNA samples was checked by running electrophoresis on 1% agarose gels and Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA) and the quantity was checked using Nanodrop ND-2000 spectrophotometer (NanoDrop Technologies, DE, USA). Finally, equal amounts of RNA from different samples were pooled for RNA sequencing.

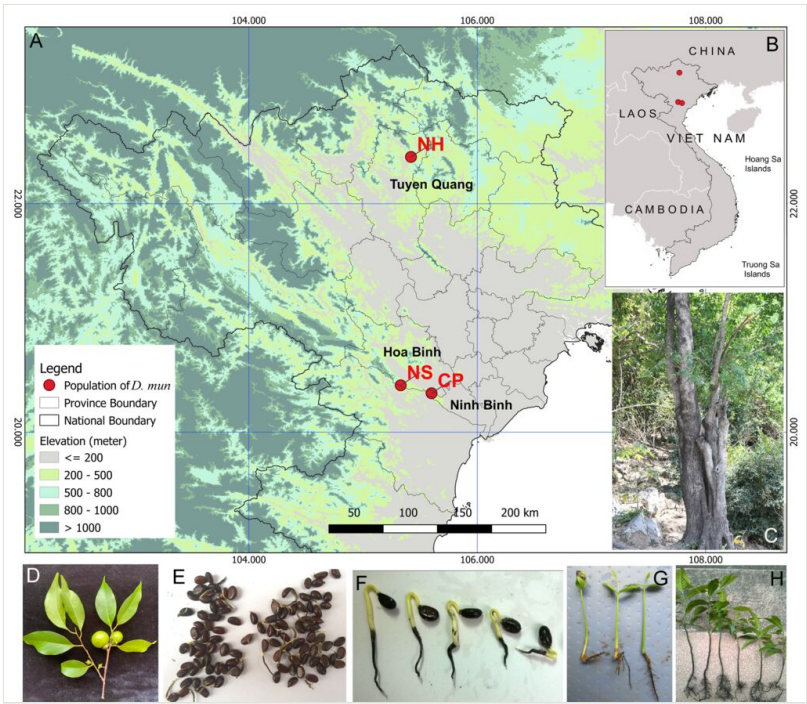


Figure 1. [doi](#)
Map of field survey locations and geographic distributions of *D. mun* in the study. Map showing the collection sites (A, B), leaves and fruits (D), adult plant (C), seeds and sprouted seeds (E, F, G) and seedling (H).

Illumina sequencing and de novo transcriptome assembly

RNA samples were sent to Breeding Biotechnologies Co., Ltd. for mRNA-seq library construction and sequencing using Illumina HiSeq™ 4000 next-generation platform technology (Illumina, Inc., CA, USA). The quality of the raw reads was checked. The adapter contaminants with the raw reads > 5% of unknown nucleotide (N) and > 20% low-quality bases (quality value < 10) were removed to obtain high-quality clean reads. The clean reads were de novo assembled to construct contigs using TRINITY (Grabherr et al. 2011). The contigs were clustered into unigenes using TGICL v.2.1 (Pertea et al. 2003).

Functional annotations

To determine the predicted function, all unigene transcripts were used and compared to the NCBI non-redundant (Nr) protein database (Deng et al. 2006) and Swiss-Prot database (Lussi et al. 2023), Gene Ontology (GO) (Hinderer and Moseley 2020), eukaryotic Orthologous Groups of proteins (KOG) (Koonin et al. 2004) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa 2004) databases using BLAST (E-value of 10^{-5}) to search for homologs. Sequences were aligned with the Protein family database (Pfam) to obtain unigene annotation information using the HMMER (Finn et al. 2013). BLAST2GO software (Conesa et al. 2005) was used to determine gene ontology (GO) annotations of assembled unigenes for the categories of biological processes and molecular functions of cellular components. The BLAST software aligned the unigene sequences to the eukaryotic Orthologous Groups (KOG) clusters. KEGG pathway annotations were analysed for the metabolic pathways and the related gene functions.

DNA extraction, microsatellite polymorphism identification

First, we considered 5,486 SSR designed loci, amongst the 100 SSR primer pairs randomly selected for PCR identification and 10 were successfully amplified by 82 *D. mun* individuals from three natural populations (Fig. 1, Table 1). The total genomic DNA was extracted from the plant DNA Kit according to the manufacturer's instructions (BioTeke, Beijing, China). The DNA purity and integrity were tested by Nanodrop ND-2000 spectrophotometer (NanoDrop Technologies, DE, USA) and then diluted to 20 ng·μl⁻¹. The PCR amplification reaction was performed in a 25 μl reaction volume, comprising 2.5 μl of template DNA, 12.5 μl of 2X Taq Master mix, 1 μl of each primer and 8 μl deionised water. The GeneAmp PCR System 9700 (Applied Biosystems, USA) was used for the PCR amplification with the following conditions: initial denaturation at 94°C for 3 minutes, 40 cycles at 94°C for 30 s, a primer-specific annealing temperature for 30 s and 70°C for 1 min; followed by a final extension at 72°C for 10 min; then maintaining the samples at 4°C until they were analysed. The PCR products were sized and subjected to relative quantification between samples on a 5300 Fragment Analyzer (Agilent) with an Agilent DNF-905 dsDNA Kit (1-500 bp) (Agilent).

Table 1.

Sampling location and genetic diversity within *D. mun* populations using 10 SSR loci.

Population code	Locations	Latitude (E)	Longitude (N)	<i>N</i>	<i>N_A</i>	<i>N_E</i>	<i>P</i> %	<i>Ho</i>	<i>He</i>	<i>F_{IS}</i>	<i>F_{IS}/IM</i>
NH	Na Hang Nature Reserve, Tuyen Quang Province	22°24'35"	105°25'11"	28	3.8	2.3	100	0.53	0.55	0.05	0.02
NS	Ngoc Son Nature Reserve, Hoa Binh Province	20°24'22"	105°20'14"	26	4	2.8	100	0.58	0.59	0.002	0.03
CP	Cuc Phuong National Park, Ninh Binh Province	20°20'16"	105°36'21"	28	4.1	2.9	100	0.57	0.6	0.034	0.01
Species level					3.9	2.8	100	0.56	0.58	0.029	0.02

Note: *N*: population size; *N_A*: means number of alleles per locus; *N_E*: means number of effective alleles; *P*%; percentage of polymorphic loci; *Ho* and *He*: mean observed and expected heterozygosities, respectively; *F_{IS}*: inbreeding coefficient; The individual inbreeding model was performed to evaluate the *F_{IS}* index for null allele frequency *F_{IS}/IM*.

Data analysis

Genetic diversity: Null alleles and other genotyping errors were detected using the MICRO-CHECKER v.2.0 software (Van Oosterhot et al. 2004), with 1000 bootstrap iterations over loci to generate the expected homozygote and heterozygote frequencies. CERCUS (Kalonowski et al. 2007) was used to estimate each locus' *PIC* (polymorphism information content) value. Variables for genetic diversity per loci and population, including the number of alleles (*N_A*), number of effective alleles (*N_E*), percentage of polymorphic loci (*P*%), the observed heterozygosities (*Ho*), the expected heterozygosities (*He*), inbreeding coefficient (*F_{IS}*) and gene flow (*Nm*) were calculated using the programme GenAlEx v.6.5 (Peakall and Smouse 2012). The individual inbreeding model was performed to evaluate the *F_{IS}* index for null allele frequency (*F_{IS}/IM*) using INEst Chybicki and Burczyk 2008). Deviations from Hardy-Weinberg equilibrium (HWE) for loci

within populations were tested in CERCUS, based on 1,000 permutations of alleles amongst individuals.

Genetic differentiation and population structure: The F_{ST} (Weir and Cockerham 1984) and G'_{ST} values (Hedrick 2005) were able to quantify the genetic divergence that exists between the populations by GenAlEx v.6.5. Using ARLEQUIN v.3.1, the significance of the F_{ST} values in population pairs was assessed at a significance level of 0.05. The results of this testing were considered significant (Excoffier et al. 2017). The analysis of molecular variance (AMOVA) was carried out with the assistance of ARLEQUIN v.3.1, which was used to carry out the significance testing for the variance components involved. The outcomes of 10,000 different permutations served as the basis for this testing. To analyse the population structure using STRUCTURE v.2.3.4, the Bayesian clustering method was used (Pritchard et al. 2000). Establishing the admixture model with correlated allele frequencies required ten distinct runs for each of the groups in the dataset (K). These runs were carried out with 500,000 Markov Chain Monte Carlo (MCMC) iterations and a burn-in time of 100,000 iterations. K ranged from 1 to 15. STRUCTURE HARVESTER (Earl and von Holdt 2011) was used to detect the number of groups that best fit the dataset, based on the K that was determined by Evano et al. (2005) to determine the ideal value of K. This was done so that the optimal value of K could be determined. When the optimal K value had been determined, the duplicated findings were aligned with the help of CLUMPP v.1.1.2 (Jakobsson and Rosenberg 2007) and the allocated cluster membership bar plots were created with DISTRUCT v.1.1.1. (Rosenberg 2003). A Discriminant Analysis of Principal Components (DAPC) was also carried out by utilising the Adegenet package that was available for the R v.4.0.2 programme to discover clusters of individuals with a common genetic ancestor (Jombart et al. 2010). The DAPC was implemented without any previous information on the population's origin. The Bayesian Information Criterion was used to establish the "ideal" cluster size and the number of clusters, denoted by the letter "K," ranged from one to twenty in the experiments (BIC). The DAPC was also responsible for performing the previous information analysis to determine how individuals should be assigned to populations. To obtain a better visual representation of the genetic clusters, the complot function of adegenet was included. The *xva*/DAPC function was used to keep the top fourteen principal components of the principal component analysis (98.5% of the variance that was conserved), as well as the seven discriminant eigenvalues. The tree topologies of all individuals were constructed, based on the Neighbour-Joining (NJ) method using NTSYS v.2.11 (Saitou and Nei 1987) and MEGA v.11.0 (Tamura et al. 2021).

Data resources

All data generated or analysed during this study are included in this published article [and its supplementary information files]. Raw data, including biological replicates, are available from the corresponding author on reasonable request.

Results

Illumina sequencing and de novo assembly

Through RNA-seq, 18,628,719 original sequences of *D. mun* were obtained with a total of 5,588,615,700 base pairs, Q20 (98.08%) and Q30 (94.51%). The number of ‘N’ bases (ambiguous bases) and GC-content were 0% and 46.82%, respectively. It showed that the high-throughput sequencing platform ha obtained a higher quantity and quality of *D. mun* sequences, which is conducive to the subsequent data assembly and meets the needs of later bioinformatics research. After the clean reads were assembled by de novo, a total of 157,553 transcripts with sequence information of 166,642,297 bp and 91,134 unigenes with sequence information of 58,649,736 bp were obtained (Table 2). Analysis of the sequence length of the transcript shows that its average length was 105.77 bp and N50 was 1,867 bp. Amongst them, short sequences of 200 to 300 bp were the majority, with 37,823 sequences accounting for 24% of the total; sequences with a length of 300 to 500 bp were 32,955 (20.92%) sequences; sequences of 500 to 1,000 bp in length were 30,838 (19.57%) sequences; 31,236 (19.83%) sequences ranged from 1,000- 2,000 bp and sequences longer than or equal to 2,000 bp accounted for 24,701 (15.68%) sequences. Unigene analysis and statistical results show that its average length was 645.55 bp N50 was 957 bp, of which the sequences of 200 to 300 bp account for 33,327 (36.57%) of the total sequence; 26,456 (29.03%) sequences were between 300 and 500 bp; 17,051 (18.71%) sequences 500 to 1,000 bp; 9,067 (9.95%) sequences 1001 to 2000 bp and the sequences of more than 2,000 bp account for 5,233 (5.74%) (Table 2). By processing a large number of sequences obtained by high-throughput RNA-seq, the integrity of unigenes data after assembly is significantly improved and the next step of analysis and statistics can be performed.

Table 2. Length distribution of assembly transcript and unigenes for <i>D. mun</i> .		
Length range (bp)	Unigene	Transcripts
200-300	33,327 (36.57%)	37,823 (24%)
300-500	26,456 (29.03%)	32,955 (20.92%)
500-1000	17,051 (18.71%)	30,838 (19.57%)
1000-2000	9,067 (9.95%)	31,236 (19.83%)
>2000	5,233 (5.74%)	24,701 (15.68%)
Total Number	91,134	157,553
Total Length	58,649,736	166,642,297
N50 Length	957	1,867
Mean Length	645.55	105.77

Functional annotation and classification of unigenes

A total of 91,134 unigenes were obtained through BLAST software in eight major databases (COG, GO, KEGG, KOG, Pfam, Swissprot, eggNOG and Nr) (Table 3). Amongst them, 32,798 unigenes were successfully annotated in Nr, accounting for 35.99% of the total number of unigenes; in KEGG, there were 12,017 successful annotations, accounting for 13.19% of the total; in SwissProt, there were 21,134 successful annotations, accounting for 23.19% of the total; 20,799 successful annotations in Pfam, accounting for 22.82% of the total; 19,929 successful annotations in GO, accounting for 21.87% of the total; 18,499 successfully annotated in KOG, accounting for 20.30% of the total; 9,666 successfully annotated in COG, accounting for 10.61% of the total and, in eggNOG, there are 30,531 successful annotations, accounting for 33.50% of the total. The number of successfully annotated sequences in all eight major databases was 33,421, accounting for 36.67%.

Table 3. Functional annotation of <i>D. mun</i> in different databases.				
Annotated database	Annotated_No.	Percentage (%)	300–1000 (bp)	≥ 1000 (bp)
COG	9,666	10.61	3,022	4,821
GO	19,929	21.87	7,990	7,066
KEGG	12,017	13.19	4,721	4,878
KOG	18,499	20.30	7,399	7,326
Pfam	20,799	22.82	7,706	9,764
Swissprot	21,134	23.19	8,631	8,991
eggNOG	30,531	33.50	12,487	11,544
Nr	32,798	35.99	13,646	11,766
All	33,421	36.67	13,852	11,794

Through the Nr library's comparison, 32,798 unigenes of *D. mun* found similar sequences in the Nr database (Suppl. material 1). The annotation matching species mainly include *Nelumbo nucifera* (6,032; 18.41%), *Macleaya cordata* (3,484; 10.63%), *Vitis vinifera* (1,689; 5.15%), *Elaeis guineensis* (1,240; 3.78%), *Phoenix dactylifera* (1,110; 3.39%), *Aquilegia coerulea* (916; 2.80%), *Quercus suber* (625; 1.91%), *Plicaturopsis crispa* (531; 1.62%), *Amborella trichopoda* (524; 1.60%) and *Juglans regia* (510; 1.56%). A large fraction indicated similarities to genes in other species (16,109; 49.16%). From the annotated information, it can be concluded that most of the sequences of *D. mun* can be matched in angiosperms. In general, from the distribution of sequence similarity, it can be seen that *D. mun* has a high matching degree in the Nr database. However, due to the lack of genome and transcriptome information of *D. mun*, no match in the database was found for some unigenes.

According to the genes successfully annotated by Nr, the GO function classification annotation was performed and the result is shown in Fig. 2. The analysis results show

that 91,134 unigenes have annotated 19,929 GO functions, accounting for 21.87% of the total number of unigenes. Divided into three major functional categories, there are 43,996 gene sequences of cellular component functional categories, accounting for 40.98% of the total; 39,750 gene sequences of biological process functional categories, accounting for 37.02% of the total and 23,617 gene sequences of molecular functional categories, accounting for 22% of the total. It can be observed that the proportion of genes annotated in the functional category of a cellular component is the largest. The three functional categories can be further divided into 52 GO functional subcategories, including 16 cellular components, 15 molecular functions and 21 biological processes. Amongst the 21 functional subcategories in biological processes metabolic processes (1,0697; 26.91%), cellular processes (9,859; 24.8%) and single-organism processes (6,642; 16.71%) have received more annotations. The percentage of annotations obtained during the biological phase in biological processes was the lowest, having only (2; 0.005%). In the category of cellular components, most of the annotations for cell (8,832; 20.07%), cell part (8,784; 19.97%), membrane (7,440; 16.91%) and organelle (6,218; 14.13%) have accounted for the extracellular region part (35; 0.08%) and nucleoid (41; 20.07%) were less annotated. In the molecular function category, there are many annotations for catalytic activity (10,393) and binding (9,895), each accounting for 44% and 41.9% of the total classification, while the annotations for metallochaperone activity (2), translation regulator activity (3) and protein tag (4) account for at least 0.008%, 0.013% and 0.017%, respectively. The above GO function annotation results show the basic situation of gene expression in *D. mun*. It can be seen that amongst the three functional subcategories, there are more genes related to metabolic activities in biological processes, indicating that *D. mun* has a strong metabolic ability.

The obtained unigene is classified and annotated by the KOG protein database and the results are shown in Fig. 3. The analysis results show that 18,499 unigenes in KOG can be matched, accounting for 20.30% of the total. The comparison results can be divided into 25 functional categories, including energy production and conversion, translation, ribosomal structure and biogenesis and different types of gene expression, such as biosynthesis, processing and secondary metabolites biosynthesis, transport and catabolism. "General function prediction only" was the primary group (5,204; 28.13%), followed by "post-translational modification, protein turnover and chaperones" (1,912; 10.34%), "signal transduction mechanisms" (1,731; 9.36%) and "translation, ribosomal structure and biogenesis" (1,139; 6.16%). The cell motility (7; 0.04%) and extracellular structures (48; 0.26%) have the least functional annotation information. *D. mun* has more gene expression in terms of transcription, translation and protein transport. In addition, there is one function unknown (893; 4.83%), whose specific biological function cannot be known.

Comparing the obtained unigenes to the KEGG database, 12,017 unigenes were annotated, accounting for 13.19% of the total. According to the metabolic pathways involved, *D. mun* can be classified into five major categories (Cellular Processes, Environmental Information Processing, Genetic Information Processing, Metabolism, Organismal Systems) and 50 subcategories (Fig. 4). Through the specific analysis and

statistics of unigenes under the relevant pathway classification in Fig. 4, it was found that the metabolism pathways account for the largest proportion of the five categories, with 5,258 unigenes, followed by pathways related to genetic information processing (2,620 unigenes), cellular processes (588 unigenes) and environmental information processing (347 unigenes). Pathways related to organic systems (233) were the fewest. The five major categories were further subdivided into subcategories. Amongst them, metabolism-related pathways were divided into 32 subcategories. In addition, pathways related to genetic information processing were divided into 12 subcategories. In the environmental information processing pathways, there are only two subcategories. In the cellular processes, there were three subcategories. Pathways related to organic systems had only one subcategory. In the KEGG metabolic pathway analysis results, the category with the most annotated genes indicated that *D. mun* has strong metabolic activities.

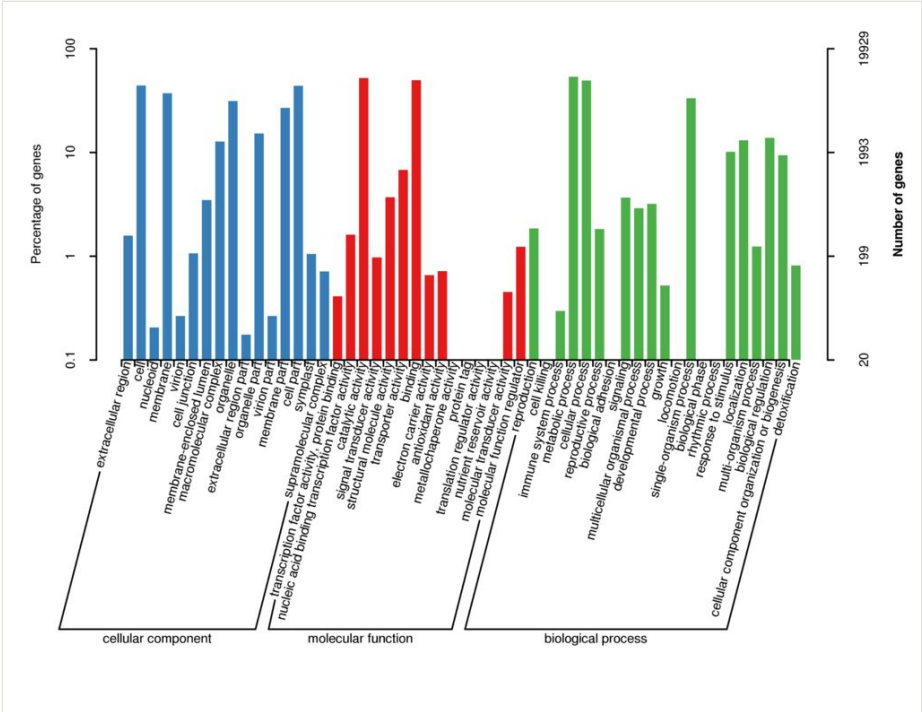


Figure 2. [doi](#)
Gene Ontology (GO) classifications of unigenes of *D. mun*. A total of 19,929 GO functions were categorised into three main categories: biological process, cellular component and molecular function. The x-axis indicates the subgroups in GO annotation, while the y-axis indicates the percentage of specific categories of genes in each main category.

Frequency and distribution of SSRs in the unigenes

All 14,300 assembled unigenes with 28,647,926 bp were examined to discover potential EST-SSRs (Suppl. material 3). A total of 9,391 SSRs were found amongst 91,134

unigenes. Of the 9,391 SSRs, 6,545 SSR-containing sequences were found, as were 2,035 sequences with more than one microsatellite locus and the number of SSRs involved in compound formation (586).

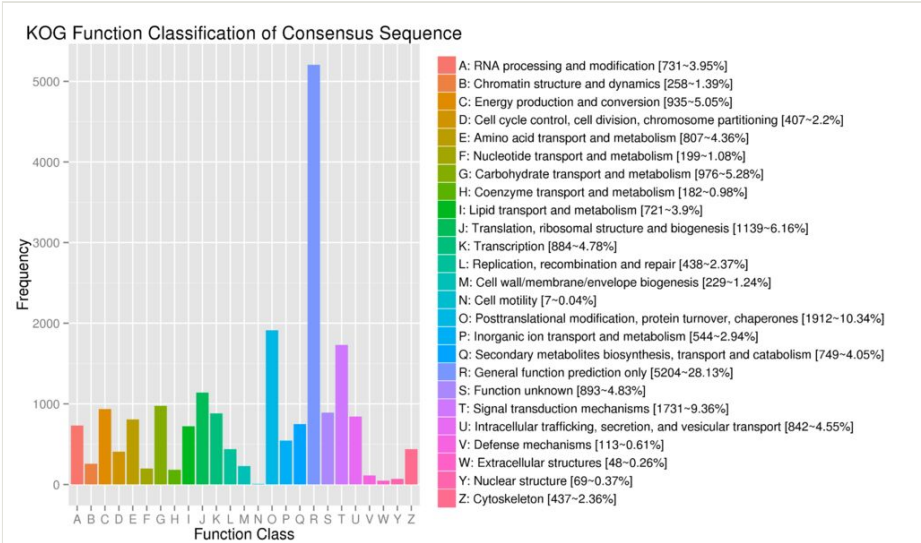


Figure 3. [doi](#)
Eukaryotic Orthologous Groups (KOG) classification of unigenes. Functional predictions of unigenes were classified into at least 25 function classes according to the KOG database.

The frequency of microsatellites in the *D. mun* unigenes was 14.33%. Of the 9391 potential SSRs, distribution to different repeat type classes was identified (Suppl. material 2): mononucleotide and dinucleotide repeats were the most abundant, with 5,030 (53.56%) and 2,384 (25.39%), respectively, followed by trinucleotide repeats (1,865; 19.86%), tetranucleotide repeats (85; 0.91%), pentanucleotide repeats (17; 0.18%) and hexanucleotide repeats (10; 0.11%). It can be seen that the main repeat types of microsatellite sites in the *D. mun* transcriptome are tri-nucleotide repeats and dinucleotide repeats.

Regarding the number of repeat units, the distributions and frequencies of nucleotide repeats are presented in Table 4. The microsatellites that contained ten repeat units were the most prevalent (1,735; 18.48%), followed by those that contained six tandem repeats (1,184; 12.61%) and five tandem repeats (1,150; 12.25%), while the remaining tandem repeats contributed less than 10% of the total SSR.

Statistical analysis of EST-SSR, based on motif types in *D. mun* transcriptome was undertaken (Suppl. material 4). In the mono-nucleotide repeats, A/T had a large proportion (4,981; 53.04%). In the di-nucleotide, the dominant nucleotide repeats were AG/CT (1,852; 19.72%). In the tri-nucleotide, AAG/CTT was the most common (720; 7.67%). In the tetra-nucleotide, AAAT/ATTT was the most common (22; 0.23%). Additionally, in the pentanucleotide and hexanucleotide repeats, the types of repeat

motifs were evenly distributed, which related to the small number of corresponding microsatellite sites.

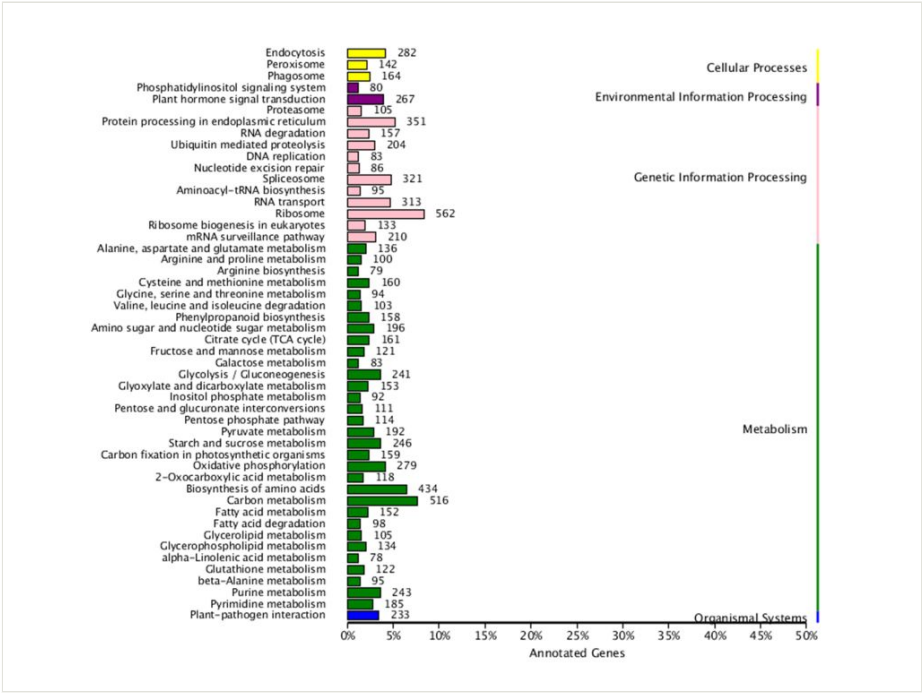


Figure 4. [doi](#)
Kyoto Encyclopaedia of Genes and Genomes (KEGG) metabolism pathway classification of unigenes. Data were classified into five main categories: metabolism, genetic information processing, environmental information processing, cellular processes and organismal systems. Bars represent the number of *D. mun* unigene matches in each category.

Table 4.
Frequency of SSRs, based on repeat types in *D. mun* transcriptome.

Number of repeat	Repeat type						Total	Percentage (%)
	Mono-	Di-	Tri-	Tetra-	Penta-	Hexa-		
5	0	0	1056	75	13	6	1,150	12.25
6	0	633	535	10	2	4	1,184	12.61
7	0	394	251	0	1	0	646	6.88
8	0	409	21	0	0	0	430	4.58
9	0	462	0	0	1	0	463	4.93
10	1354	380	1	0	0	0	1,735	18.48
11	732	104	0	0	0	0	836	8.90
12	565	2	0	0	0	0	567	6.04

Number of repeat	Repeat type						Total	Percentage (%)
	Mono-	Di-	Tri-	Tetra-	Penta-	Hexa-		
13	423	0	0	0	0	0	423	4.50
14	397	0	0	0	0	0	397	4.23
15	346	0	1	0	0	0	347	3.70
16	274	0	0	0	0	0	274	2.92
17	247	0	0	0	0	0	247	2.63
18	217	0	0	0	0	0	217	2.31
19	232	0	0	0	0	0	232	2.47
20	140	0	0	0	0	0	140	1.49
21	77	0	0	0	0	0	77	0.82
22	20	0	0	0	0	0	20	0.21
23	4	0	0	0	0	0	4	0.04
24	2	0	0	0	0	0	2	0.02
Total	5,030	2,384	1,865	85	17	10	9,391	
Percentage (%)	53.56	25.39	19.86	0.91	0.18	0.11		

Genetic diversity

Using 10 different polymorphic EST-SSRs, a study was conducted on the standard genetic diversity and the population genetic structure of three different *D. mun* populations (Table 5). Null allele frequencies were not determined at ten loci for *D. mun* at the significant level of 0.05. From the results, Table 5 shows that $N_A = 3.96$, $N_E = 2.67$, $PIC = 0.77$, $Ho = 0.56$ and $H_E = 0.57$. The F_{IS} value was 0.02. Positive F_{IS} values were found across all four loci under investigation, suggesting an excessive number of homozygotes and inbreeding. It may be deduced that the inbreeding coefficient of the populations ($F_{IT} = 0.06$) suggests an excess of homozygosity in the populations.

Table 5.
Genetic parameters in ten SSR loci for *D. mun*

Locus	Sequence of primer (5'-3')	Motif type	Size (bp)	N_A	N_E	Null allele	PIC	H_o	H_E	F_{IS}	F_{IT}	F_{ST}	P_{HWE}
SSR01	F: GAGGTGGATGAGGTGTTTCGT R: GAGAAGGTTAGGGAGCCAGC	(GGA) ₇	251	5.00	2.66	No	0.86	0.66	0.61	-0.08	-0.03	0.05	***
SSR02	F: TCCCACCAATCATCTTCTCC R: CACAGCAGAACCGCTTTGTA	(TC) ₁₁	193	4.00	3.02	No	0.75	0.85	0.66	-0.28	-0.26	0.02	ns

Locus	Sequence of primer (5'–3')	Motif type	Size (bp)	N_A	N_E	Null allele	PIC	H_o	H_E	F_{IS}	F_{IT}	F_{ST}	P_{HWE}
SSR03	F: TTTTGGTTTCGGGACTTTG R: TCACCAACAACAGGGTTCA	(TA) ₁₁	174	3.33	2.59	No	0.80	0.87	0.60	-0.45	-0.37	0.06	***
SSR04	F: AGCTTCATTCAAAGGCGCTA R: TCATTTCCCTTTGCCAAGAC	(CA) ₁₁	185	4.00	2.81	No	0.74	0.95	0.64	-0.48	-0.47	0.0	***
SSR05	F: CATGTTCAAAGCCATTGTGG R: GCAATTGCAGCACCTCATTA	(GAG) ₁₀	265	3.33	2.43	No	0.76	0.25	0.58	0.58	0.60	0.06	ns
SSR06	F: GGAGCTATCCTCCCCTCAAC R: TCAGGCCATGAGACGTTGTA	(TTTC) ₆	273	4.00	2.14	No	0.72	0.34	0.53	0.37	0.37	0.01	ns
SSR07	F: TGAGGCTTCTTTGGATGCTT R: GTCTGGAATTGGGGGAAAT	(CAT) ₈	216	3.00	2.35	No	0.62	0.76	0.57	-0.33	-0.28	0.04	**
SSR08	F: GCAACTTTGCATGTACCCCT R: AAGGCAAACAGGTTCCAATG	(AAAT) ₆	264	6.33	4.73	No	0.71	0.42	0.74	0.44	0.48	0.07	ns
SSR09	F: AGTGACGACAATAGCGGGTC R: TCTTTTGGTGGGGTTGAAAG	(CAGCAA) ₅	228	4.00	2.77	No	0.92	0.33	0.63	0.49	0.53	0.08	ns
SSR10	F: AATGAAAGACACCCACCCAG R: ATGAGACAGCCGTAATTGGG	(ACCAC) ₉	203	2.67	1.27	No	0.80	0.20	0.20	-0.02	0.03	0.05	***
Mean				3.96	2.67		0.77	0.56	0.57	0.02	0.06	0.04	

Note: N_A : number of alleles; N_E : effective alleles; PIC : polymorphism information content; H_o and H_E : observed and expected heterozygosity; F_{IS} : inbreeding coefficient, Null allele, the average null allele frequency, F_{IT} : coefficient of total inbreeding; F_{ST} : genetic differentiation index of Weir and Cockerham (1984), ns = not significant, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$

The number of polymorphic loci varied between populations for *D. mun* species. The percentage of polymorphic loci was high (100%) in all *D. mun* populations (Table 1). Alleles per locus (N_A) averaged 3.9, ranging from 3.8 in the NH population to 4.1 in the CP population for *D. mun*. The lowest value of effective alleles (N_E) was observed in the NH population ($N_E = 2.3$), whereas this value was the highest in the CP population ($N_E = 2.9$, $H_o = 0.57$ and $H_E = 0.6$). The lowest values of $H_o = 0.53$ and $H_E = 0.55$ were detected in the NH population. The $F_{IS} = 0.029$. Significantly positive F_{IS} values were observed in three populations, showing a deficiency of heterozygotes. The F_{IS}/IM values varied from 0.01 (CP) to 0.03 (NS), with an average of 0.02 (Table 1).

Genetic structure and cluster analysis

The analysis of molecular variance (AMOVA) was implemented, based on 9,999 permutations and revealed the molecular variation attributable to differentiation amongst and within the populations for *D. mun* (Table 6). The AMOVA showed that 4% of the total genetic variation occurred amongst populations and a significant amount (91%, $p < 0.001$) of the total variation occurred within populations. This finding was confirmed by the small overall F_{ST} and large overall Nm , which were 0.045 and 5.34, respectively. The F_{ST} values amongst populations varied from 0.009 (NS and CP) to 0.082 (CP and NH). Significant differentiation was observed for pairwise F_{ST} values for both the *D. mun* species ($p < 0.01$ and 0.001) (Suppl. material 5).

Table 6. Analysis of molecular variance from natural populations for <i>D. mun</i> species produced.					
	df	Sum of squares	Variance components	Total variation (%)	Fixation indices
Amongst populations	2	21.237	0.138	4	$F_{IS} = 0.046^*$ $F_{ST} = 0.045^{**}$ $F_{IT} = 0.089^{**}$
Amongst individuals within populations	79	243.537	0.136	4	$Nm = 5.34$
Within individuals	82	230.500	2.811	91	
Total	163	495,274	3.085	100	
Note: df, degree of freedom; $*P < 0.05$, $^{**}P < 0.01$, $^{***}p < 0.001$					

Without any prior information, discriminant analysis of principal components (DAPC) also uncovered three genetic groupings for *D. mun* (Fig. 5A). Cluster 1 included 14 individuals from the NS population, nine individuals from the NH population and seven individuals from the CP population (Suppl. material 6). Most individuals from the two populations of NH (11 individuals) and CP (14 individuals) were assigned to cluster 2. In cluster 2, there was also one individual in NS. Cluster 3 included most individuals in NS (11 individuals), eight in NH and seven in CP. The DAPC, with prior information on population origin, showed individuals within and between populations (Fig. 5B). The high overlap indicated low genetic differentiation between populations.

The Bayesian analysis of the assignation of individuals, based on the likelihoods, showed that the highest ΔK value (215.9) for 82 individuals revealed $K = 3$ to be an optimum number of genetic groups and showed that all individuals exhibited admixture from three groups (Fig. 5C). The proportion of ancestry linked with each genetic group might be deduced from the colour of each individual, as measured by their percentage of the population's total colour (Fig. 5D). One group (orange) was predominant in the NS populations with strong ancestry values 47.3% and two populations (NH and CP) with lower ancestry values, 29.4% and 16%, respectively (Suppl. material 7). In the second

group (blue), two populations (PH and HSP) represented the largest group (47.7% and 37.9%, respectively), followed by the NS population (16%). Similarly, in the last group (purple), the three populations (NS, NH and CP) were the same (36.8%, 32.7% and 36.3%, respectively). At $K = 2$, the *D. mun* populations were divided into two groups.

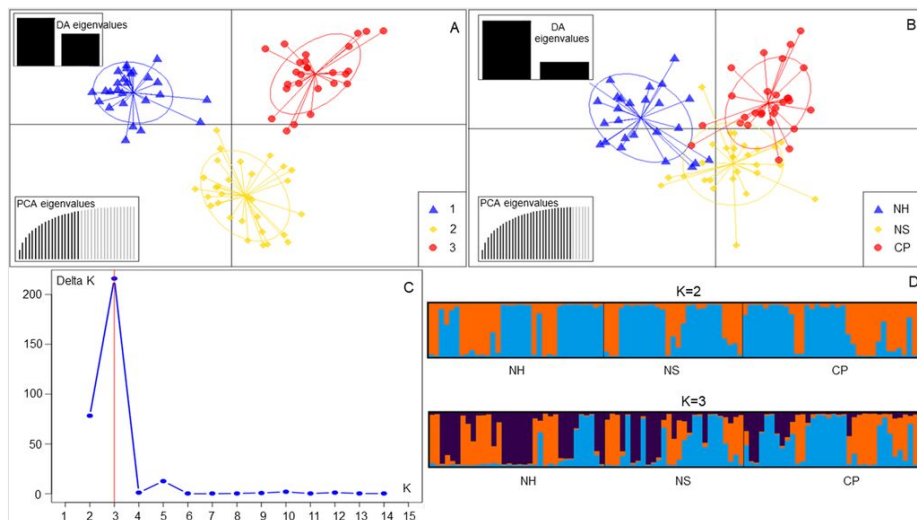


Figure 5. [doi](#)

Analysis of population structure using DAPC and STRUCTURE for three *D. mun* populations. Scatterplot of the DAPC with prior information (A), Scatterplot of the DAPC without prior information (B), Distribution of DeltaK over $K = 1-15$ (C) and Bar plots at $K = 2$ and $K = 3$ (D).

The NJ tree, based on a matrix of Nei's GDs amongst individuals using NTSYS 2.11 and MEGA 11.0, showed that the relationship of 82 *D. mun* individuals mainly had three branches. The first branch consisted of 28 individuals in the NH population and the second largest branch consisted of 54 individuals (CP and NS populations). The NJ tree was consistent with the STRUCTURE and the DAPC analysis and the relationships amongst many individuals were inconsistent with their geographic location (Fig. 6).

Discussion

The Illumina Hiseq4000 sequencing platform was used in this study to determine the transcriptome of *D. mun* and bioinformatics techniques to analyse the microsatellite sites in the transcriptome database. The transcriptome sequencing of *D. mun* provided a valuable resource for developing SSR markers to study genetic diversity, molecular marker-assisted breeding and the evolution of the Ebenaceae family. A total of 91,134 transcriptomic unigenes were obtained using the Illumina HiSeq4000 platform for *D. mun* with a mean length of 645.55 bp and N50 of 957 bp. These lengths are shorter than *Diospyros kaki* Thunb. with N50 = 1,025 bp and average length = 663 bp (Li et al. 2019), indicating that the transcriptome sequencing data were well assembled for *D. mun* in the present study. The methods used in the current study are the same. However, the

deviation might be due to the sequencing depth, assembly method and natural characteristics of the species. After assembling the uni-transcripts, the GC content (46.82%) of *D. mun* was lower than that of *D. kaki* (GC = 48.68% (female) and GC = 48.9 (male)) (Li et al. 2019), although the methodology was the same. The GC content provides information regarding the stability of genes and genomic composition for the evolution and genetic structure of the species. These results can be related to species' adaptability to different environmental conditions. The COG, GO, KEGG, KOG, Pfam, Swiss-Prot and NR databases were searched for sequence matches to determine the genetic diversity of *D. mun* and its development. Annotations of GO categories were made on the *D. mun* unigenes. The metabolic process in the biological process and cell in cellular components were the largest groups in this study, indicating important cellular and metabolic activities. These findings are comparable to those found in an earlier investigation by *D. kaki* (Li et al. 2019). A total of 18,499 unigenes were assigned to 25 KOG categories of *D. mun*. The current KOG database results are parallel to those previously reported for 26 in KOG of *D. kaki* (Li et al. 2019). For this purpose, five major categories and 50 subcategories in KEGG resources were used to study genes' biological functions and interactions. As plant hormones play crucial roles in flower sex differentiation, the plant hormone signal transduction pathway is of particular interest. A total of 267 unigenes were assigned to the plant hormone signal transduction pathway. Additionally, some dominant pathways were ribosome, oxidative phosphorylation, protein processing in the endoplasmic reticulum, glycolysis/gluconeogenesis, spliceosome, RNA transport and purine metabolism. These findings suggest that the *D. mun* unigenes discovered here have broad applicability and will prove useful for future studies of the species' functional diversity.

According to Zalapa et al. (2012), having longer sequences will enhance the likelihood of successfully developing EST-SSR primers, thus *D. mun* was chosen as a prime target for microsatellite analysis in this study. Furthermore, the results indicate 91,134 unigenes in the transcriptome of *D. mun*, which have 8,805 microsatellite loci. At the same time, there are abundant types of microsatellites in the transcriptome of *D. mun*, with six different types of nucleotide repeats, of which the main repeat type is mono-nucleotide repeats (53.57%) and di-nucleotide repeats (25.37%), which are similar to *D. kaki* (Li et al. 2019, Wang et al. 2020). Of these identified SSRs, the (A/T)_n nucleotide repeat was the most abundant.

Improvements to germplasm resources through genetic diversity have been extensively implemented in several other *Diospyros* (Deng et al. 2015, Yang et al. 2015, Wang et al. 2018, Wang et al. 2020, Samarina et al. 2021, Samarina et al. 2022). There is typically a correlation between the number of loci and populations within a species and the level of genetic diversity that exists within that species (Vu et al. 2018), the size of the geographical range that the species inhabits (Nguyen et al. 2021) and genetic exchange (Hellmann et al. 2008). According to the evaluation criteria of polymorphism ($PIC > 0.5$) (Botstein et al. 1980), the current study found evidence of high genetic variation, based on the EST-SSR ($PIC = 0.77$). Due to the relatively high PIC values, it can be deduced that the chosen loci are both highly informative and appropriate for conducting genetic

research on *D. mun*. Several genetic measurements, including the mean number of observed alleles, observed heterozygosity, expected heterozygosity and the percentage of the polymorphic band, corroborate this finding (Liang et al. 2015, Wang et al. 2018, Wang et al. 2020, Samarina et al. 2021). *D. mun* exhibited a moderate degree of genetic diversity ($H_E = 0.58$; $H_o = 0.56$) compared to other *Diospyros*, such as *D. kaki* ($H_o = 0.694$) and *D. lotus* ($H_o = 0.2$) (Samarina et al. 2021) and *D. kaki* ($H_o = 0.795$) (Wang et al. 2018). Furthermore, our results showed that the observed number of alleles for each locus in *D. mun* ($N_A = 3.9$) was lower than that compared to *D. kaki* ($N_A = 11.0$) (Samarina et al. 2021). The extensive genetic diversity within a species is a reflection of its evolutionary past and the ecogeographic environment in which it has thrived. It is common for species to maintain a high genetic diversity when they have a wide distribution pattern, a large population size, a long lifespan, a predominance of outcrossing and successional stages (Hamrick and Godt 1996, White et al. 2007). It is hypothesised that species with high genetic heterozygosity will have a greater capacity for adaptability in their environment. This hypothesis is supported by the observation that *Diospyros* species have a broad distribution and live for a very long time. Genetic heterozygosity is related to anthropogenic disturbance and can be reduced through genetic drift and increased homozygosity for common alleles due to the loss of rare alleles (Honnay and Jacquemyn 2007, Gijbels et al. 2015). Genetic heterozygosity in NS, NH and CP was moderate, suggesting that human activity may have contributed (Liu et al. 2023, Yan et al. 2024). In all populations examined, deforestation and overexploitation are potential major factors reducing genetic heterozygosity. Most of the *D. mun* trees with more than 30 cm diameter (dbh) were exploited by forestry enterprises in the 1980s and 1990s (MOST and VAST 2007). The recent populations include trees with age less than 100 years that have fragmented into small populations. The short evolutionary history and tiny population sizes of *D. mun* may account for its moderate genetic diversity. An effective population size and genetic variety decreasing due to inbreeding may have resulted from anthropogenic disturbance and population declines. Significant heterozygosity deficits across 82 trees from three populations suggest that a relatively high heterozygosity deficiency can exist within the *D. mun* populations. This shows that inbreeding exists in small populations, although this species is outcrossed (Murawski et al. 1994, Kenta et al. 2002, Lee et al. 2006).

Gene flow and random mutations contribute to genetic diversity amongst populations (Schaal et al. 2002). Low population genetic differentiation reflects high gene flow. Strong gene flow ($N_m > 1$) determined a high number of migrants for each generation and may prevent genetic differentiation amongst populations due to genetic drift (Slatkin 1987). According to the findings of our study, a gene flow rate that was comparatively high ($N_m = 6.34$) had the potential to reduce the impact of genetic drift by lowering the population's level of genetic differentiation, while simultaneously raising the level of genetic heterozygosity. Gene flow is determined via the dispersal of pollen grains and seeds (Finkeldey and Hattermer 2007, Brütting et al. 2012, Budd et al. 2015) and contributes to population genetic differences and genetic structure. *D. mun* are long-lived, predominantly outcrossed, insect-pollinated and late successional (Nguyen 2003, Ngo 2014). Their seeds are dispersed by wind and water (Nguyen 2003). Insects are

responsible for pollinating *D. mun*, while wind is responsible for spreading its seeds. The increased isolation distance will result in a reduction in the amount of pollen that is transmitted from population to population of *D. mun*. Pollen transmission is primarily associated with insects. The current population is fragmented into smaller populations. Thus, habitat fragmentation leading to increased isolation can decrease gene exchange and influence its genetic structure in recent decades. The genetic structure of *D. mun* in the present study was also detected via different clustering analyses (Structure, DAPC and NJ), in which 82 studied trees were grouped, based on their geographic origin. Therefore, the genetic structure's existence could result from gene exchange related to geographic distance and habitat fragmentation that separates populations into different genetic groups. Thus, a decrease in the dispersal of pollen and seeds might limit gene flow for *D. mun*.

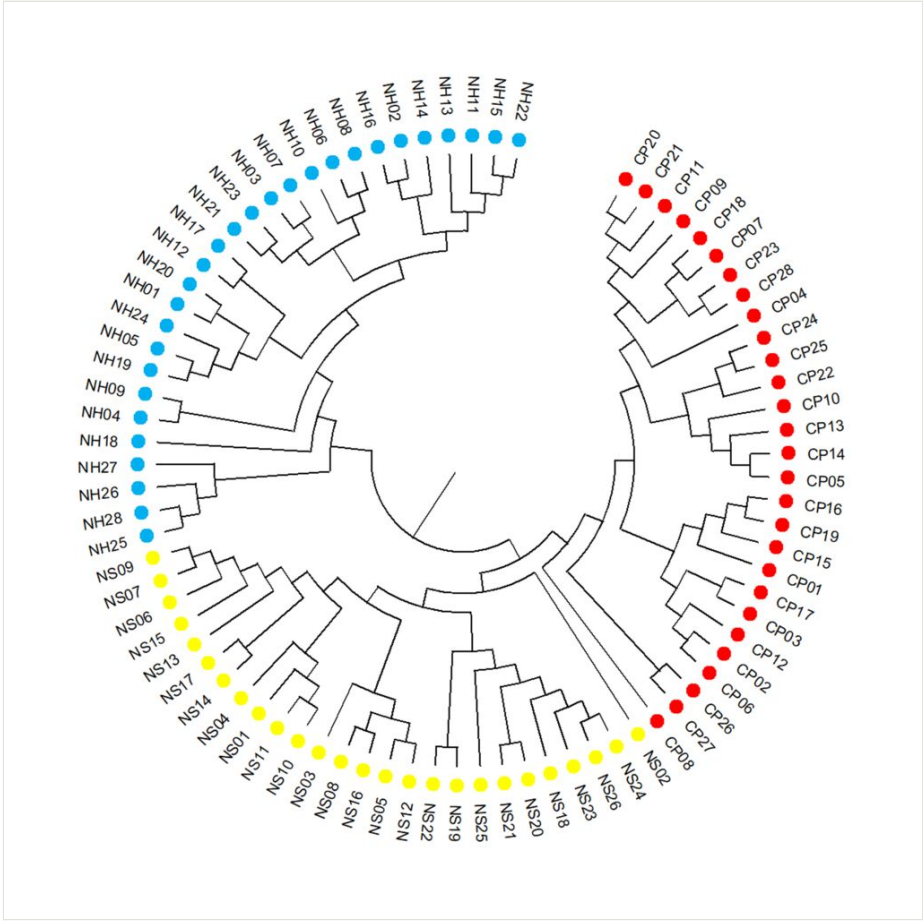


Figure 6. [doi](#)
Neighbour-Joining (NJ) tree was constructed, based on Nei's GDs of 82 genotypes of *D. mun*, based on 10 polymorphic SSR loci.

Conclusions

The *D. mun* transcriptome was sequenced in its entirety for the first time using the Illumina 4000 sequencing technology. Many ESTs were generated and potentially differently expressed genes in *D. mun* were found. A total of 9,391 EST-SSRs were discovered and 10 microsatellite sites were successful. It is clear from the results that the naturally occurring populations of *D. mun* have maintained a reasonable amount of genetic variation over time. Several SSR markers were found, which will be useful in the future marker-assisted breeding of *D. mun*. We determined a moderate genetic variation and low genetic difference among the three populations of *D. mun*. The structure and the DAPC identified three main genetic groups besides moderate genetic diversity in three populations of *D. mun* species. This research lays the foundation for conservation efforts to protect the species' genetic diversity.

Acknowledgements

The study was supported by the Institute of Ecology and Biological Resources (IEBR), Vietnam Academy of Science and Technology (VAST) under grant number CSCLO9.01/24-25.

Author contributions

Duy Dinh Vu and Thi Tuyet Xuan Bui: conception and design of this study; Thi Tuyet Xuan Bui and Duy Dinh Vu collection of samples; Thi Tuyet Xuan Bui: implementation of experiments; Duy Dinh Vu and Thi Tuyet Xuan Bui: data analysis and writing – the original draft of the manuscript, writing – review and editing. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Botstein D, White R, Skolnick M, Davis R (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* 32 (3): 314-331.
- Brütting C, Hensen I, Wesche K (2012) *Ex situ* cultivation affects genetic structure and diversity in arable plants. *Plant Biology* 15 (3): 505-513. <https://doi.org/10.1111/j.1438-8677.2012.00655.x>
- Budd C, Zimmer E, Freeland J (2015) Conservation genetics of *Magnolia acuminata*, an endangered species in Canada: Can genetic diversity be maintained in fragmented,

peripheral populations? Conservation Genetics 16 (6): 1359-1373. <https://doi.org/10.1007/s10592-015-0746-9>

- Chybicki I, Burczyk J (2008) Simultaneous estimation of null alleles and inbreeding coefficients. Journal of Heredity 100 (1): 106-113. <https://doi.org/10.1093/jhered/esn088>
- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21 (18): 3674-6. <https://doi.org/10.1093/bioinformatics/bti610>
- Deng L, Liang Q, He X, Luo C, Chen H, Qin Z (2015) Investigation and analysis of genetic diversity of *Diospyros* germplasms using SCoT molecular markers in Guangxi. PLOS One 10 (8). <https://doi.org/10.1371/journal.pone.0136510>
- Deng YY, Li JQ, Wu SF, Zhu YP, Chen YW, He FC (2006) Integrated nr database in protein annotation system and its localization. Computer Engineering 32 (5): 71-74.
- Durmaz AA, Karaca E, Demkow U, Toruner G, Schoumans J, Cogulu O (2015) Evolution of Genetic Techniques: Past, Present, and Beyond. BioMed Research International 2015: 1-7. <https://doi.org/10.1155/2015/461524>
- Earl DA, von Holdt BM (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources 4 (2): 359-361. <https://doi.org/10.1007/s12686-011-9548-7>
- Egan A, Schlueter J, Spooner D (2012) Applications of next-generation sequencing in plant biology. American Journal of Botany 99 (2): 175-185. <https://doi.org/10.3732/ajb.1200020>
- Evano G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14 (8): 2611-2620. <https://doi.org/10.1111/j.1365-294x.2005.02553.x>
- Excoffier L, Laval G, Schneider S (2017) Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evolutionary Bioinformatics 1 <https://doi.org/10.1177/117693430500100003>
- Finkeldey R, Hatterer H (2007) Tropical Forest Genetics. Tropical Forestry <https://doi.org/10.1007/978-3-540-37398-8>
- Finn R, Bateman A, Clements J, Coggill P, Eberhardt R, Eddy S, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M (2013) Pfam: the protein families database. Nucleic Acids Research 42 <https://doi.org/10.1093/nar/gkt1223>
- Gijbels P, De Hert K, Jacquemyn H, Honnay O (2015) – Reduced fecundity and genetic diversity in small populations of rewarding versus deceptive orchid species: a meta-analysis. Plant Ecology and Evolution 148 (2): 153-159. <https://doi.org/10.5091/plecevo.2015.914>
- Govindaraj M, Vetriventhan M, Srinivasan M (2015) Importance of Genetic Diversity Assessment in Crop Plants and Its Recent Advances: An Overview of Its Analytical Perspectives. Genetics Research International 2015: 1-14. <https://doi.org/10.1155/2015/431487>
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29 (7): 644-652. <https://doi.org/10.1038/nbt.1883>

- Guo DL, Luo ZR (2011) Genetic relationships of the Japanese persimmon *Diospyros kaki* (Ebenaceae) and related species revealed by SSR analysis. *Genetics and Molecular Research* 10 (2): 1060-8. <https://doi.org/10.4238/vol10-2gmr1100>
- Hamrick JL, Godt M (1996) Conservation genetics of endemic plant species. *Conservation Genetics* 281-304. https://doi.org/10.1007/978-1-4757-2504-9_9
- Hedrick P (2005) A standardized genetic differentiation measure. *Evolution* 59 (8). <https://doi.org/10.1554/05-076.1>
- Hellmann I, Mang Y, Gu Z, Li P, de la Vega F, Clark A, Nielsen R (2008) Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Research* 18 (7): 1020-1029. <https://doi.org/10.1101/gr.074187.107>
- Hinderer E, Moseley HB (2020) GOcats: A tool for categorizing gene ontology into subgraphs of user-defined concepts. *PLOS One* 15 (6). <https://doi.org/10.1371/journal.pone.0233311>
- Honnay O, Jacquemyn H (2007) Susceptibility of common and rare plant species to the genetic consequences of habitat fragmentation. *Conservation Biology* 21 (3): 823-831. <https://doi.org/10.1111/j.1523-1739.2006.00646.x>
- Jakobsson M, Rosenberg N (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23 (14): 1801-1806. <https://doi.org/10.1093/bioinformatics/btm233>
- Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* 11 (1). <https://doi.org/10.1186/1471-2156-11-94>
- Kalonowski S, Taper M, Marshall T (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology* 16 (5): 1099-1106. <https://doi.org/10.1111/j.1365-294x.2007.03089.x>
- Kanehisa M (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Research* 32 (90001). <https://doi.org/10.1093/nar/gkh063>
- Kenta T, Shimizu K, Nakagawa M, Okada K, Hamid A, Nakashizuka T (2002) Multiple factors contribute to outcrossing in a tropical emergent *Dipterocarpus tempehes*, including a new pollen-tube guidance mechanism for self-incompatibility. *American Journal of Botany* 89 (1): 60-66. <https://doi.org/10.3732/ajb.89.1.60>
- Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology* 5 (2). <https://doi.org/10.1186/gb-2004-5-2-r7>
- Lee SL, Wickneswari R, Mahani MC, Zakri AH (2006) Mating system parameters in a tropical tree species, *Shorea leprosula* Miq. (Dipterocarpaceae), from Malaysian Lowland dipterocarp forest. *Biotropica* 32: 693-702. <https://doi.org/10.1111/j.1744-7429.2000.tb00517.x>
- Liang Y, Han W, Sun P, Liang J, Wuyun T, Li F, Fu J (2015) Genetic diversity among germplasms of *Diospyros kaki* based on SSR markers. *Scientia Horticulturae* 186: 180-189. <https://doi.org/10.1016/j.scienta.2015.02.015>
- Li P, Xiao L, Du Q, Quan M, Song Y, He Y, Huang W, Xie J, Lv C, Wang D, Zhou J, Li L, Liu Q, El-Kassaby Y, Zhang D (2023) Genomic insights into selection for heterozygous

- alleles and woody traits in *Populus tomentosa*. Plant Biotechnology Journal 21 (10): 2002-2018. <https://doi.org/10.1111/pbi.14108>
- Li S, Sun P, Du G, Wang L, Li H, Fu J, Suo Y, Han W, Diao S, Mai Y, Li F (2019) Transcriptome sequencing and comparative analysis between male and female floral buds of the persimmon (*Diospyros kaki* Thunb.). Scientia Horticulturae 246: 987-997. <https://doi.org/10.1016/j.scienta.2018.11.073>
 - Liu J, Magige E, Fan P, Wambulwa M, Luo Y, Qi H, Gao L, Milne R (2023) Genetic imprints of grafting in wild iron walnut populations in southwestern China. BMC Plant Biology 23 (1). <https://doi.org/10.1186/s12870-023-04428-z>
 - Li Y, Zhang P, Chachar S, Xu J, Yang Y, Guan C (2023) A comprehensive evaluation of genetic diversity in persimmon (*Diospyros kaki* Thunb.) germplasms based on large-scale morphological traits and SSR markers. Scientia Horticulturae 313 <https://doi.org/10.1016/j.scienta.2023.111866>
 - Luo C, Zhang F, Zhang QL, Guo DY, Luo ZR (2013) Characterization and comparison of EST-SSR and TRAP markers for genetic analysis of the Japanese persimmon *Diospyros kaki*. Genetics and Molecular Research 12 <https://doi.org/10.4238/2013.january.9.3>
 - Lussi YC, Magrane M, Martin MJ, Orchard S (2023) Searching and navigating UniProt databases. Current Protocols 3 (3): e700. <https://doi.org/10.1002/cpz1.700>
 - MOST, VAST (2007) Red data book of Vietnam. Natural Sciences and Technology Publishing House [In Vietnamese].
 - Murawski D, Dayanandan B, Bawa K (1994) Outcrossing rates of two endemic *Shorea* species from Sri Lankan tropical rain forests. Biotropica 26 (1). <https://doi.org/10.2307/2389107>
 - Naval MM, Zuriaga E, Pecchioli S, Ll  cer G, Giordani E, Badenes ML (2010) Analysis of genetic diversity among persimmon cultivars using microsatellite markers. Tree Genetics & Genomes 6 (5): 677-687. <https://doi.org/10.1007/s11295-010-0283-0>
 - Ngo VN (2014) Some silvicultural characteristics of *Diospyros mun* A.Chev. ex Lecomte in Cuc Phuong National Park. Vietnam Journal of forest science 2: 3302-3307. [In Vietnamese].
 - Nguyen HN (1998) *Diospyros mun*. IUCN Red List of Threatened Species <https://doi.org/10.2305/iucn.uk.1998.rlts.t32821a9732503.en>
 - Nguyen TB (2003) Checklist of plant species of Vietnam. Agriculture Publishing House [In Vietnamese].
 - Nguyen TM, Vu DD, Dang HP, Bui XTT, Nguyen HPL, Nguyen DM (2021) Population genetic structure and demographic history of the dipterocarp species *Anisoptera costata* Korth revealed by microsatellite analysis. Planta 253 (3). <https://doi.org/10.1007/s00425-021-03584-6>
 - Peakall R, Smouse P (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. Bioinformatics 28 (19): 2537-2539. <https://doi.org/10.1093/bioinformatics/bts460>
 - Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19 (5): 651-652. <https://doi.org/10.1093/bioinformatics/btg034>
 - Pham M, Vu DD, Bei C, Bui TTX, Vu DG, Shah SNM (2024) Characterisation of the *Cinnamomum parthenoxylon* (Jack) Meisn (Lauraceae) transcriptome using Illumina

paired-end sequencing and EST-SSR markers development for population genetics. Biodiversity Data Journal 12 <https://doi.org/10.3897/bdj.12.e123405>

- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155 (2): 945-959. <https://doi.org/10.1093/genetics/155.2.945>
- Rosenberg N (2003) <scp>distruct</scp>: a program for the graphical display of population structure. Molecular Ecology Notes 4 (1): 137-138. <https://doi.org/10.1046/j.1471-8286.2003.00566.x>
- Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4: 406-425. <https://doi.org/10.1093/oxfordjournals.molbev.a040454>
- Salgotra RK, Chauhan BS (2023) Genetic Diversity, Conservation, and Utilization of Plant Genetic Resources. Genes 14 (1). <https://doi.org/10.3390/genes14010174>
- Samarina L, Malyarovskaya V, Reim S, Koninskaya N, Matskiv A, Tsaturyan G, Rakhmangulov R, Shkhalakhova R, Shurkina E, Kulyan R, Omarova Z, Omarov M, Ryndin A (2021) Genetic diversity in *Diospyros* germplasm in the Western Caucasus based on SSR and ISSR polymorphism. Biology 10 (4). <https://doi.org/10.3390/biology10040341>
- Samarina L, Malyarovskaya V, Rakhmangulov R, Koninskaya N, Matskiv A, Shkhalakhova R, Orlov Y, Tsaturyan G, Shurkina E, Gvasaliya M, Kuleshov A, Ryndin A (2022) Population analysis of *Diospyros lotus* in the Northwestern Caucasus based on leaf morphology and multilocus DNA markers. International Journal of Molecular Sciences 23 (4). <https://doi.org/10.3390/ijms23042192>
- Schaal BA, Hayworth DA, Olsen KM, Rauscher JT, Smith WA (2002) Phylogeographic studies in plants: problems and prospects. Molecular Ecology 7 (4): 465-474. <https://doi.org/10.1046/j.1365-294x.1998.00318.x>
- Slatkin M (1987) Gene flow and the geographic structure of natural populations. Science 236 (4803): 787-792. <https://doi.org/10.1126/science.3576198>
- Soriano JM, Pecchioli S, Romero C, Vilanova S, Llacer G, Gioranil E, Badenes ML (2006) Development of microsatellite markers in polyploid persimmon (*Diospyros kaki*) from an enriched genomic library. Molecular Ecology Notes 6 (2): 368-370. <https://doi.org/10.1111/j.1471-8286.2006.01236.x>
- Tamura K, Stecher G, Kumar S (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Molecular Biology and Evolution 38 (7): 3022-3027. <https://doi.org/10.1093/molbev/msab120>
- Van Oosterhot C, Hutchinson W, Wills DM, Shipley P (2004) microchecker: software for identifying and correcting genotyping errors in microsatellite data. Molecular Ecology Notes 4 (3): 535-538. <https://doi.org/10.1111/j.1471-8286.2004.00684.x>
- Vlč D, Řepková J (2017) Application of next-generation sequencing in plant breeding. Czech Journal of Genetics and Plant Breeding 53 (3): 89-96. <https://doi.org/10.17221/192/2016-cjgpb>
- Vu D, Bui TT, Nguyen M, Shah SNM, Vu D, Zhang Y, Nguyen M, Huang X (2018) Genetic diversity and conservation of two threatened dipterocarps (Dipterocarpaceae) in southeast Vietnam. Journal of Forestry Research 30 (5): 1823-1831. <https://doi.org/10.1007/s11676-018-0735-1>

- Vu DD, Ngo VN, Do TPT, Hoang TTT, Phan MQ, Bui TTX (2014) Genetic relationship of *Diospyros mun* a. Chev. Ex Lecomte populations in Vietnam based on analysis of ITS-rDNA sequences. [Vietnam Journal of Biotechnology](#) 12 (2): 269-279. [In Vietnamese].
- Vu DD, Bui TTX, Nguyen MT, Vu DG, Nguyen MD, Bui VT, Huang X, Zhang Y (2016) Genetic diversity in two threatened species in Vietnam: *Taxus chinensis* and *Taxus wallichiana*. *Journal of Forestry Research* 28 (2): 265-272. <https://doi.org/10.1007/s11676-016-0323-1>
- Vu DD, Shah SNM, Pham MP, Bui VT, Nguyen MT, Nguyen TPT (2020) De novo assembly and transcriptome characterization of an endemic species of Vietnam, *Panax vietnamensis* Ha et Grushv., including the development of EST-SSR markers for population genetics. *BMC Plant Biology* 20 (1): 358. <https://doi.org/10.1186/s12870-020-02571-5>
- Wang L, Li H, Sun P, Fu J, Suo Y, Zhang J, Han W, Diao S, Li F, Mai Y (2018) Genetic diversity among wild androecious germplasms of *Diospyros kaki* in China based on SSR markers. *Scientia Horticulturae* 242: 1-9. <https://doi.org/10.1016/j.scienta.2018.07.020>
- Wang L, Li H, Suo Y, Han W, Diao S, Mai Y, Sun P, Fu J (2020) Development of EST-SSR markers and their application in the genetic diversity of persimmon (*Diospyros kaki* Thunb.). *Trees* 35 (1): 121-133. <https://doi.org/10.1007/s00468-020-02024-4>
- Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38 (6). <https://doi.org/10.2307/2408641>
- White T, Adams W, Neale D (2007) *Forest genetics*. CABI Publishing, Boston, MA, USA <https://doi.org/10.1079/9781845932855.0000>
- Yang Y, Jing ZB, Ruan XF, Cheng JM (2015) Development of simple sequence repeat markers in persimmon (*Diospyros* L.) and their potential use in related species. *Genetics and Molecular Research* 14 (1): 609-618. <https://doi.org/10.4238/2015.january.30.2>
- Yan L, Fan P, Wambulwa M, Qi H, Chen Y, Wu Z, Milne R, Khan R, Luo Y, Gao L, Shen S, Rashid I, Khan SM, Maity D, Li D, Liu J (2024) Human-associated genetic landscape of walnuts in the Himalaya: implications for conservation and utilization. *Diversity and Distributions* 30 (4). <https://doi.org/10.1111/ddi.13809>
- Zalapa J, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E, McCown B, Harbut R, Simon P (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *American Journal of Botany* 99 (2): 193-208. <https://doi.org/10.3732/ajb.1100394>
- Zhang C, Wu Z, Jiang X, Li W, Lu Y, Wang K (2021) De novo transcriptomic analysis and identification of EST-SSR markers in *Stephanandra incisa*. *Scientific reports* 11 (1): 1059. <https://doi.org/10.1038/s41598-020-80329-7>

Supplementary materials

Suppl. material 1: Distribution of species search of unigenes against the Nr database. [doi](#)

Authors: Thi Tuyet Xuan Bui, Dinh Duy Vu

Data type: images

[Download file](#) (102.41 kb)

Suppl. material 2: Distribution type of EST-SSRs of *D. mun* [doi](#)

Authors: Thi Tuyet Xuan Bui, Dinh Duy Vu

Data type: images

[Download file](#) (13.82 kb)

Suppl. material 3: Summary of analyses of expressed sequence Tag–Simple Sequence repeat (EST-SSRs) in *D. mun* [doi](#)

Authors: Thi Tuyet Xuan Bui, Dinh Duy Vu

Data type: Genomic

[Download file](#) (12.83 kb)

Suppl. material 4: Frequency distribution of SSRs based on motif types in *D. mun* transcriptome [doi](#)

Authors: Thi Tuyet Xuan Bui, Dinh Duy Vu

Data type: Genomic

[Download file](#) (21.97 kb)

Suppl. material 5: Pairwise genetic differentiation (F_{st}) between populations for *D. mun* species [doi](#)

Authors: Thi Tuyet Xuan Bui, Dinh Duy Vu

Data type: Genomic

[Download file](#) (12.85 kb)

Suppl. material 6: Number of individuals for each population assigned. Each cluster was obtained from DAPC without prior information [doi](#)

Authors: Thi Tuyet Xuan Bui, Dinh Duy Vu

Data type: Genomic

[Download file](#) (12.36 kb)

Suppl. material 7: Percentage of ancestry for three *D. mun* populations [doi](#)

Authors: Thi Tuyet Xuan Bui, Dinh Duy Vu

Data type: Genomic

[Download file](#) (12.52 kb)