Functional & Integrative Genomics (2023) 23:51
https://doi.org/10.1007/s10142-023-00975-8

DATA NOTE

# Genome-wide characterization leading to simple sequence repeat (SSR) markers development in *Shorea robusta*

**Garima Mishra[1] · Rajendra K. Meena[1] · Rama Kant[1] · Shailesh Pandey[2] · Harish S. Ginwal[1] · Maneesh S. Bhandari[1]**

Received: 4 January 2023 / Revised: 18 January 2023 / Accepted: 19 January 2023 / Published online: 28 January 2023


## Abstract

Tropical rainforests in Southeast Asia are enriched by multifarious biota dominated by Dipterocarpaceae. In this family, *Shorea robusta* is an ecologically sensitive and economically important timber species whose genomic diversity and phylogeny remain understudied due to lack of datasets on genetic resources. Smattering availability of molecular markers impedes population genetic studies indicating a necessity to develop genomic databases and species-specific markers in *S. robusta*. Accordingly, the present study focused on fostering de novo low-depth genome sequencing, identification of reliable microsatellites markers, and their validation in various populations of *S. robusta* in Uttarakhand Himalayas. With 69.88 million raw reads assembled into 1,97,489 contigs (read mapped to 93.2%) and a genome size of 357.11 Mb (29 × coverage), Illumina paired-end sequencing technology arranged a library of sequence data of ~ 10 gigabases (Gb). From 57,702 microsatellite repeats, a total of 35,049 simple sequence repeat (SSR) primer pairs were developed. Afterward, among randomly selected 60 primer pairs, 50 showed successful amplification and 24 were found as polymorphic. Out of which, nine polymorphic loci were further used for genetic analysis in 16 genotypes each from three different geographical locations of Uttarakhand (India). Prominently, the average number of alleles per locus (*N*a), observed heterozygosity (*H*o), expected heterozygosity (*H*e), and the polymorphism information content (PIC) were recorded as 2.44, 0.324, 0.277 and 0.252, respectively. The accessibility of sequence information and novel SSR markers potentially enriches the current knowledge of the genomic background for *S. robusta* and to be utilized in various genetic studies in species under tribe Shoreae.

**Keywords** *Shorea robusta* · Genome sequencing · Microsatellite markers · Null alleles · Genetic diversity

## Introduction

Forests hold in excess of 75 percent of the earth biodiversity (Cvetković et al. 2019), wherein tropical forests are one of the key-category of luxuriant land ecosystems which incorporate world's most diverse habitat types flourishing many dominant tree species. In tropical Asia, particularly the Indian subcontinent, *Shorea robusta* Gaertn. F. (Vern: Sal), a diploid (2n = 2x = 14) outcrossing species belonging to the family Dipterocarpaceae, have fundamental ecological and evolutionary significance besides being utilized for commercial timber production worldwide (Gautam et al. 2007). Further, the species has usages consisting of medicinal, fodder and fuelwood, being consumed by the locals and forest dwelling communities (Adhikari et al. 2017). Due to substantial overexploitation and habitat fragmentation of the tropical forests, the species' range is in an alarming state, threatening the long-term maintenance of its genetic

✉ Maneesh S. Bhandari
bhandarims@icfre.org; maneesh31803@gmail.com

Garima Mishra
garimamishra183@gmail.com

Rajendra K. Meena
rajendra@icfre.org; rajnrcpb@gmail.com

Rama Kant
ramakant@icfre.org; rgpb82@gmail.com

Shailesh Pandey
pandeysh@icfre.org; shailesh31712@gmail.com

Harish S. Ginwal
ginwalhs@icfre.org

[1] Division of Genetics & Tree Improvement, Forest Research Institute, Dehradun - 248 195, Uttarakhand, Dehradun, India

[2] Forest Pathology Discipline, Division of Forest Protection, Forest Research Institute, Dehradun - 248 006, Uttarakhand, Dehradun, India

diversity and survival (Gautam and Devoe 2006). Recently, a SWOT analysis on the status of the century-old regeneration problem of *S. robusta* has been conducted, which revealed the need of molecular marker development and genetic diversity assessment for this vital species (Mishra et al. 2020). By and large, genetic knowledge of the tropical forest species is more limited than that of the temperate or boreal forests (Finkeldey and Hattemer 2007). Hence, a need exists for a comprehensive analysis of population genetics in *S. robusta*, which will be able to divulge the present status of gene flow, genetic diversity and population structure. For such genetic analysis, suitable molecular marker techniques are vital. Schulman (2007) stated, "since before the beginning of molecular markers, the use of traits in plants as markers for their genetic relationship predates genetics itself", which illustrates the usage and essentiality of molecular markers for genetic-based studies.

During the last three decades, the world has witnessed a rapid increase in the knowledge about plant genomic sequences, and the physiological and molecular role of various plant genes, which has revolutionized the population genetics and its proficiency in improvement programmes of a species (Nadeem et al. 2018). Yet, to date, very few researches have explored the genetic diversity and population structure in natural populations of *S. robusta*. Previous studies analyzed genetic diversity of *S. robusta* based on isozyme and Inter Simple Sequence Repeats (ISSR) markers (Suoheimo et al. 1999; Surabhi et al. 2017). However, an overview of the genetic and population structure is presently not available for this premier timber resource in subcontinents, and only limited information could be extracted due to the scarcity of markers. Therefore, there is a necessity to use a robust marker system for population genetic analysis, and to fulfill that, Simple Sequence Repeats (SSRs) are the markers of primary choice due to several desirable features, such as codominance, high variability, reproducibility, wide genomic coverage, extensive information, and accessibility (Powell et al. 1996; Nybom 2004; López-Gartner et al. 2009; Wang et al. 2019). Despite recent advances in molecular markers, such as Single-Nucleotide Polymorphisms (SNPs) or DNA array-based markers, SSRs hold promise as breeder-friendly markers involving limited technical or operating difficulties.

Considering the above facts, the proposed work demonstrate the first low-depth genomic sequence data of *S. robusta* with main objectives aimed to: (1) provide high-quality sequence data and enrich the current knowledge of the genomic background for *S. robusta*; (2) identify and develop novel SSR markers based on the sequence-specific information; (3) functionally annotate the designed SSRs using public databases; and finally (4) validation of the polymorphic SSR markers in *S. robusta* populations for the authentication of markers discovered.

## Material and methods

### Plant materials and DNA isolation

Based on a wide-ranged field survey, forty-eight indigenous accessions of *S. robusta* were collected from three different geographical locations in the state Uttarakhand (India), with their geospatial features (viz. longitude, latitude and altitude) shown in Supplementary Table 1. Sampling of seedlings nearby each other could be closely related and hence, less variation can be observed. Thus, sampled leaves were randomly collected from the trees representing size class (DBH) variations with 300 m distance apart, with populations distributed evenly in a wider area to capture as much diversity as possible. Samples were immediately dried up with silica gel and brought to the laboratory of Genetics and Tree Improvement Division, Forest Research Institute, Dehradun, and stored at – 80 °C. The genomic DNA was isolated from leaf tissues using Doyle and Doyle (1990) protocol with minor modifications.

### Illumina sequencing, library construction and genome assembly preparation

The arrangement of base pairs in a genomic DNA was determined using a molecular technique known as Illumina dye sequencing. The sequencing was performed by the M/s Clevergene Biocorp Private Limited (Bengaluru, Karnataka) with HiSeq X System (Illumina, San Diego, California, USA). A stringent filtering criterion was used to eliminate low-quality reads with the adapter sequences using software fastp (Chen et al. 2018), which is a data pre-processing tool used for quality control, trimming of adapters, filtering by quality, and read pruning to obtain high-quality clean reads. The sequence reads were then subjected to quality testing using the tools FastQC and MultiQC (Ewels et al. 2016), which allowed the analysis of parameters including base call quality distribution, % bases above Q20 and Q30, % GC, adapter sequence contamination, etc. The processed reads were assembled using assembler Megahit v1.1.3 (Li et al. 2015). The k-mer size range was set up from 21 to 141 with an increment of 28 using k-min, k-max, and k-step parameters. Notably, contigs shorter than 200 bp were removed from the assembly. Processed reads were mapped back to the assembled genome using assembler bowtie2 with default parameters (Langmead et al. 2012). The appropriate k-mer assembly was selected for SSRs mining on the basis of quality parameters. Subsequently, genome coverage was evaluated using the formula (https://genohub.com; https://www.illumina.com):

$$\text{Genome Coverage (GC)} = (\text{number of reads} * \text{read length})/\text{assembly size}$$

Finally, using the software Repeatmasker (https://www.repeatmasker.org/faq.html#faq3), the repeat sequences were masked.

### The SSR motif detection, primer designing and bioinformatics analysis

The program MIcroSAtellite (MISA) was used to detect and locate SSRs in the genomic DNA (Beier et al. 2017). Occurrence of repeats in the assembled genome revealed varied frequencies of di-, tri-, tetra-, penta-, and hexa-nucleotides. The program was able to identify and locate perfect microsatellites as well as the compound ones. Further, the primer pairs flanking in the region of SSRs were designed using the program PRIMER3 (https://bioinfo.ut.ee/primer3). The SSRs with at least 100 bp flanking sequence on both the ends were retained for primer sequencing.

Using NCBI BLASTX (https://blast.ncbi.nlm.nih.gov), the polymorphic SSR markers were compared to the non-redundant protein database to assess their putative functions. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database was crucial in comprehending the systematic functional data and applications of genes in biological systems (Kanehisa 2000). Further, the KEGG BRITE (KB) was utilized to comprehend the functional hierarchy, while the KEGG PATHWAY (KP) maps were used to illustrate molecular interaction and reaction. In addition, the KEGG Orthology (KO) numbers obtained from the KEGG server was used to summarize gene name, gene orthologs, functional definition of the orthologs, and the functional pathways through a stand-alone tool, i.e., Gene Annotation Easy Viewer (GAEV) (Huynh and Xu 2018). The Linux-based Krait tool was used to infer the SSRs' relative abundance (loci $Mb^{-1}$) and density (bp $Mb^{-1}$) for the detected SSRs (Du et al. 2018). Lastly, the functional enrichment analysis was done through g: Profiler (Raudvere et al. 2019).

### SSR validation, PCR amplification and data analysis

A subset of 60 primer pairs were synthesized for validation which consisted of 20 tri-, tetra-, and pentanucleotide repeats each based on the stringent parameters, such as product size 150–250 bp, % GC = 40–60% and temp 50–60 °C. The primers were tested for their amplification in a polymerase chain reaction (PCR) thermal cycler machine (Eppendorf Mastercycler Nexus). Screening and optimization of the annealing temperature of the primers were obtained by the gradient PCR ($T_m$ gradient range of ± 3 °C). The amplification was performed in a 15-µl PCR reaction mixture, containing 30 ng of template DNA, 7.5 µl of Taq mix, 0.1–1 µg of both forward and reverse primers and nuclease-free sterile water. The PCR conditions used were as follows: initial denaturation at 94 °C for 3 min, followed by 35 cycles of 94 °C for 30 s, primer-specific $T_m$ range for 30 s, annealing at 72 °C for 45 s; and a final extension at 72 °C for 3 min. The PCR products were electrophoresed and separated using 2% agarose gel buffered with 1 × TBE (Tris/borate/EDTA) along with 100 bp DNA ladder. The gel was stained with ethidium bromide (0.5 µg $ml^{-1}$) and visualized in the gel documentation system. After being subjected to PCR amplification in 15 random genotypes representing 3 different populations of *S. robusta*, positively amplified PCR products were resolved in 3% high-resolution agarose to check polymorphism (Make: Sigma-Aldrich). Finally, polymorphic primers were identified as those amplifying alleles of various sizes across the genotypes. The band profile produced by each SSR was scored manually by giving each band an estimated value for allele size, which was then modified in accordance with the repeat motifs of the primers using the allele binning tool TANDEM v1.07 (Matschiner and Salzburger 2009). Identification of scoring errors and excess of homozygotes at each locus to analyze the presence of null alleles was done through program Microchecker v2.2 (Van Oosterhout et al. 2004). Afterwards, the marker data were evaluated to characterize the primers and estimate the informativeness of SSR markers developed using allelic data, by calculating parameters, such as numbers of different alleles per locus ($N_a$), numbers of effective alleles ($N_e$), observed heterozygosity ($H_o$), expected heterozygosity ($H_e$), and the polymorphism information content (PIC), using program PowerMarker v3.25 (Liu and Muse 2005) and GenAlEx v6.5 (Peakall and Smouse 2012). Further, marker data were analyzed to depict the molecular variance (AMOVA) between different populations and within the genotypes of each population by calculating genetic differentiation ($F_{ST}$) and inbreeding coefficient ($F_{IS}$) through the program GenAlEx. The population structure of the 48 genotypes with 9 SSRs with admixture models and correlated band frequencies to determine number of sub populations (K) was assessed using STRUCTURE v2.2 (Pritchard et al. 2000; Evanno et al. 2005). The Jaccard similarity coefficient, the unweighted pair group method with arithmetic mean (UPGMA), and the SAHN clustering tool were used to determine the genetic similarity and generate a dendrogram between the genotypes by program NTSYS-pc v2.10 (Rohlf 1998).

## Results

### Illumina sequencing, assembly, SSR identification and primer design

A total of ~ 10 Gb data represented by 69.88 million raw reads were obtained from a low-depth high-throughput genome sequencing approach (Table 1). The quality of sequenced data generated was portrayed by the calculated

**Table 1** Summary statistics of shallow genome sequenced data

| Sl. No | Feature | Value |
|---|---|---|
| 1 | Total raw reads | 69,886,132 |
| 2 | Total number of bases (bp) | 10,482,919,800 |
| 3 | Clean reads proportion (%) | 93.2 |
| 4 | Total number of contigs | 197,489 |
| 5 | Assembly length (Mb) | 357.11 |
| 6 | Largest contig size (bp) | 140,425 |
| 7 | L50 (bp) | 16,369 |
| 8 | L75 (bp) | 49,235 |
| 9 | N50 (bp) | 5062 |
| 10 | N75 (bp) | 1536 |
| 11 | GC content (%) | 33.69 |
| 12 | Q20% | 98.615 |
| 13 | Q30% | 91.23 |
| 14 | Number of SSR identified through MISA | 57,702 |
| 15 | Number of SSR primers designed | 35,049 |
| 16 | Number of SSRs tested | 60 |
| 17 | Number of working SSR markers | 50 |
| 18 | Number of polymorphic SSR markers | 24 |

parameters, viz. GC content (33.69%), bases above Q20 (98.615%) and Q30 (91.23%), which were suitable for further processing. After quality filtration, cleaned paired reads were de novo assembled into 1,97,489 contigs (29 × coverage) with L50 value (16,369), L75 value (49,235), N50 value (5062), and N75 value (1536). Inclusively, based on k-mer, parameters, such as contigs size, read aligned percent, L50, L75, N50 and N75 were compared and the highest percentage (93.2%) of the aligned reads were selected for SSRs prediction. The raw sequencing data were deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) with accession number PRJNA639024.

The genome sequence data were utilized for identification of microsatellite repeats and development of SSR markers in *S. robusta* by scanning the contigs with the perl script MISA, which identified a total of 57,702 microsatellite repeats and 35,049 primer pairs were successfully designed from them. These SSRs were preceded as 'SRGMS', which stands for '*Shorea robusta* Genomic Microsatellite' marker. Additionally, repeats were analyzed for their frequency and distribution in the genome, where AT/AT repeats were most plentiful among di-nucleotides (60.60%), AAT/ATT among tri- (30.55%) and AAAT/ATTT among tetra-nucleotides (6.48%). Lastly, AAAAT/ATTTT and AAAAAT/ATTTTT occur in very low frequency among penta-(1.75%) and hexa-nucleotides (0.61%), respectively (Fig. 1a–d). Notably, relative abundance and density of each repeat type were also determined, with di-nucleotides having the highest relative abundance (75.00 loci $Mb^{-1}$) and density (1514.04 bp $Mb^{-1}$) followed by tri- (49.08 loci $Mb^{-1}$; 993.27 bp $Mb^{-1}$), tetra- (32.06 loci $Mb^{-1}$; 574.07 bp $Mb^{-1}$), penta (11.61 loci $Mb^{-1}$; 252.00 bp $Mb^{-1}$), and hexa-nucleotides (4.67 loci $Mb^{-1}$; 119.80 bp $Mb^{-1}$). In different genomic regions, di-nucleotides were the most abundant type, except in the CoDing Sequences (CDS) and exon regions, where tri-nucleotides were abundantly present (Fig. 1e).

For validation, a total of 60 primer pairs (20 each) from class tri-, tetra-, and penta-nucleotides were synthesized. A total of 24 out of 50 (48%) successfully amplified SSRs revealed a polymorphic banding pattern (Table 2; Fig. 2a, b; and Supplementary Fig. 1a–v), whereas rest ten were not amplified. Further, errors in the fragment separation and allele scoring were eliminated by binning to detect the null alleles. It revealed that out of 24 primer pairs, 15 were observed with an excess of homozygotes. Though assessing the values of parameters, such as *N*a, *H*o, *H*e, etc., for the full dataset and dataset excluding the null alleles, revealed the significant disparity among the estimates. The full datasets and gel images containing all 24 polymorphic primers have been provided in the Supplementary Table 2 and Supplementary Fig. 1. Consequently, for the authentication of the primers, nine SSR loci were evaluated across the 48 genotypes representing three *S. robusta* populations (Table 3).

## Functional annotation

The putative functions of all the polymorphic SSRs were obtained using a sequence similarity search against the non-redundant protein database NCBI's BLASTX in order to highlight its functional relevance (Table 2). Out of 1,97,489 contigs, 15,914 contigs were successfully mapped into 431 pathways, involving large number of contigs in pathways of neurodegeneration—multiple diseases (260 contigs); amyotrophic lateral sclerosis (231 contigs); Alzheimer disease (219 contigs); prion disease (162 contigs); *Salmonella* infection (145 contigs); thermogenesis (144 contigs); human papillomavirus infection (144 contigs); Coronavirus disease—COVID-19 (135 contigs); endocytosis (134 contigs); chemical carcinogenesis—reactive oxygen species (132 contigs), etc. (Supplementary Table 3).

Functional hierarchies were obtained through KB and characterized into three categories of protein families, namely (i) metabolism, (ii) genetic information processing, and (iii) signaling and cellular processes (Fig. 3). The KO numbers obtained through KEGG were annotated using GAEV, which was further characterized by g: profiler into metabolic component, cellular component, and biological component with their GO ID and *p*-value. The highest number of GO terms were involved in the biological process, i.e., 382 followed by the cellular component (91) and molecular function (61) as revealed by Manhattan-like plot. The table below the plot, which includes the data source

GO ID, term name, and *p*-value, also includes the identification that the plot highlights by hovering the circle. For example, the plot illustrates circle No. 1: the enrichment of the term GO:0,009,987 (cellular process) followed by circle No. 2: the term GO:0,008,152 (metabolic process), and so on (Fig. 4). Detailed results of 100 terms of the biological process, cellular component, and molecular function are illustrated in Supplementary Fig. 2.

### Polymorphic potential of novel marker loci and their efficacy in population genetic analysis

Polymorphic primers were utilized for the estimation of key diversity measures in 48 genotypes belonging to three distantly located populations of *S. robusta* in Uttarakhand (Table 3). In total, 22 alleles were generated with nine SSRs across the genotypes with an average of 2.44 alleles per locus. The PIC of each SSR primer pair ranged from 0.020 to 0.554, with a mean of $0.252 \pm 0.06$. The mean range of $H_o$ for the primers across all the populations was recorded between 0.021 and 1.000 with a mean of $0.324 \pm 0.10$, while $H_e$ ranged from 0.020 to 0.596 with a mean of $0.277 \pm 0.07$. Further, AMOVA revealed that most of the genetic variation (97%) was confined within a population; thus, a very low genetic differentiation ($F_{ST} = 0.029$) was observed among the populations. It is also supported by the high mean value of gene flow across the primers ($N_m = 17.90$). The range of inbreeding coefficient ($F_{IS}$) observed among the sampled populations was $-0.679$ to 0.206 with a mean value of $-0.109 \pm 0.08$.

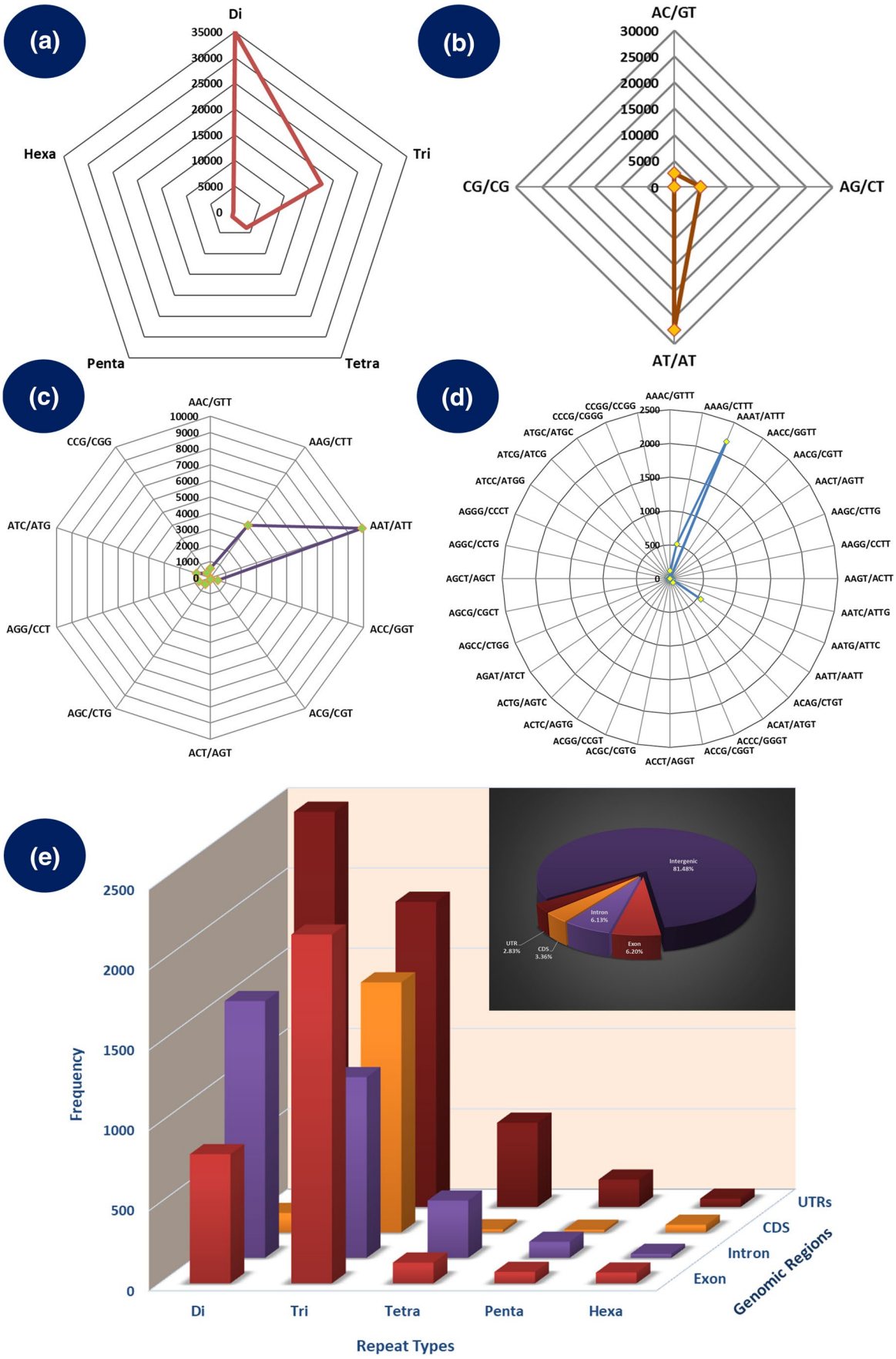
Moreover, structural analysis suggests an optimal K value of 2 [Supplementary Fig. 3a(i-iii)], which is far too low to predict any output. As a result, clearcut structure in the investigated populations of *S. robusta* was not apparent. A PCoA plot (Supplementary Fig. 3b) and UPGMA dendrogram were produced as a result of the intra-specific genetic diversity analysis using SSR markers (Fig. 5). These results demonstrated that 48 genotypes had been clearly split with a similarity coefficient of 0.79 into two distinct groups (Gp) of *S. robusta*, with GpI and GpII consisting of 44 and 4 genotypes, respectively. Notably, the former was separated with a similarity coefficient of 0.800 into two subgroups (SbGp), namely SbGpIa (43 genotypes) and SbGpIb (1 genotype), while SbGpIIa (2 genotypes) and SbGpIIb (2 genotypes) were separated from the subsequent GpII with a similarity coefficient of 0.885.

## Discussion

Outcrossing species generally have a great potential for gene flow, which assists them to maintain high levels of genetic diversity within populations (Hamrick 1983; Hamrick and Godt 1989; Tam et al. 2014). Given the persistence of genetic diversity, tree species typically acclimatize to long-term environmental change (Hedrick 2004). Compared with the anonymous markers, SSR markers yield more precise estimates of genetic diversity (Feng et al. 2016). Recently developed next-generation sequencing (NGS) platforms, such as Roche's 454 GS FLX, Illumina's Genome Analyzer (GA) and ABI's SOLiD, offer opportunities for high-throughput, cost-effective genome sequencing, and rapid marker development (Li et al. 2018). Compared with the traditional library-based and in silico methods, DNA-Seq via. Illumina is quicker with a lower cost and less dependency on existing genetic resource of target plant species for sequence-based marker development. This would also bring advancement in molecular markers-based studies on those plants which lack a genomic database (Bosamia et al. 2015). For instance, high-throughput transcriptome sequencing has been successfully employed for identifying SSRs in trees, such as *Hevea brasiliensis*, *Carapa guianensis*, *Eperua falcata*, and *Symphonia globulifera* (Brousseau et al. 2014; Sae-Lim et al. 2019); and in angiosperms, such as rose, peony, and olive (Gao et al. 2013; Yan et al. 2015; Mariotti et al. 2016). Further, numerous SSR markers have also been developed for *Shorea curtisii* (Ujino et al. 1998; Obayashi et al. 2002; Ho et al. 2006) and *Shorea leprosula* (Lee et al. 2000; Lee et al. 2004; Cao et al. 2006), but only limited microsatellite markers were detected for *S. robusta* (Pandey and Geburek 2009, 2010 and 2011). Since the species lacking genomic sequence data, which are essentially required for SSR mining. Thus, SSRs could play an important role in genetic diversity analysis, gene flow pattern, DNA fingerprinting, marker assisted selection (MAS), etc.

The present study reports discovery of 35,049 novel SSRs (24 out of 60 were validated and found polymorphic) in *S. robusta* in which ~ 10 Gb raw sequence data were generated and assembled into 1,97,489 contigs representing genome size of 357.11 Mb with a coverage of 29×. Recent past revealed that the approach has been used to develop microsatellite markers in various species viz. *S. leprosula* (Ng et al. 2009; Ng et al. 2021), *Grevillea thelemanniana* (Hevroy et al. 2013), *Macadamia integrifolia* (Nock et al. 2016), *Populus pruinosa* (Yang et al. 2017), *G. juniperina* (Damerval et al. 2019), *Exbucklandia tonkinensis* (Huang et al. 2019), etc., signifying the potential of this technology for the identification and development of novel SSRs in *S. robusta*, devoid of genome sequence information.

Genome annotation to get genome-wide information is quite effortless now since NGS has come into existence. Notwithstanding, the annotation related tasks are challenging and rely upon the accessible tools and procedures, and further to decipher the information contain in the sequenced genome. The putative functions of 24 polymorphic SSRs anticipated that the top-hit species were *Theobroma cacao*,


(a)
Di
Tri
Tetra
Penta
Hexa
35000
30000
25000
20000
15000
10000
5000
0
(b)
AC/GT
AG/CT
AT/AT
CG/CG
30000
25000
20000
15000
10000
5000
0
(c)
AAC/GTT
AAG/CTT
AAT/ATT
ACC/GGT
ACG/CGT
ACT/AGT
AGC/CTG
AGG/CCT
ATC/ATG
CCG/CGG
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
(d)
AAAC/GTTT
AAAG/CTTT
AAAT/ATTT
AACC/GGTT
AACG/CGTT
AACT/AGTT
AAGC/CTTG
AAGG/CCTT
AAGT/ACTT
AATC/ATTG
AATG/ATTC
AATT/AATT
ACAG/CTGT
ACAT/ATGT
ACCC/GGGT
ACCG/CGGT
ACCT/AGGT
ACGC/CGTG
ACGG/CCGT
ACTC/AGTG
ACTG/AGTC
AGAT/ATCT
AGCC/CTGG
AGCG/CGCT
AGCT/AGCT
AGGC/CCTG
AGGG/CCCT
ATCC/ATGG
ATCG/ATCG
ATGC/ATGC
CCCG/CGGG
CCGG/CCGG
2500
2000
1500
1000
500
0
(e)
Frequency
2500
2000
1500
1000
500
0
Di
Tri
Tetra
Penta
Hexa
Repeat Types
UTRs
CDS
Intron
Exon
Genomic Regions
Intergenic
81.48%
Intron
6.13%
Exon
6.20%
UTR
2.83%
CDS
3.36%

◂**Fig. 1** The frequency distribution of simple sequence repeat (SSR) types: (**a**) Radar indicates all types of SSR motifs; and (**b**–**d**) most predominant repeat motifs, i.e., di-, tri-, and tetra-nucleotides; and (**e**) the distribution of identified microsatellite motifs in different genomic regions of *S. robusta*

*Erythranthe guttata*, *Glycine max*, *H. brasiliensis*, *Ricinus communis*, *Citrus unshiu*, *Gossypium raimondii*, *Durio zibethinus*, *Gossypium arboreum*, *Vernicia fordii*, *Corchorus capsularis*, *Brassica napus*, *Cucurbita maxima*, *Gossypium hirsutum*, and *Cephalotus follicularis*. The KP aims to organize and computerize all the current knowledge of molecular and genetic pathways from experimental viewpoint, which implies the understanding of the molecular interaction and reaction networks. Here, KEGG database were used to perform functional annotation, delivering specifics of organismal genes and pathways besides establishing an association between them. Likewise, systematic identification of Expressed Sequence Tags (ESTs)-based SSRs (EST-SSRs) were carried out in *Pinus taeda* in California (Liewlaksaneeyanawin et al. 2004); *S. leprosula* in Indonesia (Ohtani et al. 2012); *V. fordii* and *Vernicia montana* in southwestern China and northern Laos (Xu et al. 2012), *Pinus dabeshanensis* in China (Xiang et al. 2015), *H. brasiliensis* in (Danzhou) China (Hou et al. 2017), and *Dalbergia odorifera* in China (Liu et al. 2019); whereas, putative functional SNP markers were detected for *Shorea parvifolia* in (Kuching Sarawak) Malaysia (Seng et al. 2011) and *Juniperus phoenicea* subsp. *turbinata* in Spain (Garcia et al. 2018).

KB has united a variety of interactions, such as those between genes and proteins, elements and reactions, medications and illnesses, and organisms and cells (Kanehisa et al. 2019). Parallel studies were reported in *Vatica mangachapoi* (Tang et al. 2022), *Hopea hainanensis* (Huang et al. 2022), *Neesia altissima* (Pratiwi et al. 2022), *C. capsularis* (Satya et al. 2017), *Hibiscus hamabo siebold & zuccarini* (Wang et al. 2021), *Abelmoschus esculentus* (Nieuwenhuis et al. 2021), *Helicoverpa armigera* (de la Paz Celorio-Mancera et al. 2011), *Gasterophilus nasalis* (Zhang et al. 2021), and *Operculina turpethum* (Biswal et al. 2021). Additionally, GAEV was used to annotate the KO (Iacobas et al. 2019; Emami-Khoyi et al. 2020; Nand et al. 2020; Shah et al. 2021). The biological process (382), with the highest level of involvement in the functional enrichment analysis results is highlighted here by GO ID and p-value. Lately, this kind of characterization and annotation of genes were used to predict common functions of 12,886 whole-genome duplication (WGD) in *S. leprosula* (Ng et al. 2021), examination of differentially expressed genes (Yamasaki et al. 2017), validation of immune genes (Karthikeyan et al. 2021), identification of novel prognostic biomarker (Xu et al. 2020), analyses of Integrated Gene Expression Profiling Data (IGEPA) (You et al. 2020), identification of the blood-based signatures molecules and drug targets of patients with COVID-19 (Hasan et al. 2022), and annotation of protein–protein interactions (Ieremie et al. 2022).

In the present study, a total of 35,049 SSRs were recognized, where the highest being di-nucleotides (34,969) also showed maximum relative abundance and density (75.00 loci $Mb^{-1}$; 1514.04 bp $Mb^{-1}$), followed by tri- (17,630), tetra- (3741), penta- (1011), and hexa-nucleotides (351). The SSRs repeat analysis revealed that the most prominent and abundant frequency of motifs was observed for AT/AT and AAT/ATT, similar to the study conducted on arid-zone *S. oleoides* (Bhandari et al. 2020). In other species, such as *S. curtisii,* simple CT and compound repeats of CT, CA, AT, and CTCA were observed (Ujino et al. 1998); whereas in case of *Drepanostachyum falcatum*, AG/CT and CCG/CGG were observed in maximum number (Meena et al. 2021), etc.

The characterization of genetic diversity patterns at intra- and inter-population levels is a fundamental requirement for the establishment of forest genetic resources conservation and tree improvement programmes (Stojnic et al. 2019). However, molecular tools play an important role in the efficient management and utilization of genetic assets. Thus, the usage and implication of SSR-based molecular markers increases in revealing the genetic diversity among the populations of a particular species. Notably, standardization of the isolation protocol of DNA from the samples, the quality of the markers, and the accuracy of the genotyping data, actually determines the effectiveness and success of SSRs (Liu et al. 2017). In this research, a total of 50 out of 60 primer pairs yielded 100% clear bands across three different populations of *S. robusta*. The amplification rate (83.33%) was significantly higher in comparison to *S. curtisii* (23.07%) (Ujino et al. 1998) and *Liquidambar formosana* (72%) (Chen et al. 2020) and *Parashorea malaanonan* (82%) (Abasolo et al. 2009), due to originality of the species-specific marker.

Additionally, amidst 48 accessions of *S. robusta*, 24 out of 60 markers exhibited polymorphism and showed moderate levels of polymorphism. Here, a total of 22 alleles with an average of 2.44 alleles per locus were generated that is quite lower to one of the members, i.e., *Hopea hainanensis* of a family Dipterocarpaceae, which revealed a total of 229 alleles with an average of 11.45 alleles per locus while using 20 microsatellite loci (Wang et al. 2020). In another study, 41 alleles ranging from 2.8 to 4.2 allele per locus with six microsatellite loci in *H. brasiliensis* were generated (Yu et al. 2011). In *Diospyros kaki* Thunb. (Family: Ebenaceae), the number of alleles detected ranging from 2 to 17 with an average of 8.54 with 13 SSRs (Wang et al. 2021). Further, in the neighboring genus, more than 242 samples across eight populations of both *Dipterocarpus costatus* and *Dipterocarpus alatus* were genotyped through 9 loci, where an overall 26 and 28 alleles were detected with

**Table 2** Characteristics and putative functions of 24 polymorphic SSRs (SRGMS) with E-value of *S. robusta*

| Sl. No | Locus name | Primer sequences | Product size (bp) | Repeat motifs | $T_m$ (°C) | Putative functions | E-value |
|---|---|---|---|---|---|---|---|
| | SRGMS-211 | AGACTGCCTGGAAAATGGGT<br>CCCTCCTTGTTGGTTCTGCT | 175 | (AT)10 | 55.1 | Not found | - |
| | SRGMS-254 | ACAAGAGGGAAAGATGAA GACA<br>TCTCCTACCTCTCCCCTCCT | 194 | (GAA)5 | 56.2 | Not found | - |
| | SRGMS-256 | CCATGAAGGTACACAAGATGA ACA<br>CGGGGTCCACCAACAGTAAA | 178 | (AATA)7 | 61.1 | *Theobroma cacao*<br>Alpha/beta-Hydrolases superfamily protein | 8.95e-153 |
| | SRGMS-310 | ATAACCCATGTCCTGCCCAA<br>GAATGGCCATGATTTGCCCC | 191 | (AT)6 | 61.1 | *Theobroma cacao*<br>MEI2-like 4, putative isoform 1 | 1.92e-13 |
| | SRGMS-431 | AGCACTTTTGAGGTCTAG TTCGT<br>CGTATCCCTCTCTTGGTGCC | 168 | (TA)10 | 56.0 | *Erythranthe guttata*<br>PREDICTED: uncharacterized protein LOC105961778 | 1.22e-124 |
| | SRGMS-1237 | AGAGCTTAGAAGGTTGAT GGCT<br>AGAAAAGGGGTTGGACGGAC | 166 | (AAG)5 | 57.7 | *Glycine max*<br>PREDICTED: zinc finger BED domain-containing protein RICESLEEPER 2-like | 2.22e-96 |
| | SRGMS-1370 | TCCCGTGTGGAAGTTAAGTAT AGT<br>AAGCTGGCTGCCCCATATAC | 208 | (TTC)6 | 57.7 | *Theobroma cacao*<br>PREDICTED: uncharacterized protein LOC18599617 isoform X4 | 0.001 |
| | SRGMS-1696 | CAGACAGGCTAGCAACTCGT<br>CCAGATCGAGTTCTCCACCG | 180 | (GGA)5 | 55.6 | Not Found | - |
| | SRGMS-1811 | TCGTATAACATGCGAGTGGAGT<br>TGGGTGGGAAGGGATGTTTG | 158 | (AT)7 | 57.3 | *Hevea brasiliensis*<br>putative leucine-rich repeat receptor-like serine/threonine-protein kinase At2g24130 | 0 |
| | SRGMS-1908 | TCAGAAAGCTTCCCCTCGAC<br>AGGTGGTTCAAGTCATGCCT | 166 | (TC)8 | 57.3 | *Ricinus communis*<br>PREDICTED: peroxidase 5 | 1.83e-15 |
| | SRGMS-2126 | TCAGTACACATGATGACCTTA GCT<br>AAGGTGCACACTCCAATCCT | 152 | (AT)15 | 58.0 | Not found | - |
| | SRGMS-2403 | ATTTTGCCTGCATGCAGAGG<br>TGTGTGCACTTACCCATCCA | 181 | (GA)20 | 57.3 | *Citrus unshiu*<br>hypothetical protein CUMW_112070 | 1.65e-58 |
| | SRGMS-3859 | AGATTCCTTTCGCGTCTCCC<br>CGGATCCAAGAACGGAGAGG | 189 | (ACC)5 | 61.5 | *Gossypium raimondii*<br>PREDICTED: respiratory burst oxidase homolog protein A-like isoform X2 | 2.18e-55 |
| | SRGMS-4747 | ACCCCTGCTTAAGCTTGACA<br>CGTCATGCAAATCCCTTCTCC | 176 | (TA)15 | 57.3 | *Durio zibethinus*<br>protein XRI1-like isoform X2 | 1.17e-18 |
| | SRGMS-5550 | CACAGCCTCCTCGGTAACAA<br>TTCCTGGTAAGGTTGCTGCT | 166 | (TTTA)6 | 61.1 | *Theobroma cacao*<br>PREDICTED: protein POLLENLESS 3-LIKE 2 | 0 |
| | SRGMS-5677 | TCCGTTGGAAAGGAAAGA GACA<br>AGGTTCCAGGTAGCTGCAAA | 151 | (AT)11 | 58.6 | *Durio zibethinus*<br>S-adenosylmethionine decarboxylase proenzyme-like | 1.79e-139 |
| | SRGMS-5744 | AGCATCAGTGTTCGTGTAACG<br>TTGAACTGGGGGCTCTGAAC | 170 | (ATAA)5 | 55.5 | *Gossypium arboreum*<br>PREDICTED: BTB/POZ domain-containing protein At5g48130 | 0 |
| | SRGMS-6128 | CCTGATGGAGACTTCTTAGGCA<br>GCTGGGCTGAAATATCCCGA | 152 | (AAAT)6 | 55.0 | *Vernicia fordii*<br>alpha-galactosidase | 1.08e-25 |
| | SRGMS-6690 | ACGTTGGTAGCTTTGAGGTCT<br>CTCAGAGTGAGGGACATCGC | 201 | (GAG)7 | 56.2 | *Corchorus capsularis*<br>hypothetical protein CCACVL1_18752 | 5.98e-12 |

**Table 2** (continued)

| Sl. No | Locus name | Primer sequences | Product size (bp) | Repeat motifs | $T_m$ (°C) | Putative functions | E-value |
|---|---|---|---|---|---|---|---|
| | SRGMS-9401 | TGTGATCTGCTGCCTTATGGA<br>CTTGTCCTCACACAACCCGA | 157 | (AT)8 | 60.1 | *Brassica napus* uncharacterized protein LOC106425095 | 1.60e-11 |
| | SRGMS-11234 | TGGACGAAGACATGAAGA TGGA<br>CCCCACAAACACCGAGAGAA | 187 | (GAA)7 | 60.1 | *Citrus unshiu* hypothetical protein CUMW_159590 | 2.16e-25 |
| | SRGMS-11434 | CCTTCTCTGTTCACGCTTTCC<br>ATGTGGCTATGGCGTTAGGG | 198 | (CCG)6 | 60.2 | *Cucurbita maxima* cellulose synthase A catalytic subunit 2 [UDP-forming]-like | 0 |
| | SRGMS-11787 | ACAGACAGGAACATTCTC CAGA<br>TTGGTTGGACCTGCATCTGA | 164 | (AT)12 | 60.4 | *Gossypium hirsutum* PREDICTED: F-box/kelch-repeat protein At1g23390-like | 2.49e-124 |
| | SRGMS-23562 | GAAACGAGCATCTGTCACGG<br>CTCCAGATAACCACCCTCGC | 162 | (AAGA)5 | 62.1 | *Cephalotus follicularis* DUF632 domain-containing protein/DUF630 domain-containing protein | 1.82e-15 |

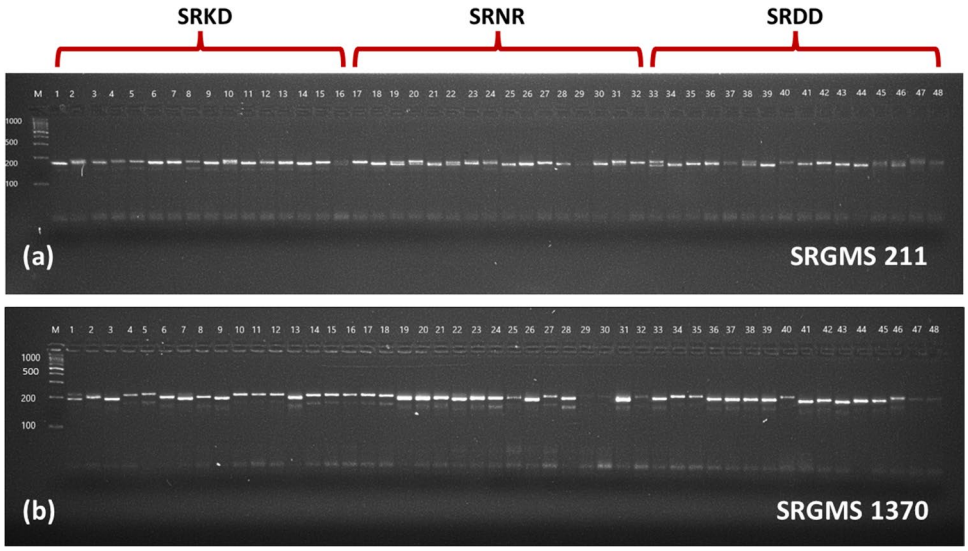



**Fig. 2** Representative sample* through SSRs showing polymorphic banding pattern in *S. robusta*: (**a**) SRGMS 211, and (**b**) SRGMS 1370. *Where, M: 100 bp DNA ladder; 1–48 representing 16 genotypes each from 3 populations

an average of 2.9 and 3.1 alleles per primer, respectively (Vu et al. 2019). All these studies confirmed that more the number of microsatellites used in a genotyping-based study, the more will be the number of polymorphic bands. The current study revealed PIC values ranged from 0.020 to 0.554 with a mean value of 0.252 for *S. robusta* (Table 3), which presumed to be low when compared to tropical and subtropical species, such as *Pinus cineraria* (PIC = 0.49 to 0.78) (Rai et al. 2017), *D. costatus* (PIC = 0.317) (Vu et al. 2019), *S. persica* (PIC = 0.630) (Monfared et al. 2018), and *D. kaki* (PIC = 0.7306) (Wang et al. 2021) but higher than *D. alatus* (PIC = 0.216) (Vu et al. 2019).

The population genetics and diversity studies are mainly based on estimating the alleles and genotype frequencies, and the changes caused by evolutionary forces, gene flow, mutations, genetic drift, and natural selection (Eriksson et al. 2001). It is necessary to assess the genetic variation levels within and among populations for understanding of the species evolutionary biology and tree improvement potentiality (Escuderoa et al. 2003). The key measures of the genetic diversity are observed ($H_o$) and expected heterozygosity ($H_e$) (Sherif and Alemayehu 2018), where $H_e$ is considered as a most suitable measure for characterizing marker loci among the different genotypes of a species (Monfared et al. 2018; Xue et al. 2018). To this date, works on genetic analysis in *S. robusta* were conducted using isozymes and ISSR markers in Nepal and India, respectively (Suoheimo et al. 1999; Surabhi et al. 2017). Moreover, few microsatellite studies have also been conducted on this species (Pandey and Geburek 2009, 2010, 2011). Our estimates of heterozygosity and number of alleles ($H_o$ = 0.021–1.000, $H_e$ = 0.020–0.596, and $N_a$ = 2.44) are comparable with the range found in *S. robusta* ($H_o$ = 0.49–0.77; $H_e$ = 0.52 to 0.89,

**Table 3** Genetic polymorphism of 9 SSR loci evaluated in three *S. robusta* populations

| Sl. no | Locus | Pop 1: SRKD (n = 16) | | | Pop 2: SRNR (n = 16) | | | Pop 3: SRDD (n = 16) | | | PIC | $F_{IS}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $N_a$ | $H_o$ | $H_e$ | $N_a$ | $H_o$ | $H_e$ | $N_a$ | $H_o$ | $H_e$ | | |
| 1 | SRGMS-1237 | 3 | 1.000 | 0.576 | 3 | 1.000 | 0.617 | 3 | 1.000 | 0.594 | 0.554 | -0.679 |
| 2 | SRGMS-3859 | 2 | 0.063 | 0.061 | 2 | 0.125 | 0.117 | 2 | 0.125 | 0.117 | 0.094 | -0.060 |
| 3 | SRGMS-5677 | 3 | 0.625 | 0.596 | 3 | 0.625 | 0.580 | 2 | 0.375 | 0.375 | 0.485 | -0.048 |
| 4 | SRGMS-5744 | 2 | 0.313 | 0.264 | 2 | 0.250 | 0.219 | 2 | 0.500 | 0.375 | 0.249 | -0.239 |
| 5 | SRGMS-6128 | 1 | 0.000 | 0.000 | 1 | 0.000 | 0.000 | 2 | 0.063 | 0.061 | 0.020 | -0.032 |
| 6 | SRGMS-9401 | 2 | 0.188 | 0.170 | 2 | 0.313 | 0.264 | 2 | 0.125 | 0.117 | 0.169 | -0.135 |
| 7 | SRGMS-11234 | 2 | 0.313 | 0.404 | 3 | 0.313 | 0.529 | 2 | 0.438 | 0.404 | 0.366 | 0.206 |
| 8 | SRGMS-11434 | 2 | 0.063 | 0.061 | 2 | 0.125 | 0.117 | 2 | 0.063 | 0.061 | 0.077 | -0.049 |
| 9 | SRGMS-23562 | 3 | 0.500 | 0.461 | 2 | 0.125 | 0.117 | 2 | 0.125 | 0.219 | 0.254 | 0.059 |
| | Range | 1–3 | 0.000–1.000 | 0.000-0.596 | 1–3 | 0.000–1.000 | 0.000–0.617 | 2–3 | 0.063–1.000 | 0.061–0.594 | 0.020–0.554 | -0.679–0.206 |
| | Mean ± SEm | 2.22 ± 0.22 | 0.340 ± 0.11 | 0.288 ± 0.07 | 2.22 ± 0.22 | 0.319 ± 0.10 | 0.285 ± 0.08 | 2.11 ± 0.11 | 0.313 ± 0.10 | 0.258 ± 0.06 | 0.252 ± 0.06 | -0.109 ± 0.08 |

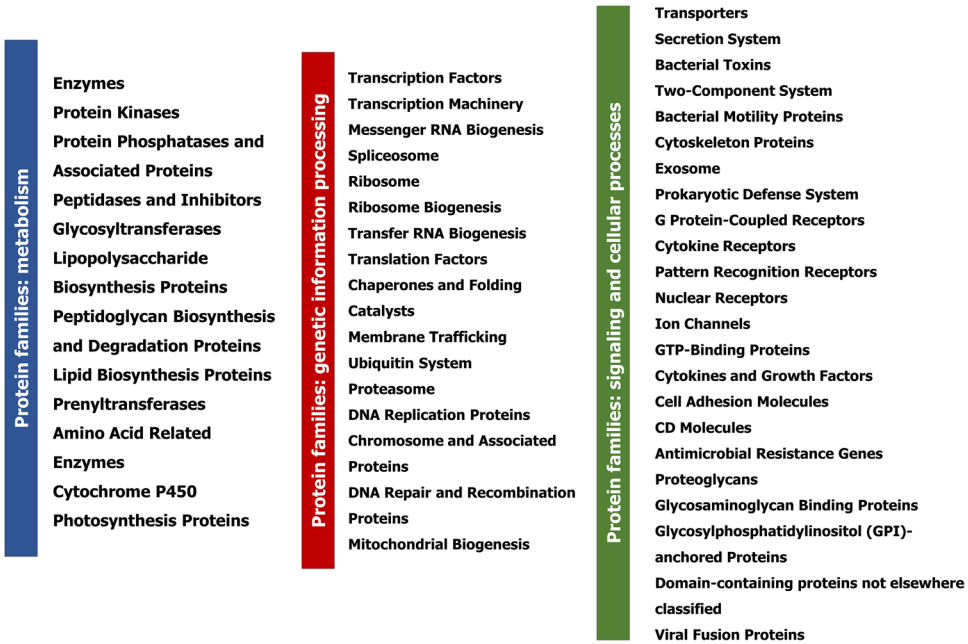
Where *n*, number of individuals collected for each population; $N_a$, number of different alleles per locus; $H_o$, observed heterozygosity; $H_e$, expected heterozygosity; PIC, polymorphism information content; and $F_{IS}$, inbreeding coefficient; and SRKD, *S. robusta* Kotdwar; SRNR, *S. robusta* Narendra Nagar; SRDD, *S. robusta* Dehradun; and SEm, Standard Error of Mean

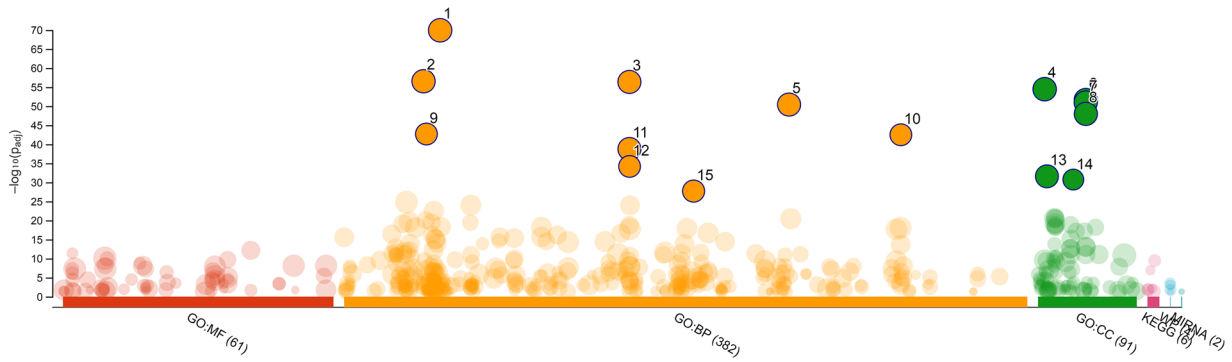
and $N_a = 11.80$) in Nepal (Pandey and Geburek 2009) and *Shorea guiso* ($H_o = 0.20–0.90$; $H_e = 0.66–0.87$, and $N_a = 15.67$) in the Philippines (Tinio et al. 2014). These measures were also equated with the members of the same family, such as *S. curtisii* ($H_e = 0.64$, $N_a = 7.9$) (Ujino et al. 1998), *Neobalanocarpus heimii* ($H_o = 0.67$, $H_e = 0.78$, and $N_a = 8.8$) (Konuma et al. 2000), *Dryobalanops aromatica* ($H_o = 0.49$, $H_e = 0.71$, and $N_a = 5.1$) (Lim et al. 2002), and *S. leprosula* ($H_o = 0.64$, $H_e = 0.70$, and $N_a = 11.4$) (Ng et al. 2004). The estimation of diversity measures ($H_o = 0–0.755$; $H_e = 0.255–0.757$) was successfully demonstrated with microsatellite markers in the neighboring genus *H. hainanensis* of China (Wang et al. 2020); *D. alatus* (gene diversity (H) = 0.223) and *D. costatus* (gene diversity (H) = 0.152) in Vietnam (Vu et al. 2019), which are closely linked with the estimated measures determined in the current study on *S. robusta*.

It has been suggested that a value lying below 0.05 indicates little genetic differentiation (Wright 1978; De Vicente et al. 2004) which implies very low genetic differentiation ($F_{ST} = 0.029$) in *S. robusta* populations. Conferring a negative value of inbreeding coefficient ($F_{IS} = -0.109$) and low $F_{ST}$, structuring and inbreeding depression was virtually not observed. Lack of significant pair-wise $F_{ST}$ indicates a pronounced gene flow among populations, due to no prominent physical barriers like mountain ridges (Pandey and Geburek 2009) during the sampling. A high rate of gene flow homogenizes the genetic differences among populations, even in the presence of intense selection (Zucchi et al. 2005). Besides, this area is characterized by continuous forests and gregarious distribution of *S. robusta* assisted by cross-pollination that supports high gene flow. Similarly, low $F_{ST}$ (0.024) and low $F_{IS}$ (0.09) indicated lesser genetic divergence despite 15 continuous and disjunct populations of this species in Nepal (Pandey and Geburek 2009). The outcomes of genetic diversity study were also supported by the structure analysis, which showed a low K value (K = 2, default generated in case of low structuring; Supplementary Fig. 3(a-iii)), as populations are not clearly defined by any single cluster. This indicates that a single or a maximum of two ancestral gene pools may result in significant genetic admixing throughout the geographical areas. Yet again, PCoA and UPGMA cluster analysis revealed similar grouping, which tends to bolster the low value of $F_{ST}$.

## Conclusions

The study demonstrates that SSR marker technique is a powerful tool for evaluating genetic diversity and relationships among the natural populations of *S. robusta*. Findings also revealed the utility of the microsatellite markers for assessing the genetic diversity estimates of this species. The novel set of genomic SSR markers in *S. robusta* were reported for




| ID | Source | Term ID | Term Name | $p_{adj}$ (query_1) |
|---|---|---|---|---|
| 1 | GO:BP | GO:0009987 | cellular process | $1.189\times10^{-70}$ |
| 2 | GO:BP | GO:0008152 | metabolic process | $2.901\times10^{-57}$ |
| 3 | GO:BP | GO:0044237 | cellular metabolic process | $4.324\times10^{-57}$ |
| 4 | GO:CC | GO:0005622 | intracellular | $3.498\times10^{-55}$ |
| 5 | GO:BP | GO:0071704 | organic substance metabolic process | $3.786\times10^{-51}$ |
| 6 | GO:CC | GO:0043229 | intracellular organelle | $2.466\times10^{-52}$ |
| 7 | GO:CC | GO:0043226 | organelle | $1.078\times10^{-51}$ |
| 8 | GO:CC | GO:0043231 | intracellular membrane-bounded organelle | $1.048\times10^{-48}$ |
| 9 | GO:BP | GO:0009058 | biosynthetic process | $2.032\times10^{-43}$ |
| 10 | GO:BP | GO:1901576 | organic substance biosynthetic process | $3.233\times10^{-43}$ |
| 11 | GO:BP | GO:0044238 | primary metabolic process | $1.725\times10^{-39}$ |
| 12 | GO:BP | GO:0044249 | cellular biosynthetic process | $6.590\times10^{-35}$ |
| 13 | GO:CC | GO:0005737 | cytoplasm | $2.613\times10^{-32}$ |
| 14 | GO:CC | GO:0032991 | protein-containing complex | $1.863\times10^{-31}$ |
| 15 | GO:BP | GO:0050896 | response to stimulus | $1.984\times10^{-28}$ |

| | |
|---|---|
| **version** | e99_eg46_p14_f929183 |
| **date** | 7/21/2020, 10:51:45 AM |
| **organism** | athaliana |

g:Profiler

**Fig. 3** Functional hierarchies obtained through KEGG BRITE

**Fig. 4** The hierarchical clustering of the genes assigned to a particular process in GO

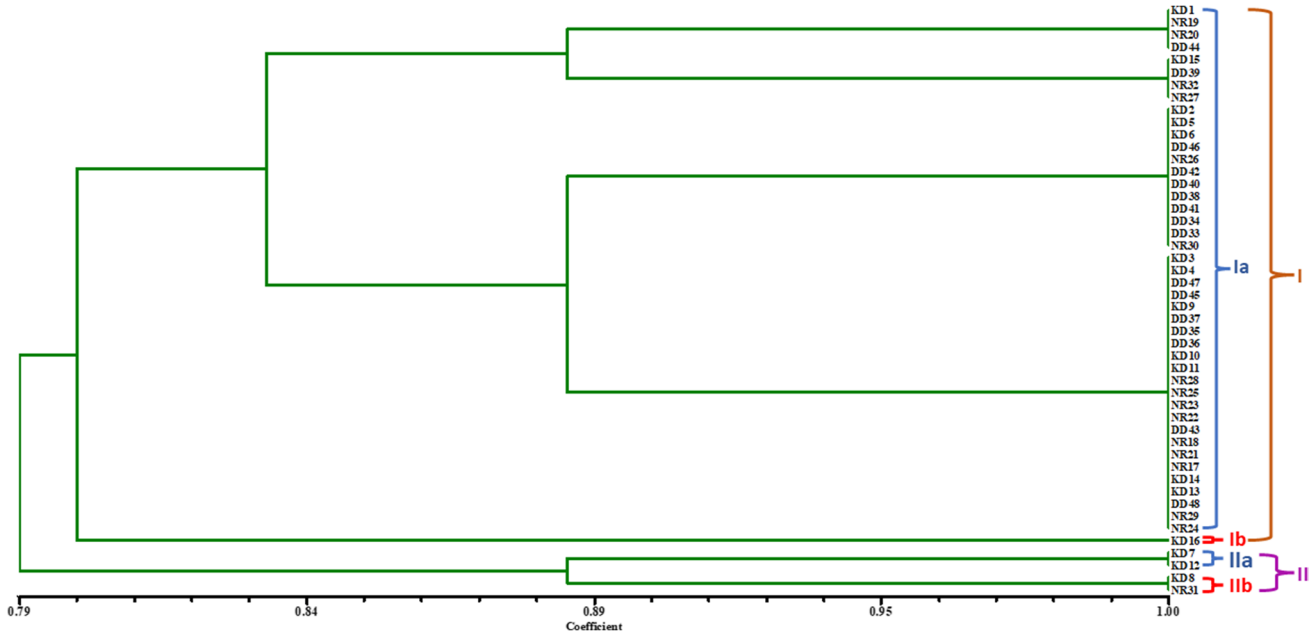



**Fig. 5** The UPGMA dendrogram unbiased measures of genetic distance among 48 genotypes representing 3 populations of *S. robusta*

the first time may serve as a useful tool for conservation and management of Dipterocarpaceae. For conservation implications, future molecular studies should cover the entire distribution range of the species, where the SSRs developed here might play a profound role in ascertaining biodiversity hotspots.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10142-023-00975-8.

**Acknowledgements** The authors are thankful to the Director, FRI for providing the research facilities and highly obliged to the state forest department, Government of Uttarakhand, for permission and support during the field surveys.

**Author contribution** GM: Wrote original draft, data collection and analysis, reviewing research papers, inference, and primary draft writing; RKM: Assist in population genetics data analysis, bioinformatics-based data analysis with GM & MSB and draft editing; MSB: Project administration and supervision, conceptualization, draft reviewing and editing; and SP, RK and HSG: Draft editing and add-on basic approaches. All the authors critically revised the final draft.

**Funding** This study was supported by the National Program for Conservation and Development of Forest Genetic Resources (NPCDFGR), CAMPA funded under the project grant No. 9–136/DGTI/NFGR-2019; dated 06th January, 2020.

**Data availability** All data files have been uploaded and clearly written in the manuscript.

**Code availability** All have been mentioned in the manuscript text.

## Declarations

**Competing interests** The authors declare no competing interests.

**Ethics approval** Not applicable.

**Consent to participate** Not applicable.

**Consent for publication** Yes, from all authors.

**Conflict of interest** The authors declare no competing interests.

## References

Abasolo MA, Fernando ES, Borromeo TH, Hautea DM (2009) Cross-species amplification of Shorea microsatellite DNA markers in *Parashorea malaanonan* (Dipterocarpaceae). Philippine J Sci 138(1):23–28

Adhikari B, Kapkoti B, Lodhiyal N, Lodhiyal LS (2017) Structure and regeneration of Sal (Shorea robusta Gaertn f.) forests in Shiwalik Region of Kumaun Himalaya. India. Indian Journal of Forestry 40(1):1–8

Babraham Bioinformatics - FastQC A Quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 24 May 2017.

Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics 33:2583–2585

Bhandari MS, Meena RK, Shamoon A, Saroj S, Kant R, Pandey S (2020) First *de novo* genome specific development, characterization and validation of simple sequence repeat (SSR) markers in Genus Salvadora. Mol Biol Rep 47(9):6997–7008. https://doi.org/10.1007/s11033-020-05758-z

Biswal B, Jena B, Giri AK, Acharya L (2021) De novo transcriptome and tissue specific expression analysis of genes associated with biosynthesis of secondary metabolites in Operculina turpethum (L.). Sci Rep 11(1):1–5. https://doi.org/10.1038/s41598-021-01906-y

Bosamia TC, Mishra GP, Thankappan R, Dobaria JR (2015) Novel and stress relevant EST derived SSR markers developed and validated in peanut. Plos One 10(6):e0129127. https://doi.org/10.1371/journal.pone.0129127

Brousseau L, Tinaut A, Duret C, Lang T, Garnier-Gere P, Scotti I (2014) High-throughput transcriptome sequencing and preliminary functional analysis in four Neotropical tree species. BMC Genomics 15(1):1–3. https://doi.org/10.1186/1471-2164-15-238

Cao CP, Finkeldey R, Siregar IZ, Siregar UJ, Gailing O (2006) Genetic diversity within and among populations of Shorea leprosula Miq. and Shorea parvifolia Dyer (Dipterocarpaceae) in Indonesia detected by AFLPs. Tree Genet Genomes 2(4):225–39. https://doi.org/10.1007/s11295-006-0046-0

Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:884–890

Chen S, Dong M, Zhang Y, Qi S, Liu X, Zhang J, Zhao J (2020) Development and characterization of simple sequence repeat markers for, and genetic diversity analysis of Liquidambar formosana. Forests 11(2):203. https://doi.org/10.3390/f11020203

Cvetković T, Hinsinger DD, Strijk JS (2019) Exploring evolution and diversity of Chinese Dipterocarpaceae using next-generation sequencing. Sci Rep 9(1):1–1. https://doi.org/10.1038/s41598-019-48240-y

Damerval C, Citerne H, Conde e Silva N, Deveaux Y, Delannoy E, Joets J, Simonnet F, Staedler Y, Schönenberger J, Yansouni J, Le Guilloux M (2019) Unraveling the developmental and genetic mechanisms underpinning floral architecture in Proteaceae. Front Plant Sci 10:18. https://doi.org/10.3389/fpls.2019.00018

de la Paz C-M, Ahn SJ, Vogel H, Heckel DG (2011) Transcriptional responses underlying the hormetic and detrimental effects of the plant secondary metabolite gossypol on the generalist herbivore *Helicoverpa armigera*. BMC Genomics 12(1):1–6. https://doi.org/10.1186/1471-2164-12-575

De Vicente MC, Lopez C, Fulton T (2004) Genetic diversity analysis with molecular marker data: learning module, vol 2. Rome and Cornell University, New York, IPGRI

Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. Focus 12:13–15

Du L, Zhang C, Liu Q, Zhang X, Yue B (2018) Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design. Bioinformatics 34(4):681–683. https://doi.org/10.1093/bioinformatics/btx665

Emami-Khoyi A, Parbhu SP, Ross JG, Murphy EC, Bothwell J, Monsanto DM, Vuuren BJV et al (2020) De novo transcriptome assembly and annotation of liver and brain tissues of common brushtail possums (*Trichosurusvulpecula*) in New Zealand: transcriptome diversity after decades of population control. Genes 11:436. https://doi.org/10.3390/genes11040436

Eriksson G, Ekberg I, Clapham D (2001) An introduction to forest genetics. Genetic Center, Department of Plant Biology and Forest Genetics, SLU

Escudero A, Iriondo JM, Torres ME (2003) Spatial analysis of genetic diversity as a tool for plant conservation. Biol Cons 113(3):351–365. https://doi.org/10.1016/S0006-3207(03)00122-8

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

Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048

Feng S, He R, Lu J, Jiang M, Shen X, Jiang Y, Wang ZA, Wang H (2016) Development of SSR markers and assessment of genetic diversity in medicinal *Chrysanthemum morifolium* cultivars. Front Genet 7:113. https://doi.org/10.3389/fgene.2016.00113

Finkeldey R, Hattemer HH (2007) Tropical Forest genetics. Springer, Berlin

Gao Z, Wu J, Liu ZA, Wang L, Ren H, Shu Q (2013) Rapid microsatellite development for tree peony and its implications. BMC Genomics 1:1–1. https://doi.org/10.1186/1471-2164-14-886

García C, Guichoux E, Hampe A (2018) A comparative analysis between SNPs and SSRs to investigate genetic variation in a juniper species (Juniperus phoenicea ssp. turbinata). Tree Genet Genomes 14(6):1–9. https://doi.org/10.1007/s11295-018-1301-x

Gautam KH, Devoe NN (2006) Ecological and anthropogenic niches of sal (Shorea robusta Gaertn. f.) forest and prospects for multiple-product forest management–a review. Forestry 79(1):81–101. https://doi.org/10.1093/forestry/cpi063

Gautam MK, Tripathi AK, Manhas RK (2007) Indicator species for the natural regeneration of *Shorea robusta* Gaertn. f.(sal). Current science 93(10):1359–61. http://www.jstor.org/stable/24099342.

Hamrick JL (1983) The distribution of genetic variation within and among natural plant populations. Conserv Genet 335–363

Hamrick JL, Godt MJ (1989) Allozyme diversity in plant species. In: Brown AHD, Clegg MT, Kahler AL, Weir BS (eds) Plant population genetics, breeding and genetic resources, pp 43–63

Hasan MI, Rahman MH, Islam MB, Islam MZ, Hossain MA, Moni MA (2022) Systems biology and bioinformatics approach to identify blood-based signatures molecules and drug targets of patient with COVID-19. Inform Med 28:100840. https://doi.org/10.1016/j.imu.2021.100840

Hedrick PW (2004) Recent developments in conservation genetics. For Ecol Manage 197(1–3):3–19. https://doi.org/10.1016/j.foreco.2004.05.002

Hevroy TH, Moody ML, Krauss SL, Gardner MG (2013) Isolation, via 454 sequencing, characterization and transferability of microsatellites for Grevillea thelemanniana subsp. thelemanniana and cross-species amplification in the Grevillea thelemanniana complex (Proteaceae). Conserv Genet Resour 5(3):887–90

Ho WS, Wickneswari R, Mahani MC, Shukor MN (2006) Comparative genetic diversity studies of *Shorea curtisii* (Dipterocarpaceae): an assessment using SSR and DAMD markers. J Trop for Sci 1:22–35

Hou B, Feng S, Wu Y (2017) Systemic identification of *Hevea brasiliensis* EST-SSR markers and primer screening. J Nucl Acids. https://doi.org/10.1155/2017/6590902

Huang C, Yin Q, Khadka D, Meng K, Fan Q, Chen S, Liao W (2019) Identification and development of microsatellite (SSRs) makers of *Exbucklandia* (Hamamelidaceae) by high-throughput sequencing. Mol Biol Rep 46(3):3381–3386. https://doi.org/10.1007/s11033-019-04800-z

Huang G, Liao X, Han Q, Zhou Z, Liang K, Li G, Yang G, Tembrock LR, Wu Z, Wang X (2022) Integrated metabolome and transcriptome analyses reveal dissimilarities in the anthocyanin synthesis pathway between different developmental leaf color transitions in *Hopea hainanensis* (Dipterocarpaceae). Front Plant Sci 3:453

Huynh T, Xu S (2018) Gene annotation easy viewer (GAEV): integrating KEGG's gene function annotations and associated molecular pathways. F1000 Res. https://doi.org/10.12688/f1000research.14012.3

Iacobas S, Ede N, Iacobas DA (2019) The gene master regulators (GMR) approach provides legitimate targets for personalized, time-sensitive cancer gene therapy. Genes 10(8):560. https://doi.org/10.3390/genes10080560

Ieremie I, Ewing RM, Niranjan M (2022) TransformerGO: predicting protein–protein interactions by modelling the attention between sets of gene ontology terms. Bioinformatics 38(8):2269–2277. https://doi.org/10.1093/bioinformatics/btac104

Kanehisa M (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30. https://doi.org/10.1093/nar/28.1.27

Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M (2019) New approach for understanding genome variations in KEGG. Nucleic Acids Res 47(D1):D590–D595. https://doi.org/10.1093/nar/gky962

Karthikeyan A, Pathak SK, Kumar A, Kumar S, Bashir A, Singh A, Sahoo NR, Mishra BP (2021) Selection and validation of differentially expressed metabolic and immune genes in weaned Ghurrah versus crossbred piglets. Trop Anim Health Prod 53(1):1–9. https://doi.org/10.1007/s11250-020-02440-1

Konuma A, Tsumura Y, Lee CT, Lee SL, Okuda T (2000) Estimation of gene flow in the tropical-rainforest tree *Neobalanocarpus heimii* (Dipterocarpaceae), inferred from paternity analysis. Mol Ecol 9(11):1843–1852. https://doi.org/10.1046/j.1365-294x.2000.01081.x

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359

Lee SL, Wickneswari R, Mahani MC, Zakri AH (2000) Genetic diversity of a tropical tree species, Shorea leprosula Miq. (Dipterocarpaceae), in Malaysia: implications for conservation of genetic resources and tree improvement 1. Biotropica 32(2):213–24. https://doi.org/10.1111/j.1744-7429.2000.tb00464.x

Lee SL, Tani N, Ng KK, Tsumura Y (2004) Isolation and characterization of 20 microsatellite loci for an important tropical tree *Shorea leprosula* (Dipterocarpaceae) and their applicability to *S. parvifolia*. Mol Ecol Notes 2:222–225. https://doi.org/10.1111/j.1471-8286.2004.00623.x

Li D, Liu C-M, Luo R, Sadakane K, Lam T-W (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674–1676

Li J, Guo H, Wang Y, Zong J, Chen J, Li D, Li L, Wang J, Liu J (2018) High-throughput SSR marker development and its application in a centipedegrass (Eremochloa ophiuroides (Munro) Hack.) genetic diversity analysis. Plos one 13(8):e0202605. https://doi.org/10.1371/journal.pone.0202605

Liewlaksaneeyanawin C, Ritland CE, El-Kassaby YA, Ritland K (2004) Single-copy, species-transferable microsatellite markers developed from loblolly pine ESTs. Theor Appl Genet 109(2):361–369. https://doi.org/10.1007/s00122-004-1635-7

Lim LS, Wickneswari R, Lee SL, Latiff A (2002) Genetic variation of *Dryobalanops aromatica* Gaertn. F. (Dipterocarpaceae) in Peninsular Malaysia using microsatellite DNA markers. For Genet 2:125–136

Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129. https://doi.org/10.1093/bioinformatics/bti282

Liu H, Xie X, Gao X, Liu H, Li Y (2017) Stability analysis of SSR in multiple wind farms connected to series-compensated systems using impedance network model. IEEE Trans Power Syst 33(3):3118–3128. https://doi.org/10.1109/TPWRS.2017.2764159

Liu FM, Zhang NN, Liu XJ, Yang ZJ, Jia HY, Xu DP (2019) Genetic diversity and population structure analysis of *Dalbergia odorifera* germplasm and development of a core collection using microsatellite markers. Genes 10(4):281. https://doi.org/10.3390/genes10040281

López-Gartner G, Cortina H, McCouch SR, Moncada MD (2009) Analysis of genetic structure in a sample of coffee (Coffea arabica L.) using fluorescent SSR markers. Tree Genet Genomes 5(3):435–46. https://doi.org/10.1007/s11295-008-0197-2

Mariotti R, Cultrera NG, Mousavi S, Baglivo F, Rossi M, Albertini E, Alagna F, Carbone F, Perrotta G, Baldoni L (2016) Development, evaluation, and validation of new EST-SSR markers in olive (Olea europaea L.). Tree Genet Genomes 12(6):1–4. https://doi.org/10.1007/s11295-016-1077-9

Matschiner M, Salzburger W (2009) TANDEM: integrating automated allele binning into genetics and genomics workflows. Bioinformatics 25:1982–1983. https://doi.org/10.1093/bioinformatics/btp303

Meena RK, Negi N, Uniyal N, Bhandari MS, Sharma R, Ginwal HS (2021) Genome skimming-based STMS marker discovery and its validation in temperate hill bamboo *Drepanostachyum falcatum*. J Genet 100(28). https://doi.org/10.1007/s12041-021-01273-7

Mishra G, Meena RK, Pandey S, Kant R, Bhandari MS (2020) A century old regeneration problem of *Shorea robusta* Gaertn. F. in south Asia: SWOT analysis. Annals of Silvicultural Research 46(1). https://doi.org/10.12899/asr-2131

Monfared MA, Samsampour D, Sharifi-Sirchi GR, Sadeghi F (2018) Assessment of genetic diversity in Salvadora persica L based on inter simple sequence repeat (ISSR) genetic marker. J Genetic Eng Biotechnol 16(2):661–7. https://doi.org/10.1016/j.jgeb.2018.04.005

Nadeem MA, Nawaz MA, Shahid MQ, Doğan Y, Comertpay G, Yıldız M, Hatipoğlu R, Ahmad F, Alsaleh A, Labhane N, Özkan H (2018) DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. Biotechnol Biotechnol Equip 32(2):261–285. https://doi.org/10.1080/13102818.2017.1400401

Nand A, Zhan Y, Salazer OR, Aranda M, Voolstra CR, Dekker J (2020) Chromosome-scale assembly of the coral endosymbiont symbiodinium microadriaticum genome provides insight into the unique biology of dinoflagellate chromosomes. bioRxiv. https://doi.org/10.1101/2020.07.01.182477

Ng KK, Lee SL, Koh CL (2004) Spatial structure and genetic diversity of two tropical tree species with contrasting breeding systems and different ploidy levels. Mol Ecol 13(3):657–669. https://doi.org/10.1046/j.1365-294X.2004.02094.x

Ng KK, Lee SL, Tsumura Y, Ueno S, Ng CH, Lee CT (2009) Expressed sequence tag–simple sequence repeats isolated from *Shorea leprosula* and their transferability to 36 species within the Dipterocarpaceae. Mol Ecol Resour 9(1):393–398. https://doi.org/10.1111/j.1755-0998.2008.02238.x

Ng KK, Kobayashi MJ, Fawcett JA, Hatakeyama M, Paape T, Ng CH, Ang CC, Tnah LH, Lee CT, Nishiyama T, Sese J (2021) The genome of *Shorea leprosula* (Dipterocarpaceae) highlights the ecological relevance of drought in aseasonal tropical rainforests. Commun Biol 4(1):1–4. https://doi.org/10.1038/s42003-021-02682-1

Nieuwenhuis R, Hesselink T, van den Broeck HC, Cordewener J, Schijlen E, Bakker L, Trivino SD, Struss D, de Hoop SJ, de Jong H, Peters SA (2021) Genome and transcriptome architecture of allopolyploid okra (*Abelmoschusesculentus*). BioRxiv. https://doi.org/10.1101/2021.11.18.469076

Nock CJ, Baten A, Barkla BJ, Furtado A, Henry RJ, King GJ (2016) Genome and transcriptome sequencing characterises the gene space of *Macadamia integrifolia* (Proteaceae). BMC Genomics 17(1):1–2. https://doi.org/10.1186/s12864-016-3272-3

Nybom H (2004) Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. Mol Ecol 13(5):1143–1155. https://doi.org/10.1111/j.1365-294X.2004.02141.x

Obayashi K, Tsumura Y, Ihara-Ujino T, Niiyama K, Tanouchi H, Suyama Y, Washitani I, Lee CT, Lee SL, Muhammad N (2002) Genetic diversity and outcrossing rate between undisturbed and selectively logged forests of *Shorea curtisii* (Dipterocarpaceae)

using microsatellite DNA analysis. Int J Plant Sci 163(1):151–158. https://doi.org/10.1086/324549

Ohtani M, Ueno S, Tani N, Lee LS, Tsumura Y (2012) Twenty-four additional microsatellite markers derived from expressed sequence tags of the endangered tropical tree *Shorea leprosula* (Dipterocarpaceae). Conserv Genet Resour 4(2):351–354. https://doi.org/10.1007/s12686-011-9546-9

Pandey M, Geburek T (2009) Successful cross-amplification of Shorea microsatellites reveals genetic variation in the tropical tree. Shorea Robusta Gaertn Hereditas 146(1):29–32. https://doi.org/10.1111/j.1601-5223.2009.02070.x

Pandey M, Geburek T (2010) Genetic differences between continuous and disjunct populations: some insights from sal (*Shorea robusta* Roxb.) in Nepal. Conserv Genet 3:977–984. https://doi.org/10.1007/s10592-009-9940-y

Pandey M, Geburek T (2011) Fine-scale genetic structure and gene flow in a semi-isolated population of a tropical tree, *Shorea robusta* Gaertn. (Dipterocarpaceae). Curr Sci 10:293–299

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. https://doi.org/10.1093/bioinformatics/bts460

Powell W, Morgante M, Andre C, Hanafey M, Vogel J, Tingey S, Rafalski A (1996) The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Mol Breeding 3:225–238. https://doi.org/10.1007/BF00564200

Pratiwi RH, Oktarina E, Mangunwardoyo W, Hidayat I, Saepudin E (2022) Antimicrobial compound from endophytic Pseudomonas azotoformans UICC B-91 of *Neesiaaltissima* (Malvaceae). Pharmacogn J 14(1):172–181. https://doi.org/10.5530/pj.2022.14.23

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155(2):945–959

Rai MK, Shekhawat JK, Kataria V, Shekhawat NS (2017) Cross species transferability and characterization of microsatellite markers in *Prosopis cineraria*, a multipurpose tree species of Indian Thar Desert. Arid Land Res Manag 31(4):462–471. https://doi.org/10.1080/15324982.2017.1338791

Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J (2019) g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 47(W1):W191–W198. https://doi.org/10.1093/nar/gkz369

Rohlf FJ (1998) NTSYS-pc: numerical taxonomy and multivariate analysis system, version 2.02e. Setauket: Applied Biostatistics Inc., Exeter Software.

Sae-Lim P, Naktang C, Yoocha T, Nirapathpongporn K, Viboonjun U, Kongsawadworakul P, Tangphatsornruang S, Narangajavana J (2019) Unraveling vascular development-related genes in laticifer-containing tissue of rubber tree by high-throughput transcriptome sequencing. Curr Plant Biol 19:100112. https://doi.org/10.1016/j.cpb.2019.100112

Satya P, Chakraborty A, Jana S, Majumdar S, Karan M, Sarkar D, Datta S, Mitra J, Kar CS, Karmakar PG, Singh NK (2017) Identification of genic SSR s in jute (*Corchorus capsularis*, Malvaceae) and development of markers for phenylpropanoid biosynthesis genes and regulatory genes. Plant Breeding 136(5):784–797. https://doi.org/10.1111/pbr.12514

Schulman AH (2007) Molecular markers to assess genetic diversity. Euphytica 158(3):313–321. https://doi.org/10.1007/s10681-006-9282-5

Seng HW, Ling PS, Lau P, Jusoh I (2011) Sequence variation in the cellulose synthase (SpCesA1) gene from Shorea parvifolia ssp parvifolia mother trees. Pertanika J Trop Agric Sci 34(2):317–23

Shah M, Jaan S, Fatima B et al (2021) Delineating novel therapeutic drug and vaccine targets for *Staphylococcus cornubiensis* NW1T through computational analysis. Int J Pept Res Ther 27:181–195. https://doi.org/10.1007/s10989-020-10076-w

Sheriff O, Alemayehu K (2018) Genetic diversity studies using microsatellite markers and their contribution in supporting sustainable sheep breeding programs: a review. Cogent Food Agric 4(1):1459062. https://doi.org/10.1080/23311932.2018.1459062

Stojnić S, Avramidou VE, Fussi B, Westergren M, Orlović S, Matović B, Trudić B, Kraigher H, Aravanopoulos A, Konnert FM (2019) Assessment of genetic diversity and population genetic structure of Norway spruce (Picea abies (l.) Karsten) at its southern lineage in Europe Implications for conservation of forest genetic resources. Forests 10(3):258. https://doi.org/10.3390/f10030258

Suoheimo J, Li C, Luukkanen O (1999) Isozyme variation of natural populations of sal (*Shorea robusta*) in the Terai region, Nepal. Silvae Genetica (Germany).

Surabhi GK, Mohanty S, Meher RK, Mukherjee AK, Vemireddy LN (2017) Assessment of genetic diversity in *Shorea robusta*: an economically important tropical tree species. J Appl Biol Biotechnol 5(2):1–1. https://doi.org/10.7324/JABB.2017.50218

Tam NM, Duy VD, Duc NM, Giap VD, Xuan BT (2014) Genetic variation in and spatial structure of natural populations of *Dipterocarpus alatus* (Dipterocarpaceae) determined using single sequence repeat markers. Genet Mol Res 13(3):5378–5386. https://doi.org/10.4238/2014.July.24.17

Tang L, Liao X, Tembrock LR, Ge S, Wu Z (2022) A chromosome-scale genome and transcriptomic analysis of the endangered tropical tree *Vaticamangachapoi* (Dipterocarpaceae). DNA Research 29(2):dsac005. https://doi.org/10.1093/dnares/dsac005

Tinio CE, Finkeldey R, Prinz K, Fernando ES (2014) Genetic variation in natural and planted populations of *Shorea guiso* (Dipterocarpaceae) in the Philippines revealed by microsatellite DNA markers. Asia Life Sci 23:75–91

Ujino T, Kawahara T, Tsumura Y, Nagamitsu T, Yoshimaru H, Ratnam W (1998) Development and polymorphism of simple sequence repeat DNA markers for *Shorea curt*isii and other Dipterocarpaceae species. Heredity 81(4):422–428. https://doi.org/10.1046/j.1365-2540.1998.00423.x

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) Micro-checker: software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes 4:535–538. https://doi.org/10.1111/j.1471-8286.2004.00684.x

Vu DD, Bui TT, Nguyen MD, Shah SN, Vu DG, Zhang Y, Nguyen MT, Huang XH (2019) Genetic diversity and conservation of two threatened dipterocarps (Dipterocarpaceae) in southeast Vietnam. J For Res 30(5):1823–1831. https://doi.org/10.1007/s11676-018-0735-1

Wang X, Chen W, Luo J, Yao Z, Yu Q, Wang Y, Zhang S, Liu Z, Zhang M, Shen Y (2019) Development of EST-SSR markers and their application in an analysis of the genetic diversity of the endangered species *Magnolia sinostellata*. Mol Genet Genomics 294(1):135–147. https://doi.org/10.1007/s00438-018-1493-7

Wang C, Ma X, Ren M, Tang L (2020) Genetic diversity and population structure in the endangered tree *Hopeahainanensis* (Dipterocarpaceae) on Hainan Island China. Plos One 15(11):e0241452. https://doi.org/10.1371/journal.pone.0241452

Wang L, Li H, Suo Y, Han W, Diao S, Mai Y, Sun P, Fu J (2021) Development of EST-SSR markers and their application in the genetic diversity of persimmon (Diospyros kaki Thunb.). Trees 35(1):121–33. https://doi.org/10.1007/s00468-020-02024-4

Wright S (1978) Evolution and the genetics of populations: a treatise in four volumes. In: Variability within and among natural populations, vol 4. University of Chicago Press, Chicago.

Xiang X, Zhang Z, Wang Z, Zhang X, Wu G (2015) Transcriptome sequencing and development of EST-SSR markers in *Pinus*

*dabeshanensis*, an endangered conifer endemic to China. Mol Breeding 35(8):1. https://doi.org/10.1007/s11032-015-0351-0

Xu W, Yang Q, Huai H, Liu A (2012) Development of EST-SSR markers and investigation of genetic relatedness in tung tree. Tree Genet Genomes 8(4):933–940. https://doi.org/10.1007/s11295-012-0481-z

Xu M, Zhu S, Xu R, Lin N (2020) Identification of CELSR2 as a novel prognostic biomarker for hepatocellular carcinoma. BMC Cancer 20(1):1–5. https://doi.org/10.1186/s12885-020-06813-5

Xue L, Liu Q, Hu H et al (2018) The southwestern origin and eastward dispersal of pear (Pyrus pyrifolia) in East Asia revealed by comprehensive genetic structure analysis with SSR markers. Tree Genet Genom 14:48. https://doi.org/10.1007/s11295-018-1255-z

Yamazaki S, Tanaka Y, Araki H, Kohda A, Sanematsu F, Arasaki T, Duan X, Miura F, Katagiri T, Shindo R, Nakano H (2017) The AP-1 transcription factor JunB is required for Th17 cell differentiation. Sci Rep 7(1):1–4. https://doi.org/10.1038/s41598-017-17597-3

Yan X, Zhang X, Lu M, He Y, An H (2015) De novo sequencing analysis of the *Rosaroxburghii* fruit transcriptome reveals putative ascorbate biosynthetic genes and EST-SSR markers. Gene 561(1):54–62. https://doi.org/10.1016/j.gene.2015.02.054

Yang W, Wang K, Zhang J, Ma J, Liu J, Ma T (2017) The draft genome sequence of a desert tree *Populuspruinosa*. Gigascience 6(9):gix075. https://doi.org/10.1093/gigascience/gix075

You J, Qi S, Du Y, Wang C, Su G (2020) Multiple bioinformatics analyses of integrated gene expression profiling data and verification of hub genes associated with diabetic retinopathy. Med Sci Monit 26:e923146. https://doi.org/10.12659/MSM.923146

Yu F, Wang BH, Feng SP, Wang JY, Li WG, Wu YT (2011) Development, characterization, and cross-species/genera transferability of SSR markers for rubber tree (*Hevea brasi*liensis). Plant Cell Rep 30(3):335–344. https://doi.org/10.1007/s00299-010-0908-7

Zhang T, Zhang K, Zhou T, Zhou R, Ge Y, Wang Z, Shao H, Zhang D, Li K (2021) *De novo* assembly and SSR loci analysis in *Gasterophilusnasa*lis (Diptera: Oestridae). Entomol Res 51(6):305–314. https://doi.org/10.1111/1748-5967.12505

Zucchi MI, Pinheiro JB, Chaves LJ, Coelho AS, Couto MA, Morais LK, Vencovsky R (2005) Genetic structure and gene flow of *Eugenia dysenterica* natural populations. Pesq Agrop Brasileira 40(10):975–980