Physiol Mol Biol Plants (July 2021) 27(7):1623–1638
https://doi.org/10.1007/s12298-021-01035-w

RESEARCH ARTICLE

Genome skimming-based simple sequence repeat (SSR) marker discovery and characterization in *Grevillea robusta*

Aman Dabral[1] · Arzoo Shamoon[1] · Rajendra K. Meena[1] · Rama Kant[1] · Shailesh Pandey[2] · Harish S. Ginwal[1] · Maneesh S. Bhandari[1]

Received: 3 March 2021 / Revised: 1 July 2021 / Accepted: 6 July 2021 / Published online: 17 July 2021

Abstract Proteaceae, a largely southern hemisphere family consisting of 80 genera distributed in Australia and southern Africa as its centres of greatest diversity, also extends well in northern and southern America. Under this family, *Grevillea robusta* is a fast-growing species got popularity in farm and avenue plantations. Despite the ecological and economic importance, the species has not yet been investigated for its genetic improvement and genome-based studies. Only a few molecular markers are available for the species or its close relatives, which hinders genomic and population genetics studies. Genetic markers have been intensively applied for the main strategies in breeding programs, especially for the economically important traits. Hence, it is of utmost priority to develop genomic database resources and species-specific markers for studying quantitative genetics in *G. robusta*. Given this, the present study aimed to develop de novo genome sequencing, robust microsatellites markers, sequence annotation and their validation in different stands of *G. robusta* in northern India. Library preparation and sequencing were carried out using Illumina paired-end sequencing technology. Approximately, ten gigabases (Gb) sequence data with 70.87 million raw reads assembled into 425,923 contigs (read mapped to 76.48%) comprising 455 Mb genome size (23 × coverage) generated through genome skimming approach. In total, 9421 simple sequence repeat (SSR) primer pairs were successfully designed from 13,335 microsatellite repeats. Afterward, a subset of 161 primer pairs was randomly selected, synthesized and validated. All the tested primers showed successful amplification but only 13 showed polymorphisms. The polymorphic SSRs were further used to estimate the measures of genetic diversity in 12 genotypes each from the states of Punjab, Haryana, Himachal Pradesh and Uttarakhand. Importantly, the average number of alleles (N_a), observed heterozygosity (H_o), expected heterozygosity (H_e), and the polymorphism information content (PIC) were recorded as 2.69, 0.356, 0.557 and 0.388, respectively. The availability of sequence information and newly developed SSR markers could potentially be used in various genetic analyses and improvements through molecular breeding strategies for *G. robusta*.

Keywords *Grevillea robusta* · Genome sequencing · Microsatellite markers · Genetic diversity

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

Aman Dabral
amandabral93@gmail.com

Arzoo Shamoon
arzoo707@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, Uttarakhand 248 195, India

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

Introduction

Exotics forestry species, viz. *Populus*, *Casuarina*, *Eucalyptus*, *Grevillea*, etc., have demonstrated their impact and utility in the plantation forestry in India. Despite the fast-growing nature, exceptionally high adaptability to diverse climatic conditions and desirable timber qualities, *Grevillea robusta* is a poorly investigated and under-utilized tree species in the agroforestry system of the country. *Grevillea robusta* with ploidy level of 2n = 20 (Sugiura 1936; Ramsay 1963; Venkata Rao 1957; Stace et al. 1998; http://ccdb.tau.ac.il/), belongs to the family Proteaceae, which comprised of more than 80 genera and about 1700 species. This species is recognized as one of the largest plant genera native to the Australasian region (McGillivray and Makinson 1993; Harwood 1997; Makinson 2000; Weston 2007). However, to fulfill the demand of timber, *G. robusta* was introduced to all the continents across the globe and predominantly cultivated in south Asia, some parts of eastern Asia, southern Africa, Latin America, Caribbean region, northwestern and southeastern America, and few regions of Europe (Orwa et al. 2009; http://www.plantsoftheworldonline.org); which shows the species potential to adapt in varied climatic and edaphic conditions (Luna 2005). In India also, *G. robusta* is widely distributed throughout the country (World Agroforestry Centre, 2002).

Globally, the agroforestry system needs diversification to cater to the need for economic development and environmental sustainability, but at the same time the scientific knowledge base in most forestry tree species is still lacking. The exploration, characterization and documentation of base populations are the foremost requirements for tree improvement programmes. In these strides, the first step is to quantify the level of genetic diversity and variability in existing genetic resources across the distribution range, which is more important if the species is exotic and tends to have a narrow genetic base. Molecular markers are the most preferred tool for carrying out genetic studies, viz. estimation of genetic diversity, population genetics, evolutionary and phylogenetic studies, marker assisted breeding, gene or genome mapping, quantitative trait loci (QTL) mapping, and marker trait association studies (Kordrostami and Rahimi 2015; Nadeem et al. 2018). Under these, microsatellites, or simple sequence repeat (SSR) markers are highly valuable and widely used tools in plant genetic studies due to their abundance in genome, polymorphism, co-dominance and high reproducibility. Based on their origin, i.e., genome or transcriptome, the SSRs could be categorized as genomic and genic SSRs, which could preferably be used for specific applications (Li et al. 2012; Vieira et al. 2016; Colburn et al. 2017; Chen et al. 2020). Despite the importance and wide applicability of SSRs, their application is limited to the few trees owing to the unavailability of genome sequence information in most forestry species. However, due to the advent of advanced sequencing technology, it is now feasible to generate the sequence data (genomic or transcriptome) for SSRs mining in any species. Genome skimming is the most rapid and cost-effective sequencing approach to target high copy fractions of the genome through random shearing and multiplexing (Xia et al. 2018; Nevill et al. 2020).

It is hypothesized that the tree improvement programme of any exotic species has immense scientific importance to get the current perspectives of genetic variability, structural dynamics, adaptation and evolution. Therefore, it must be commenced with proper evaluation and genetic characterization of the base populations. Notably, exploring the factors affecting species improvement will help in the evaluation of the economic value of a trait of interest. As exotic *G. robusta* is often found in the plantation zone in an inhabited site (roadside, avenues, government institutions, schools, hospitals, industrial areas, etc.) either through social forestry practices or as an agroforestry species, which do not form any natural population in India. Thus, the term “stand” is used instead of population for *G. robusta*, which revealed ‘a unit of trees that is relatively homogeneous in age, structure, composition and physical environment’ (Smith 1962; Oliver and Larson 1996). Therefore, the availability of the low-depth genome sequence information of *G. robusta* opens new doors to unlocking the genetic potential of species for use as bioenergy crops. In addition, specific genetic information enables “smart” breeding and selection alternatives, along with the straightforward genetic manipulation of *G. robusta*. Given these facts, the present study aimed to: (1) generate genome sequence information using high throughput next-generation sequencing (NGS) approach and its functional annotation; (2) develop and validate the novel species-specific SSR markers; and (3) utilize them for characterizing distantly located stands of *G. robusta* in northern India.

Material and methods

Sample collection and DNA extraction

Field surveys were conducted for the collection of leaf samples, and a total of 48 genotypes (12 genotypes per stand (Std) of *G. robusta*), i.e. trees were sampled from four states, namely Punjab (Std 1: GRPB), Haryana (Std 2: GRHR), Himachal Pradesh (Std 2: GRHP) and Uttarakhand (Std 2: GRUK) in northern India, along with their geospatial attributes, viz. longitude, latitude and altitude (Supplementary Table 1). Leaf tissues were instantly desiccated with silica gel and brought to the laboratory of the Genetics and Tree Improvement division, Forest Research Institute, Dehradun, and stored at − 80 °C till further use. Total genomic DNA was extracted using the protocol given by Doyle and Doyle (1990) with minor modifications.

Genome sequencing and SSRs mining

The genome sequence data of about 10 Gb were generated using Illumina Protocols through a service provider (Clevergene Biocorp Private Limited, Bengaluru, Karnataka). Raw sequence data were subjected to quality check by recording the parameters, such as base call quality distribution, % bases above Q20 and Q30, % GC, and adapter contamination using program FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC (Ewels et al. 2016). The quality-based filtration and trimming of low-quality reads were done using the program Trim galore (https://github.com/FelixKrueger/TrimGalore). After quality trimming, the reads above 70 bp were assembled de novo with varying Kmer sizes, viz. 49, 58, 67, 76, 85 and 94 using the program ABySS genome assembler (Simpson et al. 2009). Based on the quality parameters, viz. number of contigs, % reads aligned, L50, L75, N50 and N75 values, the best Kmer assembly was selected for further SSRs mining. Afterward, genome coverage was determined as per the formula (https://genohub.com; https://www.illumina.com):

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

The repeat sequences in the assembly were masked using the program Repeatmasker (https://www.repeatmasker.org/faq.html#faq3). The assembled contigs were scanned to mine out microsatellite repeat motifs using Perl scripts-based program MIcroSAtellite (MISA) identification tool (Beier et al. 2017; https://pgre.ipk-gatersleben.de/misa). The relative abundance (loci Mb^{-1}) and density (bp Mb^{-1}) of identified SSRs were derived using the Linux-based Krait tool (Du et al. 2018). Further, the distribution of identified microsatellite motifs in different genomic regions were manually done from the draft file—General Feature Format (GFF) generated through Augustus Linux-based program (Hoff et al. 2019). Primer sequences were designed with default parameters using a web-based Primer3 program (https://bioinfo.ut.ee/primer3).

Functional annotation

To enhance the potential utility of the SSR markers, the contigs in which they nested were subjected to homology search against the non-redundant protein database using NCBI BLASTX (Johnson et al. 2008; https://blast.ncbi.nlm.nih.gov/Blast.cgi). The Kyoto Encyclopedia of Genes and Genomes (KEGG) database was further used to understand the functional utilities of the sequence data in the biological system (Kanehisa 2000; https://www.genome.jp/kegg/). Notably, the KEGG PATHWAY (KP) maps were drawn to represent molecular interaction and reaction; whereas, KEGG BRITE (KB) was used to understand the functional hierarchies of biological objects. In addition, KEGG ORTHOLOGY (KO) numbers obtained from the KEGG server were further annotated through the standalone tool, i.e. Gene Annotation Easy Viewer (GAEV) to summarize the parameters, such as gene name, gene orthologs and functional pathways (Huynh and Xu 2018). Finally, functional enrichment analysis of the genes obtained through KO was done through g:Profiler (Raudvere et al. 2019).

SSR amplification and data analysis

A subset of 161 primer pairs was synthesized for their validation through the polymerase chain reaction (PCR) amplification with genomic DNA of *G. robusta* in thermal cycler machines (Make: Eppendorf Mastercycler Nexus). The primer pairs were first subjected to gradient PCR to screen and optimize the annealing temperature (T_m). Amplification through PCR was carried out with a 15 µl PCR reaction mixture, containing 40 ng of template DNA, 1.5 µl of 10 × PCR buffer, 1.75 mM $MgCl_2$, 0.2 mM dNTPs, 100 nM of each forward and reverse primer, 0.6 units of *Taq* DNA polymerase and nuclease-free sterile water. The cycling conditions included an initial denaturation at 95 °C for 5 min, then 35 cycles of 95 °C for 1 min, primer-specific T_m (52.0 °C to 62.0 °C) for 1 min, 72 °C for 1 min; and a final extension at 72 °C for 10 min. The PCR products were separated using 2% agarose gel buffered with 1 × TBE (Tris/borate/EDTA) along with a 100 bp DNA ladder. The gel was stained with ethidium bromide (0.5 µg ml^{-1}) and visualized in a gel documentation system (Make: UVP). The primer pairs producing a clear and distinct band within the expected product size were considered to be positively amplified, which were further evaluated for polymorphism by subjecting them to PCR amplification in about 25 random samples. The PCR products were resolved in 4% high-resolution agarose (Make: Sigma-Aldrich) and the primer pairs displaying multiple sized alleles across the genotypes were marked as polymorphic.

The polymorphic primer pairs were used to genotype the individuals collected from diverse locations and the band profile generated with each SSR was scored manually by assigning an approximate allele size to each band. The allelic data were further adjusted as per the periodicity of

SSR repeat motifs using allele binning program Tandem ver. 1.07 (Matschiner and Salzburger 2009). To minimize the scoring and amplification errors, null alleles were detected using the program Microchecker ver. 2.2 (Van Oosterhout et al. 2004). Eventually, the marker data were analyzed to calculate the genetic characteristics, 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 principal coordinate analysis (PCoA) using GenAlEx ver. 6.5 (Peakall and Smouse, 2012). The analysis of molecular variance (AMOVA), inbreeding coefficient (F_{IS}) and genetic differentiation (F_{ST}) were also investigated using the program GenAlEx. The polymorphism information content (PIC) of the tested markers was calculated using program PowerMarker ver. 3.25 (Liu and Muse 2005). Further, the stand structure of the 48 genotypes with 13 loci was assessed using STRUCTURE ver. 2.2 (Pritchard et al. 2000). Based on admixture models and correlated band frequencies, the number of sub populations (K) were determined from 1 to 10 (Evanno et al. 2005). The Jaccard similarity coefficient was used to determine the genetic similarity between the genotypes using NTSYS-pc ver. 2.10 (Rohlf 1998). Finally, using the unweighted pair group method with arithmetic mean (UPGMA) and the SAHN clustering tool, cluster analysis was used to generate a dendrogram based on the similarity matrix data.

Results

Genome sequencing, assembly, SSRs identification and primer design

The library preparation and sequencing were carried out using Illumina paired-end low depth sequencing technology, and approximately ten gigabases (10 Gb) sequence data were generated with 70.87 million raw reads (read length = 150) assembled into 425,923 contigs (23 × coverage), which show read mapped value of 76.48% and genome size of 455 Mb (Table 1). The quality parameters, viz. GC content (41.52%), bases above Q20 (99.54%) and Q30 (94.97%) depicted that the quality of sequence data generated were reasonable and suitable for further processing. After filtration and trimming, the quality reads with the size more than 70 bp were de novo assembled into the contigs. Based on varying Kmer, parameters such as contigs sizes, L50 and N50 values, and percentage of aligned reads, assembly with Kmer 94 was selected for analysis. In total, 425,923 high-quality contigs were obtained with the L50 and N50 values of 122,470 and 1170, respectively. The cleaned raw data were deposited in the Sequence Read Archive (SRA) database of the NCBI with an accession number PRJNA612735.

Table 1 Summary statistics of genome sequenced data

Sl. No	Features	Value
1	Total raw reads	70,878,358
2	Total number of bases (bp)	10,631,753,700
3	Clean reads proportion (%)	76.48
4	Total number of contigs	425,923
5	Assembly length (bp)	455,331,789
6	Largest contig size (bp)	389,860
7	L50	122,470
8	L75	242,531
9	N50	1170
10	N75	778
11	GC content (%)	41.52
12	Q20%	99.54
13	Q30%	94.97
14	Number of SSR identified through MISA	13,335
15	Number of SSR primers designed	9421
16	Number of SSRs tested	161
17	Number of working SSR markers	161
18	Number of polymorphic SSR markers	13

A total of 13,335 SSRs were mined out in the assembled genomic data of *G. robusta.* The di- and tri-nucleotide repeats were the most dominant in the genome with the frequency of 80.69% and 17.82%, respectively. Among the repeat motifs, AG/CT was most prevalent in di- and AAG/CTT in tri-nucleotides. Whereas, tetra-, penta- and hexa-nucleotides repeats occurred in very low frequency (Fig. 1a–d). Furthermore, the relative abundance and density of each repeat types were also determined, which revealed di-nucleotides had the highest relative abundance (9.65 loci Mb^{-1}) and density (140.18 bp Mb^{-1}) followed by tri- (5.31 loci Mb^{-1}; 80.38 bp Mb^{-1}), tetra- (4.01 loci Mb^{-1}; 65.94 bp Mb^{-1}) and penta-nucleotides (1.47 loci Mb^{-1}; 29.43 bp Mb^{-1}). The primer pairs were successfully designed for 9421 microsatellite loci containing repeat length $\geq$ 12 bases and possessed optimal flanking regions. Eventually, the distribution of identified microsatellite motifs in different genomic regions of *G. robusta* revealed that the di-nucleotides SSRs were the most abundant (58.7%) category, followed by tri- (35.5%) and tetra-nucleotides (5.7%). The different genomic regions, such as genic, transcript, intron and exon also confirmed that di-nucleotides were the most abundant type. However, in the CoDing Sequences (CDS) region, tri-nucleotides were the most abundant (61.36%) (Fig. 1e).

The SSRs were prefaced as 'GRGMS', which stands for '*Grevillea robusta* Genomic Microsatellite' marker. Based

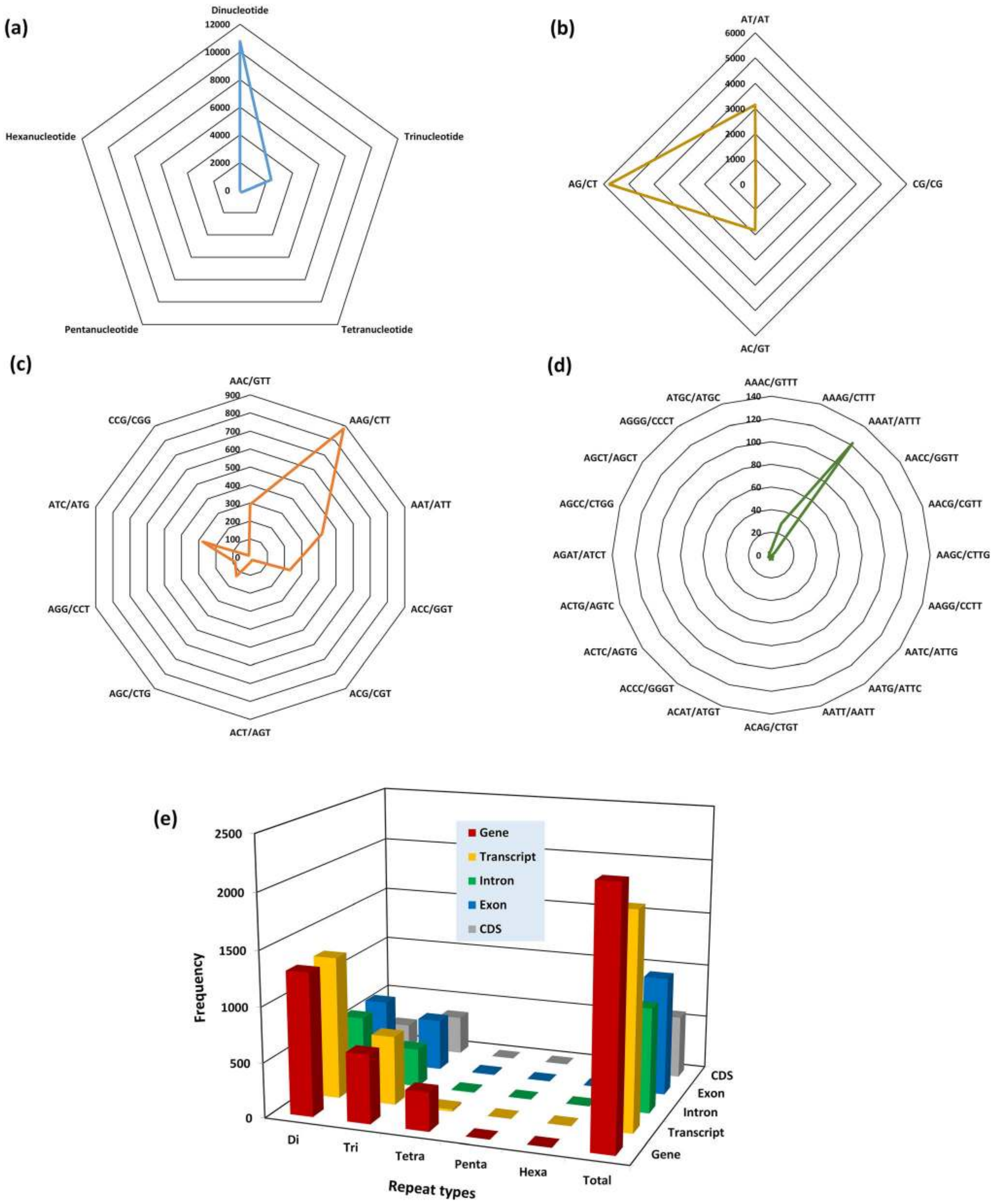


Fig. 1 The simple sequence repeat (SSR) types generated through Illumina sequencing: **a** Radar indicates frequency of all types of SSR motifs; and **b–d** most predominant repeat motifs, i.e. di-, tri- and tetra-nucleotides; and **e** the distribution of identified microsatellite motifs in different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

on the frequency of microsatellite repeats, proportionate numbers of primer pairs were selected in each class for validation. In total, 161 microsatellite primer pairs were tested for PCR amplification in *G. robusta*, and all were successfully amplified but only 13 primer pairs revealed a polymorphic banding pattern [Fig. 2a–b and Supplementary Fig. 1a–k].

Functional annotation

Sequence similarity search was done against non-redundant protein databases configured in NCBI's BLASTX to derive the putative functions of all the polymorphic SSRs (Table 2). Herein, the KEGG database resources (such as KP, KB and KO) were used to understand the high-level functional hierarchies, revealing information about the biological pathways and their utilities in the genomic data. Further, the KP was used to examine the metabolic pathways and related functions which are likely to be encoded in the genome of *G. robusta*; hence, 9488 out of 425,923 contigs were successfully mapped into 391 pathways. The maximum number of contigs were involved in ribosome (216 contigs) followed by RNA transport (154 contigs), spliceosome (142 contigs), oxidative phosphorylation (126 contigs), protein processing in endoplasmic reticulum (126 contigs), endocytosis (96 contigs), ribosome biogenesis in eukaryotes (93 contigs), cell cycle (91 contigs), cysteine and methionine metabolism (74 contigs), purine metabolism (69 contigs), amino sugar and nucleotide sugar metabolism (69 contigs), etc. (Supplementary Table 2). Similarly, functional hierarchies derived through KB were characterized into three categories (protein families), namely (i) metabolism, (ii) genetic information processing and (iii) signalling and cellular processes (Fig. 3). Finally, the list of KO numbers attained through KEGG analysis were further annotated through the GAEV. As a result, biological pathways are sorted by the number of associated genes (Supplementary Table 3), where the highest were involved in metabolic pathways (2234 genes) followed by biosynthesis of secondary metabolites (1140 genes), carbon metabolism (248 genes), plant hormone signal transduction (130 genes), etc.

Afterward, functional enrichment analysis of genes obtained through KEGG were done through g:Profiler, which accomplishes statistical enrichment analysis to find the information through predefined parameters, such as Gene Ontology (GO) terms, biological pathways, regulatory DNA elements, protein–protein interaction networks, etc. The results based on their GO ID and *p*-values were further characterized into three categories, namely metabolic component, cellular component and biological component. Importantly, the highest number of GO terms (337) were involved in the biological process, followed by cellular component (99) and molecular function (67), as revealed by Manhattan-like plot (Fig. 4) derived through g:Profiler. In this plot, hovering the circle highlights with an identifier, which was referred in the table (below plot) showing the detailed information, such as data source GO ID and term name with the corresponding p-value. For example, cellular process (*p*-value = 3.652 × 10^{-64}) is shown by GO:0009987 followed by metabolic (*p*-value = 3.287 × 10^{-58}) and cellular metabolic process (*p*-value = 4.835 × 10^{-58}). This revealed clustering of the genes in descending order based on the p-value assigned to a particular process and a list of top 100 terms representing the process is elaborated in detail (Supplementary Fig. 2).

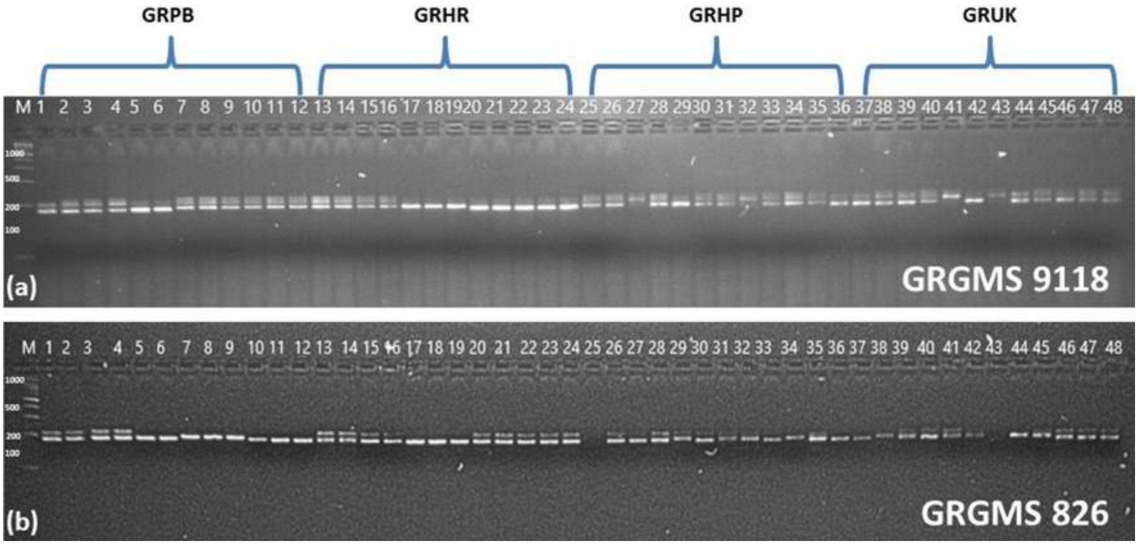

GRPB
GRHR
GRHP
GRUK
M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
1000
500
200
100
(a)
GRGMS 9118
M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
1000
500
200
100
(b)
GRGMS 826

Fig. 2 Representative sample* through SSRs showing polymorphic banding pattern in *G. robusta*: **a** GRGMS 9118, and **b** GRGMS 826. *Where, M: 100 bp DNA ladder; 1–48 representing 12 genotypes each from 4 stands

Table 2 Characteristics and putative functions of 13 polymorphic SSR types (GRGMS) with E-value of *G. robusta*

Sl. No	Locus name	Primer sequence	Product size (bp)	Repeat motifs	T_m (°C)	Putative functions	E–value
1	GRGMS 9118	CACCCTTCCTACCCCAAATGT TCGGTTTGAAGCATCACCCA	180	(TA)6	58	Not found	–
2	GRGMS 332	GGAGAACGAAGACGACCCAG GGTAGGAAGGTCGCGGTTAA	203	(AT)6	58.4	*Camellia sinensis* var. sinensis hypothetical protein TEA_009066	5e − 44
3	GRGMS 7755	GGGGAATGTGTATGCATTGGG AGCTTCATTTCCTTTCCCCCT	199	(GTAT)5	60.4	Not found	–
4	GRGMS 826	ACCTCTGCATGTGTTAGCCA GCGGCAGCGTAGTAACAGTA	180	(TA)7	60.5	*Vitis vinifera* hypothetical protein VITISV_004764	1e − 23
5	GRGMS 2493	CAGGAGGAGAAATGGGAAGGA AGACACTTACCACTGTCGGC	182	(AG)6	58	Not found	–
6	GRGMS 2500	CGTAAAACCCCTCCCCCTAC TTCCATGGTGTTGTGGAGGG	187	(TC)6	52.7	*Nelumbo nucifera* PREDICTED: pentatricopeptide repeat-containing protein At2g22410, mitochondrial-like	6e − 23
7	GRGMS 10729	GGCGGTTTCAGTGGTTCTTT AACACACTCACTCTCTGGGC	180	(AG)6	52.5	Not found	–
8	GRGMS 2198	TGCGTGCCTTTCACTCAAGA GCCATAATTTCGGTCGGCAG	202	(AT)8	53	Not found	–
9	GRGMS 2148	CGTTGAAAAAGAGCGCGGTT TTAGTGGAGAGGTCGTGGGT	184	(TA)8	52	Not found	–
10	GRGMS 4184	GGAGAAGAGCCAAATTCTGCT AGTTGCGTATGTTTCAGTCCT	166	(ATTT)5	62	Not found	–
11	GRGMS 8376	TGCCCAGGAGCCTAAAATCC TGGGTTCCAATTGTCACAAGC	163	(AAAT)5	61.8	Not Found	–
12	GRGMS 9833	TCCCTTGTGGTTCCTTCTGC GACAGTGGCCACTTAACATGC	180	(AT)8	60.5	*Cephalotus* follicularis CRAL_TRIO domain-containing protein/ CRAL_TRIO_N domain-containing protein	2e − 30
13	GRGMS 8539	TCGGCAACATAAACCCACCT ACGATCACCTATCAAAACGTGA	200	(AAAT)5	53	Not found	–

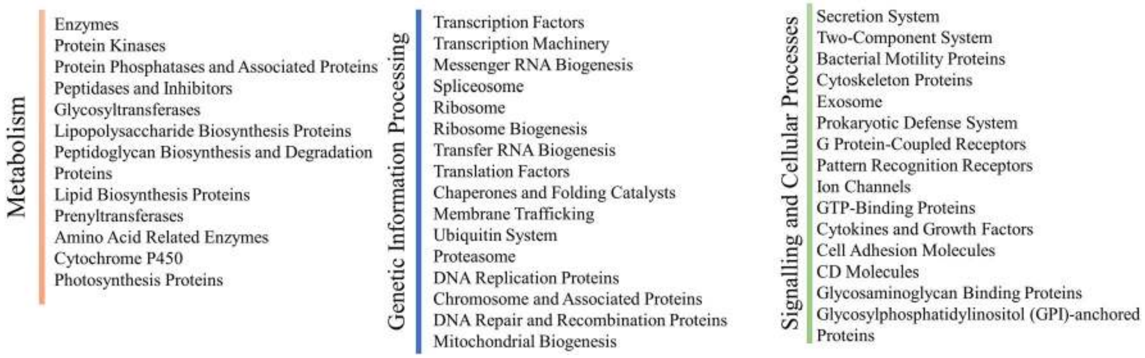
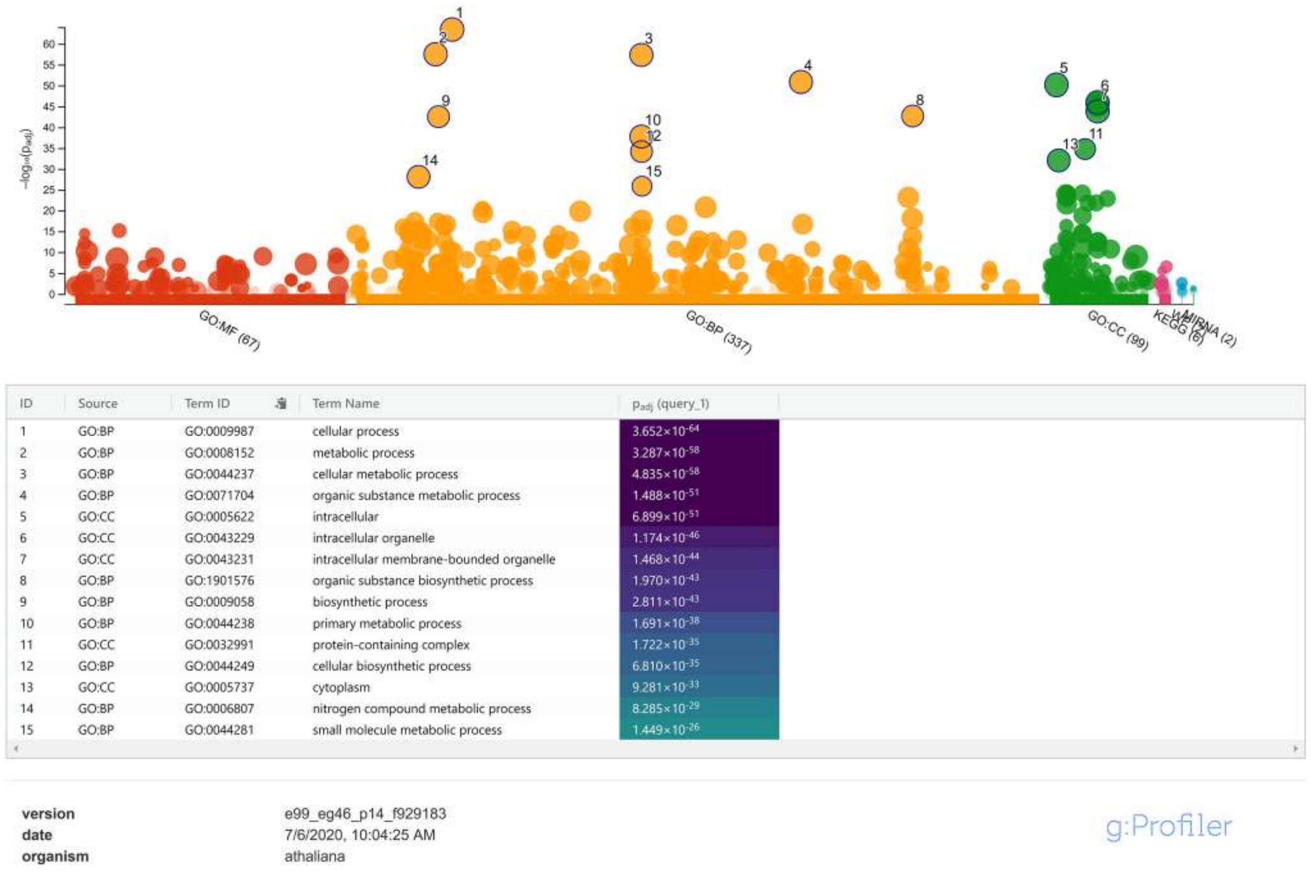

Metabolism
Enzymes
Protein Kinases
Protein Phosphatases and Associated Proteins
Peptidases and Inhibitors
Glycosyltransferases
Lipopolysaccharide Biosynthesis Proteins
Peptidoglycan Biosynthesis and Degradation Proteins
Lipid Biosynthesis Proteins
Prenyltransferases
Amino Acid Related Enzymes
Cytochrome P450
Photosynthesis Proteins
Genetic Information Processing
Transcription Factors
Transcription Machinery
Messenger RNA Biogenesis
Spliceosome
Ribosome
Ribosome Biogenesis
Transfer RNA Biogenesis
Translation Factors
Chaperones and Folding Catalysts
Membrane Trafficking
Ubiquitin System
Proteasome
DNA Replication Proteins
Chromosome and Associated Proteins
DNA Repair and Recombination Proteins
Mitochondrial Biogenesis
Signalling and Cellular Processes
Secretion System
Two-Component System
Bacterial Motility Proteins
Cytoskeleton Proteins
Exosome
Prokaryotic Defense System
G Protein-Coupled Receptors
Pattern Recognition Receptors
Ion Channels
GTP-Binding Proteins
Cytokines and Growth Factors
Cell Adhesion Molecules
CD Molecules
Glycosaminoglycan Binding Proteins
Glycosylphosphatidylinositol (GPI)-anchored Proteins

Fig. 3 Functional hierarchies obtained through KEGG BRITE



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:BP	GO:0009987	cellular process	3.652×10^{-64}
2	GO:BP	GO:0008152	metabolic process	3.287×10^{-58}
3	GO:BP	GO:0044237	cellular metabolic process	4.835×10^{-58}
4	GO:BP	GO:0071704	organic substance metabolic process	1.488×10^{-51}
5	GO:CC	GO:0005622	intracellular	6.899×10^{-51}
6	GO:CC	GO:0043229	intracellular organelle	1.174×10^{-46}
7	GO:CC	GO:0043231	intracellular membrane-bounded organelle	1.468×10^{-44}
8	GO:BP	GO:1901576	organic substance biosynthetic process	1.970×10^{-43}
9	GO:BP	GO:0009058	biosynthetic process	2.811×10^{-43}
10	GO:BP	GO:0044238	primary metabolic process	1.691×10^{-38}
11	GO:CC	GO:0032991	protein-containing complex	1.722×10^{-35}
12	GO:BP	GO:0044249	cellular biosynthetic process	6.810×10^{-35}
13	GO:CC	GO:0005737	cytoplasm	9.281×10^{-33}
14	GO:BP	GO:0006807	nitrogen compound metabolic process	8.285×10^{-29}
15	GO:BP	GO:0044281	small molecule metabolic process	1.449×10^{-26}

version	e99_eg46_p14_f929183
date	7/6/2020, 10:04:25 AM
organism	athaliana

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

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

Polymorphic primers were utilized to calculate the key marker characteristics and diversity measures through genotyping of 48 genotypes of *G. robusta* (Table 3). Further, errors in the fragment separation and allele scoring were removed by binning as per their repeat motifs. Accordingly, marker data were analyzed and none of the thirteen primer pairs showed evidence of null allele. Hence, full data sets were used for estimating genetic diversity measures. In total, 35 alleles were generated with 13 SSRs across the genotypes with an average of 2.69 alleles per locus, which were further checked for polymorphism information content (PIC). Results revealed that the PIC value of each SSR primer pair ranged from 0.161 to 0.557, with a mean value of 0.356. Overall, H_o for the primers was recorded in the range of 0 to 1 with a mean of 0.557, while H_e was ranged between 0 to 0.726 with a mean of 0.388. Importantly, AMOVA revealed that most of the genetic variation (94.23%) was confined among the individuals within a stand, and only 5.77% variance was recorded among the stands. Owing to the cross-pollinated behavior, very low genetic differentiation ($F_{\mathrm{ST}} = 0.075$) and high gene flow (Nm = 4.54) were recorded among the stands. Consequently, the value of inbreeding coefficient ($F_{\mathrm{IS}} = -0.399$) also indicates an excess of heterozygotes in the sampled populations.

In addition, structural analysis reveals an optimum K value of 2 [Supplementary Fig. 3a(i–iii)], which is very low to predict any output. Hence, no structuring was detected in the sampled stands of *G. robusta*. The intraspecific genetic diversity analysis using SSR markers further yielded a PCoA plot (Supplementary Fig. 3b) and UPGMA dendrogram (Fig. 5), which showed 48 genotypes have been split clearly with a similarity coefficient of 0.870 into two groups (Gp) of *G. robusta*, i.e. GpI and GpII consisted of 13 and 35 genotypes, respectively. The former one was subdivided with a similarity coefficient of 0.890 into two subgroups (SbGp), i.e. SbGpIa (11 genotypes) and SbGpIb (2 genotypes representing GRPB Std). While the later Gp was split with a similarity coefficient of 0.874 into SbGpIIa (30 genotypes) and SbGpIIb (5 genotypes representing GRHR, GRHP and GRUK Std).

Table 3 Genetic polymorphism of 13 SSR loci in four *G. robusta* stands (Std)

Sl. No	Locus	Std 1: GRPB (n = 12)			Std 2: GRHR (n = 12)			Std 3: GRHP (n = 12)			Std 4: GRUK (n = 12)			PIC	F_{IS}
		N_a	H_o	H_e	N_a	H_o	H_e	N_a	H_o	H_e	N_a	H_o	H_e		
1	GRGMS 9118	3	0.833	0.653	3	0.333	0.625	2	0.750	0.497	2	0.750	0.497	0.557	-0.040
2	GRGMS 332	2	1.000	0.500	2	1.000	0.500	2	1.000	0.500	2	1.000	0.500	0.375	-1.000
3	GRGMS 7755	4	0.417	0.462	2	0.750	0.469	3	0.300	0.265	3	0.455	0.368	0.362	-0.128
4	GRGMS 826	2	0.333	0.278	2	0.750	0.469	2	0.636	0.434	2	0.545	0.397	0.323	-0.385
5	GRGMS 2493	2	0.600	0.420	2	0.111	0.401	1	0.000	0.000	2	0.500	0.375	0.275	0.048
6	GRGMS 2500	2	0.417	0.330	1	0.000	0.000	2	0.167	0.153	2	0.182	0.165	0.161	-0.098
7	GRGMS 10729	2	0.200	0.180	1	0.000	0.000	2	0.583	0.413	2	0.417	0.330	0.228	-0.173
8	GRGMS 2198	3	0.917	0.601	2	0.167	0.153	2	0.143	0.133	2	0.100	0.095	0.314	-0.015
9	GRGMS 2148	4	0.833	0.726	2	1.000	0.500	3	0.333	0.438	3	0.636	0.541	0.550	-0.152
10	GRGMS 4184	3	0.833	0.569	2	1.000	0.500	2	0.333	0.278	2	0.727	0.463	0.394	-0.483
11	GRGMS 8376	2	0.750	0.469	3	0.700	0.505	2	0.571	0.408	2	0.800	0.480	0.470	-0.357
12	GRGMS 9833	2	1.000	0.500	2	0.889	0.494	2	0.714	0.459	2	1.000	0.500	0.373	-0.850
13	GRGMS 8539	2	0.500	0.375	2	0.182	0.165	2	0.500	0.375	2	0.300	0.255	0.254	-0.212
	Range	2–4	0.200–1.000	0.180–0.726	1–3	0.000–1.000	0.000–0.625	1–3	0.000–1.000	0.000–0.500	2–3	0.100–1.000	0.095–0.541	0.161–0.557	-1.000–0.048
	Mean	2.538	0.664	0.466	2.000	0.529	0.368	2.077	0.464	0.335	2.154	0.570	0.382	0.356	-0.324

Where, n: number of individuals collected for each population, N_a: number of alleles, H_o: observed heterozygosity, H_e: expected heterozygosity, PIC: polymorphism information content, and F_{IS}: inbreeding coefficient; and GRPB = *G. robusta* Punjab, GRHR = *G. robusta* Haryana, GRHP = *G. robusta* Himachal Pradesh, and GRUK = *G. robusta* Uttarakhand

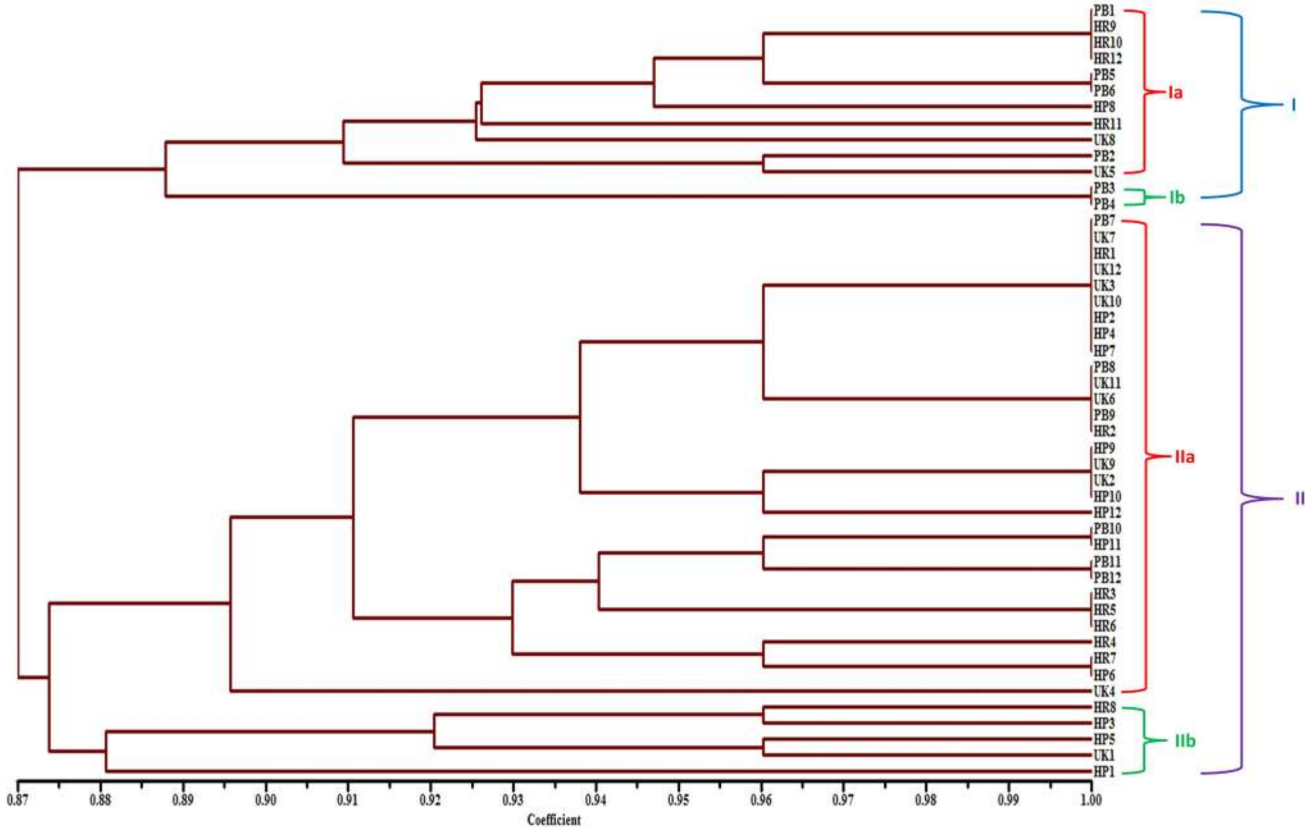


Fig. 5 The UPGMA dendrogram unbiased measures of genetic distance among 48 genotypes representing 4 stands of *G. robusta*

Discussion

Wider applicability of molecular markers in forest genetics revealed that microsatellites particularly SSRs have a vital role in population genetic analysis, stand structuring and tree breeding, which facilitates the conservation and tree improvement programme of a particular species. Despite the economic importance and fast-growing timber species with wider adaptability and distribution, *G. robusta* is relatively less investigated for its genetic analysis and improvement. Consequently, limited genomic resources have been generated in *G. robusta*, particularly the sequence data (genome or transcriptome or expressed sequence tags (ESTs)) and sequence-based markers. For instance, five SSR markers were developed in *G. robusta* through the microsatellite enriched genomic library method (Mantello et al. 2011). Notably, the development of SSR markers through low-depth high throughput Illumina-based NGS technology is one of the widely used efficient and cost-effective methods in any taxa that devoid the sequence information (Zou et al. 2014; Abdelkrim et al. 2018; Li et al. 2018; Nadeem et al. 2018; Taheri et al. 2018). This approach has been recently used to develop microsatellite markers in various species, viz. *G. 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) and *Salvadora oleoides* (Bhandari et al. 2020).

In the current study, 70.87 million raw reads (approximately 10 Gb sequence data comprises read length = 150) were generated, and de novo assembled into 425,923 contigs (read mapped = 76.4%) representing genome size of 455 Mb with a coverage of 23×. A total of 13,335 SSRs were successfully identified, where maximum relative abundance and density were obtained for di-nucleotides (9.65 loci Mb^{-1}; 140.18 bp Mb^{-1}) followed by tri-, tetra- and penta-nucleotides. It revealed that most of the non-genic region contains a large set of di-nucleotides repeats, which is similar to other taxon studies (Karaoglu et al. 2005; Patil et al. 2021; Sigang et al. 2021; Zhu et al. 2021). The SSR repeat analysis revealed that di- and tri-nucleotide repeats were almost ubiquitous in the genome, and these classes were further dominated by repeat motifs AG/CT and AAT/CTT, respectively. In other species, such as *S. oleoides* highest number of di- and tri-nucleotides repeats were found for AT/AT and AAT/ATT, respectively

(Bhandari et al. 2020); whereas, in the case of *Drepanostachyum falcatum*, AG/CT and CCG/CGG were observed in maximum number (Meena et al. 2021). The distribution of genomic SSRs were also determined in *G. robusta*, and di-nucleotides were the most abundant category except in CDS, which was mostly dominated by tri-nucleotides. Further, penta- and hexa-nucleotides SSRs were not present to be distributed in any of the genomic regions, as they are scantly (only 5) developed through low-depth high-throughput Illumina-based sequencing. It is evident from the distribution pattern that most of the SSRs (52.70%) are present in the protein coding regions of the genome sequenced. It might be possible that data generated here is useful for transcriptomics, proteomics and metabolomics study for the neighboring species, which can change the tree improvement dynamics under Genus *Grevillea* (PRJNA612735). A similar sort of studies were done in *Arabidopsis thaliana* and *Oryza* sp. (Lawson et al. 2006), *Brassica rapa* (Hong et al. 2007), *Musa paradisiaca* (Biswas et al. 2020); bovid sp. (Qi et al. 2018); *Leishmania*, *Drosophila* sp., avian and primates' fauna (Srivastava et al. 2019), and *Moschus* sp. (Qi et al. 2020).

Out of 9421 successfully designed SSR primer pairs, a subset of 161 was validated through PCR amplification and 13 of which were found polymorphic (Table 1). The low level of polymorphism finding is most consistent with a single or 2–3 source(s)/provenance(s) introduction of *G. robusta* in Indian sub-continental climatic conditions from Australia. Further, polymorphic estimates of sampled genotypes were selected from geographically four distinct locations (Punjab, Haryana, Himachal Pradesh and Uttarakhand), but population source(s) for existing individual(s) is/are limited. Similarly, biogeographical population structuring and low genetic diversity of invasive *Phyla canescens* revealed a single northwest Argentine population dispersal and naturalization, which might be due to human-assisted activity (Xu et al. 2015). Moreover, despite several introductions of *Impatiens glandulifera* throughout Europe, limited genetic diversity was observed (Hagenblad et al. 2015). Likewise, the absence of genetic structure has also been observed for *Macfadyena unguis-cati* (Prentis et al. 2009) and *Olea europaea* (Besnard et al. 2007), where are both believed to have a single introduction into its exotic range.

The NGS has made it simpler to get genome-wide information and it has moved the exploration center into genome annotation. However, the challenging tasks involved in annotation depend on the available tools and techniques, and further to interpret the data contained in the sequencing. The NCBI's non-redundant protein database BLASTX was used to predict the putative functions of 13 polymorphic SSRs loci and the top-hit species were *Camellia sinensis*, *Nelumbo nucifera*, *Pistacia vera* and *Vitis vinifera* (Table 2). Subsequently, the KEGG enrichment analysis was conducted, where the KEGG database resources (KP, KB and KO)-based annotation provided the functional role of the scaffolds and the pathways in which they are involved. The KP is a collection of manually drawn pathway maps signifying the understanding of the molecular interaction and reaction networks. The KP pathway-based scrutiny is helpful to further cognize the biological functions and gene interactions. Based on KP, about 9488 contigs were successfully mapped into 391 metabolic pathways, which include galactose metabolism, pyruvate metabolism, plant-pathogen interaction, porphyrin, chlorophyll metabolism, glycolysis/gluconeogenesis, glycerophospholipid metabolism, plant hormone signal transduction, amino sugar, nucleotide sugar metabolism, etc. Next, KB incorporates different types of relationships including, genes and proteins, compounds and reactions, drugs and diseases, and organisms and cells, etc. (Kanehisa and Sato 2019). Similar studies were reported in *Hevea brasiliensis* (Li et al. 2012), *Ipomoea batatas* (Wang et al. 2010; Xie et al. 2012), *Solanum trilobatum* (Lateef et al. 2018), *Gasterophilus nasalis* (Zang et al. 2021), *Operculina turpethum* (Biswal et al. 2021) and *Juglans mandshurica* (Yan et al. 2021). Importantly, the KO was also annotated through GAEV; hence, pathways with their number of associated genes were also obtained (Iacobas et al. 2019; Emami-Khoyi et al. 2020; Nand et al. 2020; Shah et al. 2021). Here, the functional enrichment analysis of the genes were done through g:Profiler, where the results characterized into GO ID and p-value with highest involved in the biological process (337), followed by cellular component (99) and molecular function (67). Recently, this kind of characterization and annotation of genes were used to predict therapeutic drugs against COVID-19 (Tan et al. 2021), validation of immune genes (Karthikeyan et al. 2021), identification of novel prognostic biomarker (Xu et al. 2020), and analyses of Integrated Gene Expression Profiling Data (IGEPA) (You et al. 2020).

Analyzing genetic diversity is one of the key prerequisites in any species conservation and genetic improvement programme. Genomic diversity provides an adaptive and evolutionary potential to a species, and unraveling this information facilitates species protection, conservation and management. In the present study, all 161 tested primer pairs were successfully validated in *G. robusta*. The amplification rate was significantly higher as compared to *Liquidambar formosana* (72%) (Chen et al. 2020). The genotyping of 48 accessions of *G. robusta* with 13 SSRs generated 35 alleles with an average of 2.69 alleles per locus, which is similar to that recorded in Brazilian sampled individuals of *G. robusta* using isoenzymes (2.93 alleles per locus) (Sousa et al. 2018). However, this is lower to one of the members of a Proteaceae family

(*Conospermum undulatum*) in which an average of 11.45 alleles per locus were recorded for 20 microsatellite loci (Delnevo et al. 2019). In another species, namely *G. macleayana*, 41 alleles were generated by 6 microsatellite loci in a range of 2.8–4.2 alleles per locus (England et al. 2002). In two species of the genus, viz. *Persoonia longifolia* and *P. elliptica*, genotyping with SSR loci have generated an average of 14 and 13 alleles per locus, respectively (Stingemore et al. 2013). Further, 71 alleles were detected with an average of 5.9 per locus from 12 SSR loci amplified among 22 cultivars of *M. integrifolia* (Nock et al. 2016). All these studies revealed that the number of polymorphic alleles increases in a population with an increase in sample size and the number of marker loci. The PIC value of SSRs lies between 0.160 and 0.557 with a mean of 0.356 (Table 3). The PIC analyzed for the sampled genotypes was low for *G. robusta*, when compared to *Populus deltoides* (PIC = 0.535, 0.77) (Chen et al. 2020; Sharma et al. 2018), *E. camaldulensis* (PIC = 0.44 to 0.93) and *E. tereticornis* (PIC = 0.36 to 0.93) (Arumugasundaram et al. 2011). The numerous studies of SSR markers in different species, i.e. *P. tomentosa* (Du et al. 2012) and *P. euphratica* (Wang et al. 2011), yielded reproducible polymorphic bands and showed that they provide a powerful and reliable molecular tool for analyzing genetic diversity and relationships among and within species (Feng et al. 2016).

Forest maintained the key equilibrium among different taxon, implying several evolutionary processes likely to impact the genetic diversity of a forestry tree species (Porth and El-Kassaby 2014), which is required to be analyzed either for conservation genetics or/and tree improvement programme. In a population genetic analysis, H_o and H_e are considered as most suitable measures to characterize marker loci and populations (Sherif and Alemayehu 2018; Monfared et al. 2018; Xue et al. 2018). Before our study, genetic variation was examined in 23 populations of *G. robusta* across the natural range in Australia using 20 isozyme loci (Harwood et al. 1997), where the mean expected heterozygosity was recorded as low ($H_e = 0.105$) with a low level of genetic differentiation. Whereas, in another isoenzymes-based study conducted in Brazilian populations, high levels of genetic diversity ($H_o = 0.3962$ and $H_e = 0.5140$) was recorded (Sousa et al. 2018). Earlier, one of the inter simple sequence repeats (ISSR)-based studies carried out with a relatively smaller numbers of genotypes from one (GRUK) of the four presently sampled locations also revealed comparable results ($H_o = 0.290$ and $H_e = 0.305$) (Parmar et al. 2019). However, microsatellite markers are a widely accepted molecular tool for measuring genetic diversity and divergence among different genotypes and/or populations of a species. In Australia, the estimation of diversity measures was successfully demonstrated with SSR markers in the species, such as *C. undulatum* ($H_o = 0.000$ to 1.000; $H_e = 0.117$ to 0.919) (Delnevo et al. 2019); *M. integrifolia* ($H_o = 0.571$, $H_e = 0.626$) (Nock et al. 2016); *P. longifolia* ($H_o = 0.04$ to 0.88; $H_e = 0.04$ to 0.84) and *P. elliptica* ($H_o = 0.46$ to 0.93; $H_e = 0.42$ to 0.88) (Stingemore et al. 2013). In India, the species was introduced from Australia, which is now naturalized and popularized through farm plantings across the country. The results revealed that the Indian *G. robusta* stands possess moderate levels of genetic diversity ($H_o = 0.557$ and $H_e = 0.388$), which is relatively higher than their natural populations in Australia (due to choice of marker used) but lower than the Brazilian sampled genotypes.

Interestingly, the results revealed that the level of genetic differentiation for *G. robusta* in India is still very low ($F_{ST} = 0.075$), which could be attributed to the various natural as well as manual processes. This was supported by structure analysis, which showed a low K value (K = 2, default generated in case of low structuring), as most of the *G. robusta* stands are not clearly defined by any single cluster with a significant proportional membership coefficient (> 0.7). This means significant genetic admixture across the location might be a consequence of a single or not more than two ancestral gene pools (i.e. seed source / planting stock introduced in India from Australia or any other country is probably not more than two provenances). Again, similar sets of admixing were shown by PCoA and UPGMA cluster analysis, which tend to justify the low value of F_{ST}. Concomitantly, the genetic differentiation among Australian populations (Harwood et al. 1997) and Brazilian germplasm (Sousa et al. 2018) of *G. robusta* was also recorded as low. Thus, it could be inferred that the gene flow across the spatially separated populations was sufficiently maintained and not greatly affected by the ecological and geographical attributes. Other studies on genus *Grevillea*, where five populations of an endangered shrub *G. iaspicula* were taken to measure interpopulation genetic differentiation (F_{ST} or D = 0.04–0.32), was high; whereas, the inbreeding coefficient for maternal ($F_{IS} = 0.03$) and progeny ($F_{IS} = -0.09$) suggested that gene flow was limited even among the populations separated by only a few kilometers (Hoebee and Young 2001). In the Indian context, the negative value of the inbreeding coefficient ($F_{IS} = -0.399$) and the deviation between observed and expected heterozygosity have indicated the excess of the heterozygotes which could be ascribed to the predominant out-crossing and self-incompatibility mechanism reported in *G. robusta*.

Naturally, the genetic exchange was disbursed due to predominantly out-crossing breeding behaviour owed to the protandry and self-incompatibility mechanism, long distance pollen or seed dispersal by birds and insects, etc.

(Kalinganire et al. 2000, 2001). Since the species is exotic in India, a limited amount of genetic diversity is available to be maintained and transferred, and gene flow was mainly accomplished by the manual transmission of seeds or seedlings across the geographical region. Importantly, despite the narrow genetic base, the species has not undergone significant genetic structuring and inbreeding depression, which demonstrated high genomic plasticity against the selection pressure exerted by several forces, such as limited genetic diversity, phenological changes, extreme environmental conditions, overexploitation, poor seed setting, etc., in an exotic environment.

Conclusion

Irrefutably, the microsatellite marker discovery through NGS-based genome skimming has appeared as the most rapid, efficient and cost-effective approach and holds a great promise in executing genetic studies in any species, irrespective of the availability of their genomic resources. About 10 Gb genomic data generated herein for *G. robusta* has enabled the discovery of 9421 GRGMS markers, and a subset of these have successfully been validated and utilized for genotyping the stands distributed in India. All the evaluated SSRs were highly polymorphic, and demonstrated moderate genetic diversity ($H_o = 0.557$ and $H_{e-} = 0.388$) and very low genetic differentiation ($F_{ST} = 0.075$) among the sampled genotypes. The sequence information and the SSR markers generated here will serve as powerful tools for measuring the genetic diversity, marker assisted selection (MAS), gene map construction, phylogenetic and evolutionary revelation of family Proteaceae in the coming years.

Funding Indian Council of Forestry Research and Education (ICFRE), Dehradun under the project grant No. 10-1/2017-2018/Budget & Audit (Part-IInd); dated 04th July, 2017.

References

Abdelkrim J, Robertson B, Stanton JA, Gemmell N (2018) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. Biotechniques 46(3):185–192. https://doi.org/10.2144/000113084

Arumugasundaram S, Ghosh M, Veerasamy S, Ramasamy Y (2011) Species discrimination, population structure and linkage disequilibrium in *Eucalyptus camaldulensis* and *Eucalyptus tereticornis* using SSR markers. PLoS ONE 6(12):28252. https://doi.org/10.1371/journal.pone.0028252

Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics 33:2583–2585. https://doi.org/10.1093/bioinformatics/btx198

Besnard G, Henry P, Wille L, Cooke D, Chapuis E (2007) On the origin of the invasive olives (*Olea europaea* L., Oleaceae). Heredity 99(6):608–619. https://doi.org/10.1038/sj.hdy.6801037

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 gene associated with biosynthesis of medicinally active metabolites in a high valued medicinal plant *Operculina turpethum* (L.). Res Sq. https://doi.org/10.21203/rs.3.rs-312726/v1

Biswas MK, Bagchi M, Biswas D, Harikrishna JA, Liu Y, Li C, Sheng O, Mayer C, Yi G, Deng G (2020) Genome-wide novel genic microsatellite marker resource development and validation for genetic diversity and population structure analysis of Banana. Genes 11(12):1479. https://doi.org/10.3390/genes11121479

Chen C, Chu Y, Ding C, Su X, Huang Q (2020a) Genetic diversity and population structure of black cottonwood (*Populus deltoides*) revealed using simple sequence repeat markers. BMC Genet 21(1):1–12. https://doi.org/10.1186/s12863-019-0805-1

Chen S, Dong M, Zhang Y, Qi S, Liu X, Zhang J, Zhao J (2020b) 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

Colburn BC, Mehlenbacher SA, Sathuvalli VR (2017) Development and mapping of microsatellite markers from transcriptome sequences of European hazelnut (*Corylus avellana* L.) and use for germplasm characterization. Mol Breed 37(2):16. https://doi.org/10.1007/s11032-016-0616-2

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) Unravelling the developmental and genetic mechanisms underpinning floral architecture in Proteaceae. Front Plant Sci. https://doi.org/10.3389/fpls.2019.00018

Delnevo N, Piotti A, van Etten EJ, Stock WD, Byrne M (2019) Isolation, characterization, and cross-amplification of 20 microsatellite markers for *Conospermum undulatum* (Proteaceae). APPS 7(8):e11283. https://doi.org/10.1002/aps3.11283

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

Du Q, Wang B, Wei Z, Zhang D, Li B (2012) Genetic diversity and population structure of Chinese white poplar (*Populus tomentosa*) revealed by SSR markers. J Hered 103(6):853–862. https://doi.org/10.1093/jhered/ess061

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 (*Trichosurus vulpecula*) in New Zealand: transcriptome diversity after decades of population control. Genes 11:436. https://doi.org/10.3390/genes11040436

England PR, Usher AV, Whelan RJ, Ayre DJ (2002) Microsatellite diversity and genetic structure of fragmented populations of the rare, fire-dependent shrub *Grevillea macleayana*. Mol Ecol 11(6):967–977. https://doi.org/10.1046/j.1365-294X.2002.01500.x

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRU CTU RE: 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, Kaller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048. https://doi.org/10.1093/bioinformatics/btw354

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

Hagenblad J, Hülskötter J, Acharya KP, Brunet J, Chabrerie O, Cousins SA, Dar PA, Diekmann M, De Frenne P, Hermy M, Jamoneau A (2015) Low genetic diversity despite multiple introductions of the invasive plant species *Impatiens glandulifera* in Europe. BMC Genet 16(1):1–16. https://doi.org/10.1186/s12863-015-0242-8

Harwood CE, Moran GF, Bell JC (1997) Genetic differentiation in natural populations of *Grevillea robusta*. Aust J Bot 45(4):669–678. https://doi.org/10.1071/BT96073

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):887–890. https://doi.org/10.1007/s12686-013-9918-4

Hoebee SE, Young AG (2001) Low neighbourhood size and high interpopulation differentiation in the endangered shrub *Grevillea iaspicula* McGill (Proteaceae). Heredity 86(4):489–496. https://doi.org/10.1046/j.1365-2540.2001.00857.x

Hoff KJ, Stanke M (2019) Predicting genes in single genomes with AUGUSTUS. Curr Protoc Bioinform 65(1):e57. https://doi.org/10.1002/cpbi.57

Hong CP, Piao ZY, Kang TW, Batley J, Yang T, Hur Y, Bhak J, Park B, Edwards D, Lim YP (2007) Genomic distribution of simple sequence repeats in *Brassica rapa*. Mol Cells 23(3):349. https://www.researchgate.net/publication/6191416

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

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

Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL (2008) NCBI BLAST: a better web interface. Nucleic Acids Res 36(suppl_2):W5-W9. https://doi.org/10.1093/nar/gkn201

Kalinganire A, Harwood CE, Slee MU, Simons AJ (2001) Pollination and fruit-set of *Grevillea robusta* in western Kenya. Austral Ecol 26(6):637–648. https://doi.org/10.1046/j.1442-9993.2001.01139.x

Kalinganire A, Harwood CE, Slee MU, Simons AJ (2000) Floral structure, stigma receptivity and pollen viability in relation to protandry and self-incompatibility in silky oak (*Grevillea robusta* A. Cunn.). Ann Bot 86(1):133–148. https://doi.org/10.1006/anbo.2000.1170

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 (2019) KEGG mapper for inferring cellular functions from protein sequences. Protein Sci. https://doi.org/10.1002/pro.3711

Karaoglu H, Ying Lee CM, Meyer W (2005) Survey of simple sequence repeats in completed fungal genomes. Mol Bio Evo 22(3):639–649. https://doi.org/10.1093/molbev/msi057

Karthikeyan A, Pathak SK, Kumar A et al (2021) Selection and validation of differentially expressed metabolic and immune genes in weaned Ghurrah versus crossbred piglets. Trop Anim Health Prod 53:14. https://doi.org/10.1007/s11250-020-02440-1

Kordrostami M, Rahimi M (2015) Molecular markers in plants: concepts and applications. Genet. 3rd Millenn 13:4024–403

Lateef A, Prabhudas SK, Natarajan P (2018) RNA sequencing and de novo assembly of *Solanum trilobatum* leaf transcriptome to identify putative transcripts for major metabolic pathways. Sci Rep 8:15375. https://doi.org/10.1038/s41598-018-33693-4

Lawson MJ, Zhang L (2006) Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. Genome Biol 7(2):1–1. https://doi.org/10.1186/gb-2006-7-2-r14

Li D, Deng Z, Qin B, Liu X, Men Z (2012) De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis Muell*. Arg.). BMC genomics, 13(1):1–14. http://www.biomedcentral.com/1471-2164/13/192

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):0202605. doi:https://doi.org/10.1371/journal.pone.0202605

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

Luna RK (2005) Plantation trees. International Book Distributors, India, pp 397–399

Makinson RO (2000) '*Grevillea*. Flora of Australia 17A.' (Australian Biological Resources Study: Canberra/CSIRO: Melbourne). https://doi.org/10.1007/978-3-540-32219-1_42

Mantello C, Kestring DR, Sousa VA, Aguiar AV, Souza AP (2011) Development and characterization of microsatellite loci in *Grevillea robusta*. BMC Proc 5(7):1–2. https://doi.org/10.1186/1753-6561-5-S7-P16

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

McGillivray DJ, Makinson RO (1993) Grevillea*,* proteaceae: a taxonomic revision. Melbourne University Press

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

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. JGEB 16:661–667. 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 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

Nevill PG, Zhong X, Tonti-Filippini J, Byrne M, Hislop M, Thiele K et al (2020) Large scale genome skimming from herbarium material for accurate plant identification and phylogenomics. Plant Methods 16:1. https://doi.org/10.1186/s13007-019-0534-5

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 Genom 17(1):1–12. https://doi.org/10.1186/s12864-016-3272-3

Oliver CD, Larson BA (1996) Forest stand dynamics, update edition. Yale School of the Environment Other Publications. 1. https://elischolar.library.yale.edu/fes_pubs/1

Orwa C, Mutua A, Kindt R, Jamnadass R, Simons A (2009) Agroforestry tree database: a tree reference and selection guide version 4.0 (http://www.worldagroforestry.org/af/treedb/)

Parmar P, Dabral A, Meena RK, Pandey S, Kant R, Bhandari MS (2019) Genetic diversity analysis in *Grevillea robusta* using ISSR molecular markers. Indian for 145(3):260–265

Patil PG, Singh NV, Bohra A, Raghavendra KP, Mane R, Mundewadikar DM, Babu KD, Sharma J (2021) Comprehensive characterization and validation of chromosome-specific highly polymorphic SSR markers from Pomegranate (*Punica granatum* L.) cv. Tunisia Genome. Front. Plant Sci 12:337. https://doi.org/10.3389/fpls.2021.645055

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

Porth I, El-Kassaby YA (2014) Assessment of the genetic diversity in forest tree populations using molecular markers. Divers 6(2):283–295. https://doi.org/10.3390/d6020283

Prentis PJ, Sigg DP, Raghu S, Dhileepan K, Pavasovic A, Lowe AJ (2009) Understanding invasion history: genetic structure and diversity of two globally invasive plants and implications for their management. Divers Distrib 15(5):822–830. https://doi.org/10.1111/j.1472-4642.2009.00592.x

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

Qi WH, Lu T, Zheng CL, Jiang XM, Jie H, Zhang XY, Yue BS, Zhao GJ (2020) Distribution patterns of microsatellites and development of its marker in different genomic regions of forest musk deer genome based on high throughput sequencing. Aging (albany NY) 12(5):4445. https://doi.org/10.18632/aging.102895

Qi WH, Jiang XM, Yan CC, Zhang WQ, Xiao GS, Yue BS, Zhou CQ (2018) Distribution patterns and variation analysis of simple sequence repeats in different genomic regions of bovid genomes. Sci Rep 26:8(1):1–3. https://doi.org/10.1038/s41598-018-32286-5

Ramsay HP (1963) Chromosome numbers in the Proteaceae. Aust J Bot 11(1):1–20. https://doi.org/10.1071/BT9630001

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

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

Sharma S, Dobhal S, Thakur S (2018) Analysis of genetic diversity in parents and hybrids of *Populus deltoides* Bartr. using microsatellite markers. Appl. Biol Res 20(3):262–270. https://doi.org/10.5958/0974-4517.2018.00036.8

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

Sigang F, Hao H, Yong L, Pengfei W, Chao Z, Lulu Y, Xiuting Q, Qiu L (2021) Genome-wide identification of microsatellite and development of polymorphic SSR markers for spotted sea bass (*Lateolabrax maculatus*). Aquac Rep 20:100677. https://doi.org/10.1016/j.aqrep.2021.100677

Simpson JT, Wong K, Jackman SD, Schein JE et al (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123. https://doi.org/10.1101/gr.089532.108

Smith DM (1962) The practice of silviculture, 7th edn. Wiley, New York, p 578

Sousa VAD, Kalil Filho AN, Martins EG, Shimizu JY, Albertin F (2018) Gene diversity in *Grevillea* populations introduced in Brazil and its implication on management of genetic resources. Rev Arvore. https://doi.org/10.1590/1806-90882018000200005

Srivastava S, Avvaru AK, Sowpati DT, Mishra RK (2019) Patterns of microsatellite distribution across eukaryotic genomes. BMC Genom 20(1):1–4. https://doi.org/10.1186/s12864-019-5516-5

Stace HM, Douglas AW, Sampson JF (1998) Did 'paleo-polyploidy' really occur in Proteaceae? Aust J Bot 11(4):613–629. https://doi.org/10.1071/SB98013

Stingemore JA, Nevill PG, Gardner MG, Krauss SL (2013) Development of microsatellite markers for two Australian *Persoonia* (Proteaceae) species using two different techniques. APPS 1(10):1300023. https://doi.org/10.3732/apps.1300023

Sugiura T (1936) Studies on the chromosome numbers in higher plants, with special reference to cytokinesis. I Cytol 7(4):544–595. https://doi.org/10.1508/cytologia.7.544

Taheri S, Abdullah TL, Yusop MR, Hanafi MM et al (2018) Mining and development of novel SSR markers using next generation sequencing (NGS) data in plants. Molecules 23:399. https://doi.org/10.3390/molecules23020399

Tan S, Chen W, Xiang H et al (2021) Screening druggable targets and predicting therapeutic drugs for COVID-19 via integrated bioinformatics analysis. Genes Genom 43:55–67. https://doi.org/10.1007/s13258-020-01021-8

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

Venkata Rao C (1957) Cytotaxonomy of the Proteaceae. Proc Linn Soc NSW 82:257–271

Vieira MLC, Santini L, Diniz AL, Munhoz CDF (2016) Microsatellite markers: what they mean and why they are so useful. Genet Mol Biol 39(3):312–328. https://doi.org/10.1590/1678-4685-GMB-2016-0027

Wang Z, Fang B, Chen J et al (2010) De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*). BMC Genom 11:726. https://doi.org/10.1186/1471-2164-11-726

Wang J, Li Z, Guo Q, Ren G, Wu Y (2011) Genetic variation within and between populations of a desert poplar (*Populus euphratica*) revealed by SSR markers. Ann for Sci 68(6):1143–1149. https://doi.org/10.1007/s13595-011-0119-6

Weston, PH (2007) Proteaceae. In: Flowering plants eudicots. Springer, Berlin, Heidelberg, pp 364–404

World Agroforestry Centre (2002). Agroforestree database. Nairobi, Kenya: ICRAF. http://www.worldagroforestrycentre.org/Sites/TreeDBS/AFT/AFT.htm

Xia Y, Luo W, Yuan S, Zheng Y, Zeng X (2018) Microsatellite development from genome skimming and transcriptome sequencing: comparison of strategies and lessons from frog species. BMC Genom 19:886. https://doi.org/10.1186/s12864-018-5329-y

Xie F, Burklew CE, Yang Y et al (2012) De novo sequencing and a comprehensive analysis of purple sweet potato (*Impomoea batatas* L.) transcriptome. Planta 236:101–113. https://doi.org/10.1007/s00425-012-1591-4

Xu C, Tang S, Fatemi M, Gross CL, Julien MH, Curtis C, Van Klinken RD (2015) Population structure and genetic diversity of invasive *Phyla canescens*: implications for the evolutionary potential. Ecosphere 6(9):1–21. https://doi.org/10.1890/ES14-00374.1

Xu M, Zhu S, Xu R et al (2020) Identification of CELSR2 as a novel prognostic biomarker for hepatocellular carcinoma. BMC Cancer 20:313. 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

Yan F, Xi RM, She RX, Chen PP, Yan YJ, Yang G, Dang M, Yue M, Pei D, Woeste K, Zhao P (2021) Improved de novo chromosome-level genome assembly of the vulnerable walnut tree *Juglans mandshurica* reveals gene family evolution and possible genome basis of resistance to lesion nematode. Mol Ecol Resour. https://doi.org/10.1111/1755-0998.13394

Yang W, Wang K, Zhang J, Ma J, Liu J, Ma T (2017) The draft genome sequence of a desert tree *Populus pruinosa*. Gigascience 6(9):p.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

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 *Gasterophilus nasalis* (Diptera: Oestridae). Entomol Res. https://doi.org/10.1111/1748-5967.12505

Zhou C, He X, Li F, Weng Q, Yu X, Wang Y, Li M, Shi J, Gan S (2014) Development of 240 novel EST-SSRs in *Eucalyptus* L'Hérit. Mol Breed 33(1):221–225. https://doi.org/10.1007/s11032-013-9923-z

Zhu M, Feng P, Ping J et al (2021) Phylogenetic significance of the characteristics of simple sequence repeats at the genus level based on the complete chloroplast genome sequences of Cyatheaceae. Authorea. https://doi.org/10.22541/au.161587690.08363674/v1