De Novo Transcriptome Assembly and Development of EST-SSR Markers for Pterocarpus Santalinus L. f. (Red sanders) , a Threatened and Endemic Timber Tree of India

Sindhu Agastikumar
Institute of Forest Genetics and Tree Breeding
Maheswari Patturaj
Institute of Forest Genetics and Tree Breeding
Aghila Samji
Institute of Forest Genetics and Tree Breeding
Balasubramanian Aiyer
Institute of Forest Genetics and Tree Breeding
Aiswarya Munnusamy
Institute of Forest Genetics and Tree Breeding
Nithishkumar Kannan
Institute of Forest Genetics and Tree Breeding
Vijayakumar Arivazhagan
Institute of Forest Genetics and Tree Breeding
Rekha R Warrier
Institute of Forest Genetics and Tree Breeding
Ramasamy Yasodha (yasodha@icfre.org)
Institute of Forest Genetics and Tree Breeding https://orcid.org/0000-0001-9992-9992

Research Article

Keywords: Chromosome number, Genome size, SSR markers, Pterocarpus, transcription factors

Posted Date: December 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1129482/v1

Version of Record: A version of this preprint was published at Genetic Resources and Crop Evolution on April 14th, 2022. See the published version at https://doi.org/10.1007/s10722-022-01385-8.

Abstract

The endemic and precious timber *Pterocarpus santalinus* L. f. (Red sanders) is a drought hardy tree species for conservation in peninsular India due to its high risk of illegal timber harvest. It is only found in Eastern Ghats of India, and has become threatened owing to overexploitation of its valuable timber. The development of genomic resources, particularly simple sequence repeat (SSR) markers, is essential for strict implementation of in situ conservation measures and application of DNA information based red sanders genetic resource management. However, a lack of genomic data and efficient molecular markers limit the study of its spatial and temporal population genetic structure, identification of diversity hotspots and tree improvement. The current study aims at comprehensive molecular characterization of red sanders and the somatic chromosome counts, flow cytometry and EST-SSR analyses. The results revealed that red sanders is diploid with 2n=20 and the 2C genome size was 0.7872 ± 0.0561pg for the first time in this species. A total of 3128 EST-SSRs were detected based on 25,854 *de novo* assembled unigenes from transcriptome data and primer sets designed for 1953 SSRs. Fifty-nine EST-SSR markers were evaluated for polymorphism in the natural populations of red sanders and 13 were found to be suitable for genetic analysis. Two major transcription factor families bHLH and ERF, responsible for abiotic stress and secondary metabolite synthesis were analysed which would provide the foundation for further research on production of medicinally important biocompounds.

Introduction

Pterocarpus santalinus L. f. (Fabaceae) often known as red sanders, is an endemic species found only in the Eastern Ghats of Andhra Pradesh, India. Geographically it occupies approximately 5160 km^2. Red Sanders is a leguminous tree that grows in tropical dry deciduous forests. In Andhra Pradesh, India and it is limited to the hill ranges of Seshachalam, Lankamala, Veligonda, and Palakonda in the districts of Cuddapah, Kurnool, Chittoor, Nellore and Prakasham (Pullaiah et al. 2019). It has also been documented in isolated locations in the adjoining states of Tamil Nadu and Karnataka, at elevations ranging from 200 to 900 m. In its native habit, red sanders grows in quartzite and shale geological formations (Raju 1999).

P. santalinus is a deciduous tree yielding extremely durable timber with wide variety of uses (Donga et al 2017). The commercially valuable heartwood is red to purplish black in colour with interlocked grains, whereas the sapwood is pearly white in colour with little or no economic value. The heartwood has therapeutic properties and is also used to make carved wood products including musical instruments. The occurrence of wavy grained wood in nature is unusual and unpredictable; wood with such grain pattern is regarded as a unique commodity of high value.

Endemic species are more vulnerable for loss in genetic diversity than other taxa due to their limited geographic range (Coelho et al. 2020). Extreme exploitation and selective logging of mature trees in natural forests had lead to loss of genetic diversity and impairs the genetic structure of the populations. Human-imposed genetic diversity loss of red sanders has been faster in the last 50 years than at any previous period in history (Sarangi 2010). Biologists, ecologists, and biogeographers have been paying close attention to *P. santalinus* conservation (Kukrety et al. 2013; Pullaiah et al. 2019). Studies on endemism patterns provide critical information for conservation efforts, such as identifying hotspots and delineating seed zones.

Identifying priority regions or hotspots where the genetic diversity is most threatened is essential when conservation resources are limited. Conservation units and their sizes are generally based on the concept of dynamic gene conservation which emphasizes the maintenance of evolutionary processes within tree populations to safeguard their potential for continuous adaptation. Such units are continuously managed with essential silvicultural operations, to facilitate genetic processes within the target tree populations. These units become the source of seeds for new plantations.

In such situation, spatially dispersed molecular data that incorporates genetic information of populations become vital, at a scale where regional decisions on conservation activities and investment are made. As population genetic studies have progressed, there has been an increasing appreciation for the value of regional patterns of genetic variants in adaptive and neutral molecular markers in the identification of strong seed zones for genetic conservation and restoration efforts. As DNA based markers have become more widely available, several indigenous tree species have been the subject of genetic studies (Hartvig et al 2018; Castro et al 2021). Researchers can analyse the present influence of drift and gene flow on populations using polymorphic markers inherited from both parents. In many trees with wider distribution, conservation efforts were guided by markers such as simple sequence repeats (Debbabi et al. 2021). Furthermore, assessing genetic structure and diversity combining both adaptive and neutral variations allows researchers to better understand the role of different microevolutionary processes in the formation of genetic diversity hotspots, which has major ecological, evolutionary, and conservation implications (Schueler et al. 2013). Certainly, it would assist in shedding insight on the processes that shaped contemporary genetic diversity hotspots at both the population and species levels. The majority of seed zonation and seed transfer recommendations have been based on ecological classifications, meteorological data, ecophysiological data, and field performance of provenances (Gomory et al.1998).

Despite the great economic importance and scarcity of information for conservation efforts, the use of genetic markers to analyse genetic variation in red sanders is highly restricted. Microsatellites or simple sequence repeat (SSR) markers generated from expressed sequence tags (EST) are increasingly being utilized to evaluate genetic diversity due to their relatively rapid development, cost-effectiveness and are less susceptible to null alleles (Uchiyama et al. 2013). They are developed from expressed regions of the genome with known or suggested functions with the potential to differentiate closely related individuals. Next generation sequencing based RNA-Seq approaches help to analyze the expression and function of genes in different plant species and generate wide variety of SSRs. RNA-Seq using Illumina is rapid, cheaper, and less reliant on existing genomic information than traditional library based techniques. In forest trees high throughput transcriptome sequencing has been successful in discovering SSR markers (Cornacini et al. 2021). These markers have a greater mutation frequency than random mutations, making them an excellent choice for building genomic maps, genotyping, and studying genetic diversity. Although reported to be less polymorphic than genomic SSRs, these markers are superior in functional diversity in relation to adaptive variation and interspecific transferability. Until recently, very few studies on DNA markers have been developed for red sanders (Usha et al. 2013) and, so far, there is no report of SSR markers available for the species. Furthermore, there is no comprehensive publication on red sander genetic diversity based on molecular markers, indicating that this species has not received adequate attention for conservation and improvement. Consequently, applications of molecular markers for seed sources identification, conservation

of genetic resources and genetic improvement are highly inadequate. Hence, this study began with the objective to identify the exact number of chromosomes and genome size which is the foundation for any molecular breeding research in red sanders. The second objective was to generate and verify EST-SSR markers for use in red sanders, and to our knowledge, this is the first study to develop EST-SSRs by deep sequencing of transcriptome in *Pterocarpus* species. The transcriptome data obtained in this study will lay a foundation for further genetic analysis of *P. santalinus* and will be helpful for the discovery and functional annotation of new genes, mapping and marker based breeding.

Materials And Methods

Plant material

The details of plant material used in this study are given in Table 1. Young and fresh leaf sample from one seedling individual of *P. santalinus* was taken for RNA sequencing. To evaluate the polymorphisms of the EST-SSR markers developed from the transcriptome datasets and analyze the population genetic diversity of *P. santalinus*, samples were collected from a total of 42 individuals of *P. santalinus* from eight natural populations (2-11 individuals per population) in Andhra Pradesh, including Chittor East (CE), Puttor (PU), Mamandur North (MN), Mamandur South (MS) Tirupathi (TR), Balapalli (BP) Chammala (CH) and Srikalahasti (SKH) (Table 1). Seedlings from the first five populations were used for chromosome count and DNA content estimation. The field data collections in all the natural distribution areas were undertaken along with the Andhra Pradesh State Forest Department authorities. All the leaf samples were silica dried and stored in room temperature before DNA isolation. In addition, three related species in *Pterocarpus* (*P. marsupium* Roxb., *P. dalbergioides* Roxb. ex DC and *P. indicus* Willd.) were used for testing the transferability of the polymorphic EST-SSR markers developed from *P. santalinus*. In each species three accessions were used for cross amplification purposes.

Chromosome counting

Mitotic chromosome preparations from root tips were performed using seedlings from five different sources of red sanders. Young seedlings (3 months old) were root pruned to induce more adventitious roots and maintained in pots under greenhouse conditions in sand rich medium. The root tips (1–1.5 cm in length) were collected in three different months (April, July and October) in the morning hours (10.30 – 11.30). The sand free root tips were immediately immersed in a saturated solution of α-bromonaphthalene and incubated in darkness at room temperature for 3 hours. The pretreated tips were subsequently fixed in Farmer's fixation solution (Ethanol:Glacial acetic acid, 3:1) overnight at 4 °C. Selected root tips were hydrolyzed in 1N HCl for 13 minutes at 60°C for maceration followed by enzyme treatment. Meristems were dissected using a scalpel blade, and digested in a mixture of 2.5%(v/v) cellulase from *Aspergillus* (Sigma-Aldrich) and 2.5%(v/v) pectinase from *Aspergillus aculeatus* (Sigma-Aldrich) in citrate buffer (pH 4.8) for 20 minutes at room temperature. Enzyme treated roots were then stained using leuco basic fuchsin under the dark condition for 30 minutes. Mitotic chromosome spreading was done by squashing method. Roots dyed with leuco basic fuchsin were immediately placed on a clean slide and added a drop of acetocarmine and squashed under the cover slip by pressing manually with thumb and observed using a light field microscope (Carl Zeiss, Oberkochen, Germany) equipped with AxioCam camera. The cells with well spread chromosomes were

selected for counting followed by recording pencil marking of individual chromosomes. For each plant, at least five different root tips were used (one root tip per slide), and 5–10 metaphase plates were examined.

Flow Cytometry

Approximately 0.5 to 1 mg of young *red* sanders leaf tissue was excised and placed on a sterile petri plate. The tissue soaked in the extraction buffer was chopped gently to release cells. The homogenate was filtered through a sterile nylon membrane filter cup, and the flow collected in a sterile Eppendorf tube. The tube was vortexed and incubated at room temperature for 5 mins in Galbraith's buffer (Galbraith et al. 1983). The sample was then acquired into the flow cytometer (BD FACSVerse, BD Biosciences, USA). *Solanum lycopersicum* L. 'StupickéPolní Rane' and *Pisum sativum* L. (pea) were used as references. The sample rate was adjusted to 40-50 nuclei/second and the DNA G_1 nuclei peaks were positioned by adjusting instrument gain settings. An FSC – SSC dot plot was generated to resolve the events based on size and granularity. A PI area to height plot was used to discriminate the doublets and the PI Area histogram to resolve the frequency distribution of the populations from singlets. The threshold discriminator was set on the PI channel to exclude all PI negative events. Measurements were taken up to 10000 particles. The mean channel number of the G_1 sample peak was determined, and the DNA ploidy of *P. santalinus* was calculated. Three replicates were analysed with each replicate representing a separate individual. 2C value of sample was estimated according to the following formulae:

$$2C\ \text{value of sample} = \frac{\text{Sample}\ \frac{G0}{G1\text{peak}}\text{Mean}}{\text{Sample}\ \frac{G0}{G1\text{peak}}\text{Mean}} \times 2C\ \text{value of standard (pg)}$$

DNA isolation

Genomic DNA was isolated from 42 individuals of *P. santalinus* belong to 8 natural populations for validation of EST-SSR primer pairs designed in this study. Each DNA sample was extracted from 100 mg of silica dried leaf tissue. Leaves were homogenized with liquid nitrogen and DNA isolated using ArborEasy® DNA Isolation Kit (patent product of Institute of Forest Genetics and Tree Breeding, Coimbatore, India) according to the manufacturer's instructions with Proteinase K treatment. The quality of isolated genomic DNA was examined using 0.8% agarose gel electrophoresis, and the concentration was determined using a NanoDrop 8000 spectrophotometer (Nanodrop Technologies, USA). DNA samples were subsequently diluted for the PCR working concentration of 8.0 ng/µL and stored at -20 °C for further use.

RNA isolation, cDNA library preparation and sequencing

Leaf sample was homogenized with liquid nitrogen and total RNA was extracted from each plant using the TRIzol-based RNA extraction kit (Thermo Fischer Scientific, USA). To ensure the accuracy of the data, the purity, concentration, and nucleic acid absorption peaks of the isolated RNA were detected using a Nanodrop spectrophotometer, and the RNA integrity was accurately tested with an Agilent 2100 instrument. Three µg of RNA from three biological replicates were pooled together and the sequencing library was prepared using the NEBNext Ultra RNA Library Prep Kit for Illumina (NEB, Ipswich, MA, USA) according to the manufacturer's

instruction. Size selection was performed according to the manufacturer's protocol with the addition of AMPure XP beads (Beckman Coulter) for obtaining final library size of 400-600bp. After the library quality assessment, cDNA libraries were sequenced on a flow cell using an Illumina HiSeq 2500 system (Illumina, San Diego, CA, USA) at Clevergene Biocorp Pvt. Ltd., Bengaluru, India. Raw data were stored in the sequence read archive (SRA) of NCBI (https://submit.ncbi.nlm.nih.gov/subs/sra/) with the project id PRJNA782228.

Transcriptome assembly generation

Before assembly, the raw reads were filtered to remove poly A/T, low-quality sequences, and empty reads or reads with more than 10% of bases having Q < 30. Raw reads of transcriptome sequences were processed with Cutadapt 1.16.5 (Martin 2011) and Trimmomatic 0.36.5 (Bolger et al. 2014) tools to remove the adapter sequences and to trim low quality bases, respectively. The assembly of a de novo transcriptome using clean reads was performed using the short-read assembling program Trinity 2.9.1 software with parameters setting of minimum contig length of 200 bp (Haas et al. 2013). Trinity super transcript (trinity gene splice modeler) was used to cluster transcripts to obtain final unigenes. In order to validate the completeness of the reconstructed transcriptome analysis was performed with BUSCO v 5.0.0 software (Simao et al. 2015).

Functional annotation of unigenes

The open reading frames (ORFs) and coding sequences (CDS) were predicted from the unigenes using TransDecoder v 5.5.0 program (Haas et al. 2013). Unigenes and ORFs were annotated using BLASTX and BLASTP alignment tools (BLAST+ v 2.12.0 software) with the e- value < 10^{-5} against NCBI non- redundant protein (NR), UniProtKB, Gene Ontology (GO) and KEGG Orthology (KO) databases. The orthologous groups and protein function annotation were identified by WebMGA online server using COG, KOG and Pfam databases with default parameter (e < 10^{-3}) (http://weizhong-lab.ucsd.edu/webMGA/server/).

Identification of single copy orthologs (SCOs) and molecular phylogeny

Phylogenetic relationship of *P. santalinus* was assessed with 7 Dalbergieae tribe species (nearest relatives), 4 Fabaceae members and 3 *Populus* species as out groups were chosen. The transcriptome sequences of *Dalbergia* species (*Dalbergia cochinchinensis* Pierre ex Laness - GIHU01, *D. frutescens* (Vell.) Britton - GIHP01, *D. melanoxylon* Guill. & Perr. - GIHQ01, *D. Miscolobium* - GIHR01, *D. oliveri* Gamble - GIHS01 and *D. sissoo* Roxb. - GIHT01) and three *Populus* species (*Populus adenopoda* Maxim - GIKX01, *P. gradidentata* Michx. - GIKW01 and *P. tomentosa* Carr. - GHTT01) were retrieved from NCBI transcriptome shotgun assembly sequence database (https://www.ncbi.nlm.nih.gov/genbank/tsa/). The trinity assembled transcriptome sequences of *Enterolobium cyclocarpum* (Jacq.) Griseb. (ERS399691), *Faidherbia albida* (Delile) A. Chev. (ERS399692), *Sesbania sesban* (L.) Merr. (ERS399703), *S. macrantha* Welw. ex E.Phillips & Hutch. (ERS399702) and *Tipuana tipu* (Benth.) Kuntze (ERS399705) were retrieved from tropiTree online portal (https://ics.hutton.ac.uk/tropiTree/) (Russell et al. 2014). The longest ORF sequences of all transcriptomes were retrieved using TransDecoder v 5.5.0 program (Haas et al. 2013). Orthofinder v 2.5.4 (Emms and Kelly 2015) tool was used to identify the SCO sequences. Each SCO groups were aligned using MUSCLE v5 program (Edgar 2004). The aligned sequences were concatenated into single alignment files and the phylogenetic tree was constructed using MEGA v 11.0 following maximum likelihood (ML) method with 1000

bootstrap replications, Jones-Taylor-Thornton (JTT) model (Tamura et al 2021). A single heuristic search was performed by Nearest Neighbor interchange (NNI) method.

Molecular classification of bHLH and ERF family transcription factors

The protein coding longest ORF sequences from the transcriptome sequences of *P. santalinus* and *T. tipu* were BLAST against *Arabidopsis thaliana* using the plant transcription factor database (http://planttfdb.gao-lab.org/). The protein sequences of bHLH, ERF transcription factor families were screened for phylogenetic analysis. The conserved domains and motifs of screened transcription factors were identified using MEME Suite v. 5.4.1 software (https://meme-suite.org/meme/) (Bailey et al., 2006) and NCBI Batch Web CD- search tool (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) with default parameters. Full length amino acid sequences were aligned by MAFFT v 7.0 program using g-ins-i algorithm with BLOSUM 62 scoring matrix (Katoh and Standley 2013). The molecular phylogeny tree was constructed using the IQ TREE v 2.1.2 program (Minh et al. 2020) using default parameters with 1000 bootstrap replications. The groups/ clades of bHLH and ERF family transcription factors were named according to the classification followed in *A. thaliana* (Heim et al. 2003; Nakano et al. 2006). All sequences were renamed and listed in Supplementary File 1.

Detection of EST-SSR markers and designing of primers

EST-SSR markers were detected in the assembled unigenes using the MicroSAtellite v.2.1 software (http://pgrc.ipk-gatersleben.de/misa/). Parameters were set with a minimum number of 7, 5, 5, 4, and 4 repeat units for identification of di-, tri-, tetra-, penta-, and hexa-nucleotide motifs, respectively and no mono-nucleotide repeats were considered. The maximum number of bases for two SSRs in an interrupted compound microsatellite was 100. EST-SSRs were detected and mined among the unigenes with length >1000 bp.

Unigenes with a sequence of more than 150 bp before and after the SSR region with known putative functions were used for primer design by Primer v3.0 (https://bioinfo.ut.ee/primer3-0.4.0/). Primer designing criteria consisted of primer length 18–23 bp, with optimum value 20 bp; T_m 57–63°C, with optimum value 60°C; GC content 40–60%, with the optimum value 50%; maximum T_m difference between forward and reverse primer 1.5°C and product size range 100 - 300 bp optimum value 150 bp. Validation of these SSRs was carried out by synthesising 59 primer sets selected randomly.

PCR amplification and validation of EST-SSRs

The PCR amplification was performed in a 10 µL reaction volume, which included 1.5µL of template DNA (8 ng/µL), 0.5µL of the forward primer (10 µmol/L), 0.5µL of the reverse primer (10 µmol/L), 5.0µL of AmpliTaqGold™ 360 master mix, and 2.5 µL of double distilled water. PCR amplification was performed in a Veriti 96-well Thermal Cycler (Applied Biosystems™, USA) using the following temperature profile: pre-denaturation at 95 °C for 10 min; 35 cycles of denaturation at 95 °C for 45 s, the appropriate annealing temperature for 35 s, extension at 72 °C for 55 s; and a final extension at 72 °C for 7 min; and preservation at 4°C. PCR products were resolved on vertical gel electrophoresis system (Cleaver scientific Ltd, UK) using 7%

non-denaturing polyacrylamide gel with 1x TAE buffer and then detected by silver staining (Huang et al. 2018).

The size and number of alleles present per marker was scored by AlphaEaseFC software (AlphaInnotech, San Leandro, CA, USA). Total number of polymorphic alleles, polymorphic information content (PIC) and heterozygosity (He) were calculated by Power Marker v 3.25 (Liu and Muse 2005).

Results And Discussion

Chromosome counting and nuclear DNA content analysis

The chromosome number of *P. santalinus* ascertained in this study was 2*n* = 20 (Fig. 1). However, the chromosome count observed in this study is different from earlier report (Bhaskar, 1981). Previous study on chromosome counts in *P. santalinus* showed 2*n* = 22 after observing seedlings obtained from 2 individuals trees growing in Mysore, which is a non native location. Recently Moraes et al. (2020) studied the chromosome number evolution in *Pterocarpus* clade and reported somatic chromosome count of 20 and 40 for species such as *Centrolobium tomentosum* Guillemin ex Benth. (2*n* = 20), *Stylosanthes hamata* (L.)Taub. (2*n* = 20), *S. scabra* Vog. cv. Seca (2*n* = 40), *S. seabrana* B.L.Maass & 't Mannetje (2*n* = 20) and *S. viscose* Swartz (2*n* = 20). Somatic chromosome counts (parsed n) for different *Pterocarpus* species showed that it varied from 10, 11, 12, 17 and 22 (Rice et al. 2015; Supplementary File 1) indicating the possibilities of polyploidy and dysploidy. Nevertheless the species *T. tipu*, a sister species belonging to Pterocarpus clade was reported to have 2n=20 (Coleman and Menezes, 1980). Chromosome changes within species (dysploidy) were reported in dalbergioid legumes, which includes Pterocarpus clade and highlighted the importance of studying multiple populations of the species (Moraes et al. 2020). Accordingly, in this study after assessing five different populations, it was substantiated that the somatic chromosome number in *P. santalinus* was 20. Many publications suggest that *T. tipu* is phylogenically closely linked to the genus *Pterocarpus*, therefore such investigations may support these findings (Lavin et al. 2005; Klitgard et al. 2013; Hong et al. 2020).

The 2C nuclear DNA content of *P. santalinus* was estimated as 0.7872 ± 0.0561 pg, where the internal standards *Solanum lycopersicum L.* and *Pisum sativum* L. DNA content was 2.0 and 9.8 pg, respectively using flow cytometry (Fig. 2). The C-value (gametic DNA content) and the Cx value (genome content) were calculated based on the estimated 2C nuclear DNA content of the studied plant and were expressed in terms of Mbp. Being diploid, the C-value and the 1Cx value were the same and half the 2C nuclear DNA content, i.e., 0.3936 pg or 393.6 Mbp. Thus, the 2C Nuclear DNA content is 0.7872 pg or 787.2 Mbp. In dalbergias, the estimated genome size ranged from 1.43–1.98 Gb (Hung et al. 2020), however no information available for *Pterocarpus* species. The genome size of *P. santalinus* has been described for the first time, which will serve as the foundation for whole genome sequencing and tree improvement.

Transcriptome assembly

Transcriptome research is one of the most essential tools for studying species biological processes. In the current study, using RNA-seq technology on the Illumina HiSeq 2500 platform, leaf transcriptome of *P.*

santalinus was characterized with the objective of identifying EST-SSRs. Despite its economic importance (Pullaiah et al. 2019), there is a genetic information gap in the genus *Pterocarpus*, and no SSR markers are available. Totally 17,930,663 raw reads were generated and after the removal of adapters and sequences with poor quality, 16,182,076 sequences in the range of 20 to 151 bp were retained. *De novo* assembly produced 29,816 transcripts and 25,854 unigenes were obtained with total bases of 20,943,459 bp having mean transcript length of 810 bp with the N50 contig length of 1126 bp. Among the assembled transcripts about 90 % had >500bp length (Table 2). Read mapping of the raw reads showed that 90.2% were aligned by Trinity 2.9.1 software. These results corroborated with the earlier findings on species such as *Camelina sativa* L. (Liang et al. 2013), *Myrciaria dubia* Kunth McVaugh (Castro et al. 2020) and *Ulmus wallichiana* Planch. (Singh et al. 2021). The use of PacBio SMRT sequencing technology for transcriptome sequencing resulted in high-quality transcripts longer than 1000 bp in many plant species, allowing third-generation technologies to be introduced for species with little genomic information (Wu et al. 2020)

Functional annotation and classification

The 25,854 super transcripts were functionally annotated using NR, UniProtKB, KEGG, KOG databases (Supplementary Table 1). These annotated sequences provide a platform for further research into *P. santalinus* genetic diversity. By comparing the transcript sequence to UniProtKB with homologous species, most of the transcripts (75.8%) had best matches with *Medicago truncatula* Gaertn., a legume species and the second top-hit (9.3%) species was *Actinidia chinensis var.chinensis* Planch. (Supplementary File 1) thus reflecting the scarcity of reports on the transcriptome of related *Pterocarpus* species. Among 25,854 unigenes 93.25% were assigned a specific or general function gene ontology terms, 63.26% of transcripts involved in molecular function, 56.39% as cellular components and 45.79% involved in biological process (Fig. 3a). The results of functional annotation in red sanders clearly indicated that the proportion of unigenes annotated in this study is higher than transcriptome data deficient plant species such as *Panax vietnamensis* Ha et Grushv. (Vu et al. 2020). Among all the unigenes involved in molecular function 6056 unigenes were responsible for catalytic activity, 5187 unigenes were responsible for protein binding and 4860 unigenes involved in response to stimulus. Among cellular components 8267 were involved in membrane bound organelles, 4445 were organelle parts and 3135 involved in plastids. KEGG annotation showed 1899 unigenes involved in the metabolism, 1921 unigenes in genetic information processing, 421 unigenes in signalling & cellular processing and 492 in environmental information processing (Fig 3b; Supplementary Table 1). Function annotation of the non-redundant unigenes was conducted by searching against the main databases. A total of 93.25% (24,110) of the non-redundant unigenes were annotated in UniProtKB, 20,608 unigenes in GO, 18,476 unigenes in Pfam, 8,754 unigenes in KEGG and 8,152 unigenes in KOG. Further, functional classification of *P. santalinus* was carried out with search against the Clusters of COG database, and 5163 unigenes allocated to 25 classes (Supplementary File 1). The highest category was general function prediction (702, 13.6%), followed by amino acid transport and metabolism (270, 5.2%), and energy production and conversion (258, 5.0%). In KOG terms, the three major represented predictions were 233 transcripts to serine/threonine protein kinase family (KOG1187), 53 to E3 ubiquitin ligase family (KOG4628) and 47 to UDP-glucuronosyl and UDP-glucosyl transferase (KOG1192) (Supplementary Table 1). KEGG annotation provided 209 categories in which metabolism were represented maximum (Supplementary File 1). The major pathways include biosynthesis of secondary metabolites, biosynthesis of amino acids, inositol

phosphate metabolism, oxidative phosphorylation, RNA transport, ubiquitin mediated proteolysis, sulphur relay system, MAPK signalling pathway etc. These data reveals the active metabolic processes as well as synthesis of diverse metabolites in red sanders. In *P. santalinus* the wood has high value phytochemicals like alkaloids, phenols, saponins, glycosides, flavonoids, triterpenoids, sterols and tannins (Pullaiah et al. 2019) and these compounds help in adaptation for environmental stresses, such as drought. In the current study, we have recorded majority of unigenes for biosynthesis of secondary metabolites could enhance the utility value of red sanders.

Identification of single copy orthologs and phylogenetic tree construction

In many of the plant species, due to the non-availability of genomic resources, single copy protein coding genes are used as efficient markers for phylogenetic analysis (Li et al. 2017). In this study, all possible SCOs were identified for phylogenetic tree construction to assess the relationship between *P. santalinus* and 12 other Fabaceae members by combining the publicly available transcriptomes. Three *Populus* species were included as outgroups. The Fabaceae members included 7 Dalbergieae tribe species (6 species belonging to *Dalbergia* genus and *T. tipu*, a Pterocarpus clade species) and *Sesbania* (2 species), *Enterolobium cyclocarpum* (Jacq.) Griseb. and *Faidherbia albida.* Orthofinder identified 56,631 orthogroups from 15 species targeted for analysis. Within species group, an orthogroup is described as a collection of genes derived from a single gene of the last common ancestor (Emms and Kelly, 2015). The largest orthogroup contained 8419 genes (O50= 8419) and 2683 orthogroups shared genes of all the 15 species. However, complete single copy genes were identified from 19 orthogroups only and were concatenated to generate the phylogenetic tree. The phylogenetic relationship of *P. santalinus* and the other plant species is shown in Fig. 4. In the orthologues phylogram, a clear separation between the outgroup *Populus* and Fabaceae was observed. As expected, the *P. santalinus* was placed within Pterocarpus clade along with *T. tipu* and confirmed the recent report on their status as sister taxa (Hong et al. 2020). Further all the members of *Dalbergia* genus grouped together followed by two *Sesbania* species and two additional Fabaceae members. While studying the Dalbergieae, it was suggested that Pterocarpus clade needs thorough phylogenetic analysis for their polyphyletic origin of species (Cordoso et al. 2013).

Classification of transcriptional factors

From the sequences of *P. santalinus* and *T. tipu* totally 606 and 486 transcription factor (TF) families were identified, respectively. Among these bHLH and ERF TFs were predominant with 105 and 77 respectively. The phylogenetic tree of bHLH family and ERF family transcription factors were represented in Fig. 5a and Fig. 5b. The phylogeny trees of bHLH and ERF were subdivided into 18 and 15 subgroups each (Supplementary File 1) respectively, and majority of the groups had both the taxa. Generally the bHLH family proteins in plants are grouped into 15 to 32 classes and increase in genome sequencing projects identify novel protein sequences. Overwhelming evidences available to show that the bHLH genes have high relevance to abiotic stress tolerance and secondary metabolite biosynthesis including anthocyanins (Goossens et al. 2017; Guo et al. 2021). Similarly, the ERF transcription factors are another major group serve as important regulators in biotic and/or abiotic stress responses, disease resistance, plant hormone signal transduction, and metabolite regulation (Xie et al. 2019). Red sanders being the store house of plethora of phytochemicals with bioactive potential and have applications in pharmaceutical, ayurveda, cosmetics, liquor, and textile industries (Bulle et

al. 2016; Pullaiah et al. 2019). In-depth research on biosynthesis of key metabolites through RNA-Seq analysis would bring in more information on production of industrially important derivatives. The information generated in this study on transcription factor families would have intriguing avenues for translational research in *P. santalinus,* specifically in the areas of regulations of specialized metabolism.

Identification of EST-SSRs from *P. santalinus* transcriptome

SSRs are frequently used in genetic analysis of plant species, and EST-SSRs are functional molecular markers with the benefits of easier and more efficient production, and more interspecific transferability when compared to genomic SSRs (Wu et al. 2020). EST-SSRs can be used to map the gene rich regions of chromosomes and obtain gene expression data because of their close link to functional genes. In this study, all 25,854 unigenes were employed to mine potential EST-SSRs in *P. santalinus* and 3128 SSRs have been identified. Among these, a total of 3564 (13.8%) unigene sequences were found to encode 1953 potential EST-SSRs which had sufficient flanking regions to design the primer sets for PCR amplification (Supplementary Table 2). Of these, 746 unigenes contained more than one SSR and 344 SSRs were present in compound formation.

There were 156 motif sequence types detected in 3128 identified SSRs having di, tri, tetra, penta, and hexa-nucleotide repeats with 3, 10, 12, 30 and 101 types respectively. Tri-nucleotide repeat motifs were the most abundant (1885, 60.26%) followed by di-nucleotide repeat motifs (818, 26.15%) and hexa-nucleotide repeat motifs (258, 8.25%), whereas penta- (105, 3.36%) and tetra-nucleotide repeat motifs (62, 1.98%) were found to be rare (Table 2). Abundance of tri-nucleotide repeats is common among the transcriptome of Legume members (Roorkiwal and Sharma, 2011; Wang et al. 2018), and many other species. Among the repeat types, AG/CT, AAG/CTT, AAC/GTT, ATC/ATG were predominant motifs and similar results were reported in most other Fabaceae plant species (Wang et al. 2014; Huang et al. 2016; Liu et al. 2019),

Polymorphic validation of EST-SSRs in *P. santalinus* and transferability across related species

All the 59 primers (Supplementary Table 3) were screened with four randomly selected individuals of *P. santalinus* and selected 13 primer pairs (Table 3; Supplementary File 1) which amplified clearly and generated polymorphic alleles among the four selected individuals. All the 13 polymorphic primer pairs were used to analyze the 42 individuals from eight different geographic locations. Allele size was scored with AlphaEaseFC software (Alpha Innotech, USA). The 13 EST- SSR markers produced 151 polymorphic alleles. The samples from Chamala population had highest percentage of polymorphism (84.6) and the lowest was from Balapalli population (46.2) (Table 3). PIC value ranged from 0.7937 to 0.9185 with a mean value of 0.8639, genetic diversity ranged from 0.7792 to 0.9235 with a mean value of 0.8763 and heterozygosity ranged from 0.0714 to 0.5 with a mean value of 0.3223. The EST-SSR markers developed in this study are proven to be useful for genetic variability assessment of red sanders populations.

Cross-species transferability of 13 polymorphic SSR markers were examined in *P. marsupium*, *P. dalbergioides* and *P. indicus* and it was observed that all the 13 markers produced clear, sharp bands of expected product size, indicating the conserved sequences across species. Such cross-species transferability is a common phenomenon in legume plants (Datta et al. 2013; Singh et al. 2020) which increase the

repository of SSR markers for related species where minimum genomic information available. Many species of the genus *Pterocarpus* have international significance and listed under the IUCN Red List of Threatened Species. These novel co-dominant markers would find their applications in gene flow and population genetic structure of *P. santalinus*, for identifying genetic diversity hotspots, genecological zones and development of suitable conservation guidelines. Using number SSR markers and accessions from entire natural distribution area, genetic diversity and population structure are being studied to obtain better knowledge about the structure and evolution of *P. santalinus*.

Conclusions

To the best of our knowledge, this is the first comprehensive report on confirmation of somatic chromosome number, estimation of genome size, generation of transcriptome data and development of EST-SSR markers in *P. santalinus*. This study provided an important genetic resource which would shed light on the role of major transcription factor families such as bHLH and ERF for elucidating their role in various secondary metabolite biosynthetic pathways in *P. santalinus* especially for those with medicinal value and allow for drug development. Additionally red sanders being an illegally traded commodity timber, the EST-SSR markers have become valuable resource for timber tracking through fingerprinting.

Declarations

Funding

The financial support received from National Biodiversity Authority, Government of India (No. Tech/Genl/22/149/17/18-19/290) is gratefully acknowledged.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

Chromosome studies, conduct of PCR experiments, data analysis and, drafting the article were performed by Sindhu Agasthikumar, Aghila Samji, Nithishkumar Kannan and Vijayakumar Arivazhagan; Genomic DNA isolation for the samples was carried out by Aiswarya Munusamy; Conduct of RNA isolation and bioinformatic analysis was done by Maheswari Patturaj; Field collection of samples and seeding development was carried out by Balasubramanian Aiyer; Genome size estimation was performed by Rekha R Warrier; Conceived the concept, designed the analysis and critical revision of the article was undertaken by Yasodha Ramasamy. All authors read and approved the final version of the manuscript.

Acknowledgements

The authors acknowledge the funding support from the National Biodiversity Authority, Government of India. Field support for collection of leaf and seed samples by the Andhra Pradesh Forest Department is gratefully acknowledged.

Data Availability

The datasets generated during the current study are available in the sequence read archive (SRA) of NCBI (https://submit.ncbi.nlm.nih.gov/subs/sra/) with the ID PRJNA782228.

References

1. Bailey T, Williams N, Misleh C, Li W (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34:369-373. https://doi.org/10.1093/nar/gkl198
2. Bhaskar V (1981) The chromosome number of *Pterocarpus santalinus* L.f. (red sanders). Ind J Forestry 4(4): 335
3. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120. https://doi.org/10.1093/bioinformatics/btu170
4. Bulle S, Reddy VD, Hebbani AV, Padmavathi P, Challa C, Puvvada PK, Repalle E, Nayakanti D, Narasimhulu CA, Nallanchakravarthula V (2016) Nephro-protective action of *P. santalinus* against alcohol-induced biochemical alterations and oxidative damage in rats. Biomed Pharmacother 84:740–746
5. Cardoso D, Pennington RT, de Queiroz LP, Boatwright JS, Van Wyk BE, Wojciechowski MF, Lavin M (2013) Reconstructing the deep-branching relationships of the papilionoid legumes. S Afr J Bot 89:58–75
6. Castro DAM, Costa TS, Cardoso AS, Ramos HCC, López JA, Diniz LEC (2021) Genetic structure analysis of *Mauritia flexuosa* natural population from the Lençóis Maranhenses region using microsatellite markers. Scientia Agricola 79(1). https://doi.org/10.1590/1678-992X-2020-0112
7. Castro JC, Maddox JD, Rodríguez HN, Castro CG, Imán-Correa SA, Cobos M, Adrianzén PM (2020) Dataset of de novo assembly and functional annotation of the transcriptome during germination and initial growth of seedlings of *Myrciaria dubia* "camu-camu". Data in Brief 31: 105834
8. Coelho N, Gonçalves S, Romano A (2020) Endemic plant species conservation: Biotechnological approaches. Plants 9(3):345
9. Coleman JR, DeMenezes EM, (1980) Chromosome numbers in *Leguminosae* from the state of São Paulo Brazil. Rhodora 82(831):475-481
10. Cornacini MR, Manoel RO, Alcantara MAM, Moraes ML, Silva EA, Neto LGP, Sebbenn AM, Rossini BC, Marino CL (2021) Detection and application of novel SSR markers from transcriptome data for *Astronium fraxinifolium* Schott, a threatened Brazilian tree species. Mol Biol Rep 48(4): 3165-3172. https://doi.org/10.1007/s11033-021-06338-5
11. Datta S, Mahfooz S, Sahil P, Choudhary AK, Chaturvedi SK, Nadarajan N (2013) Conservation of microsatellite regions across legume genera increases marker repertoire in pigeonpea. Aust J Crop Sci 7(13):1990-1997
12. Debbabi SO, Mnasri SR, Bhasker, Amar FB, Naceur B, Montemurro C, Miazzi MM (2021) Applications of Microsatellite Markers for the Characterization of Olive Genetic Resources of Tunisia. Genes 12(2): 286
13. Donga S, Moteriya P, Pande J, Chanda S (2017) Development of quality control parameters for the standardization of *Pterocarpus santalinus* L. f. leaf and stem. J Pharmacogn Phytochem 6(4):242–252

14. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797
15. Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157. https://doi.org/10.1186/s13059-015-0721-2
16. Galbraith DW, Harkins KR, Maddox JM, Ayres NM, Sharma DP, Firoozabady E (1983) Rapid flow cytometric analysis of the cell cycle in intact plant tissues. Science 220(4601): 1049-1051
17. Gömöry D, Hynek V, Paule L (1998) Delineation of seed zones for European beech (*Fagus sylvatica* L.) in the Czech Republic based on isozyme gene markers. Annales des sciences forestières. 55(4): 425-436
18. Goossens J, Mertens J,Goossens A (2017) Role and functioning of bHLH transcription factors in jasmonate signalling. J Exp Bot 68(6):1333-1347
19. Guo J, Sun B, He H, Zhang Y, Tian H, Wang B (2021) Current Understanding of bHLH Transcription Factors in Plant Abiotic Stress Tolerance. Int J Mol Sci 22(9):4921
20. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494-512
21. Heim MA, Jakoby M, Werber M, Martin C, Weisshaar B, Bailey PC (2003) The basic helix–loop–helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. Mol Biol Evol 20(5):735-747
22. Hartvig I, So T, Changtragoon S, Tran HT, Bouamanivong S, Theilade I, Kjær ED, Nielsen LR (2018) Population genetic structure of the endemic rosewoods *Dalbergia cochinchinensis* and *D. oliveri* at a regional scale reflects the Indochinese landscape and life history traits. Ecol Evol 8(1):530-545
23. Hong Z, Wu Z, Zhao K, Yang Z, Zhang N, Guo J, Tembrock LR, Xu D (2020) Comparative analyses of five complete chloroplast genomes from the genus *Pterocarpus* (Fabacaeae). Int J Mol Sci 21(11):3758
24. Huang J, Guo X, Hao X, Zhang W, Chen S, Huang R, Gresshoff PM, Zheng Y (2016) *De novo* sequencing and characterization of seed transcriptome of the tree legume *Millettia pinnata* for gene discovery and SSR marker development. Mol Breed 36:1-15
25. Huang L, Deng X, Li R, Xia Y, Bai G, Siddique KH, Guo P (2018) A fast silver staining protocol enabling simple and efficient detection of SSR markers using a non-denaturing polyacrylamide gel. J Vis Exp 134:e57192
26. Hung TH, So T, Sreng S, Thammavong B, Boounithiphonh C, Boshier DH, MacKay JJ (2020) Reference transcriptomes and comparative analyses of six species in the threatened rosewood genus *Dalbergia*. Sci Rep 10(1):1-14
27. Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol (30):772-780
28. Klitgård B, Forest F, Booth TJ, Saslis-Lagoudakis CH (2013) A detailed investigation of the *Pterocarpus* clade (*Leguminosae: Dalbergieae*): *Etaballia* with radially symmetrical flowers is nested within the papilionoid-flowered *Pterocarpus*.S. Afr J Bot vol 89:128-142

29. Kukrety S, Jose S, Alavalapati JRR (2013) Exploring stakeholders' perceptions with analytic hierarchy process: A case study of red sanders (*Pterocarpus santalinus* L.) restoration in India. Restor Ecol 21(6):777-784
30. Lavin M, Herendeen PS, Wojciechowski MF (2005) Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. SystBiol 54:575-594
31. Li Z, De La Torre AR, Sterck L, Cánovas FM, Avila C, Merino I (2017) Single copy genes as molecular markers for phylogenomic studies in seed plants. Genome Biol Evol 9:1130-47
32. Liang C, Liu X, Yiu SM, Lim BL (2013) *De novo* assembly and characterization of *Camelina sativa* transcriptome by paired end sequencing. BMC Genomics 14:146
33. Liu FM, Hong Z, Yang ZJ, Zhang NN, Liu XJ, Xu DP (2019) De Novo transcriptome analysis of *Dalbergia odorifera* T. Chen (Fabaceae) and transferability of SSR markers developed from the transcriptome. Forests 10:1-16
34. Liu K, Muse SV (2005) Power Marker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21(9):2128-2129
35. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal 17:10-12
36. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Haeseler A, Lanfear R (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530-1534
37. Moraes AP, Vatanparast M, Polido C (2020) Chromosome number evolution in dalbergioid legumes (Papilionoideae, Leguminosae). Braz J Bot 43:575-587
38. Nakano T, Suzuki K, Fujimura T, Shinshi H (2006) Genome-wide analysis of the ERF gene family in *Arabidopsis* and rice. Plant Physiol 140(2):411-432
39. Pullaiah T, Balasubramanya S, Anuradha M (2019) *Red Sanders: Silviculture and Conservation*. Springer Singapore
40. Raju K, Nagaraju A (1999) Geobotany of red sanders (*Pterocarpus santalinus*) – a case study from the south-eastern portion of Andhra Pradesh. Environ Geol 37:340-344
41. Rice A, Glick L, Abadi S, Einhorn M, Kopelman NM, Salman-Minkov A, Mayzel J, Chay O, Mayrose I (2015) The Chromosome Counts Database (CCDB) — a community resource of plant chromosome numbers. New Phytol 206(1):19-26. https://doi: 10.1111/nph.13191
42. Roorkiwal M, Sharma PC (2011) Mining functional microsatellites in legume unigenes. Bioinformation 7(5):264-270
43. Russell JR, Hedley PE, Cardle L, Dancey S, Morris J, Booth A, Odee D, Mwaura L, Omondi W, Angaine P, Machua J, Muchugi A, Milne I, Dawson IK (2014) Tropitree: an NGS-based EST–SSR resource for 24 tropical tree species. PLoS One 9(7):e102502
44. Sarangi PK (2010) Conservation, production and marketing of Red Sanders, *Pterocarpus santalinus* L f. Ind For 136(5):569-579
45. Schueler S, Kapeller S, Konrad H, Geburek T, Mengl M, Bozzano M, Koskela J, Lefevre F, Hubert J, Kraigher H, Longauer R (2013) Adaptive genetic diversity of trees for forest conservation in a future

climate: a case study on Norway spruce in Austria. Biodivers Conserv 22:1151-1166

46. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210-3212
47. Singh A, Majeed A, Bhardwaj P (2021) Transcriptome characterization and generation of marker resource for Himalayan vulnerable species, *Ulmus wallichiana*. Mol Biol Rep 48:721-729
48. Singh D, Singh CK, Tribuvan KU, Tyagi P, Taunk J, Tomar RSS, Kumari S, Tripathi K, Kumar A, Gaikwad K, Yadav RK (2020) Development, characterization, and cross species/genera transferability of novel EST-SSR markers in lentil, with their molecular applications. Plant Mol Biol Rep 38(1):114-129
49. Tamura K, Stecher G, Kumar, S (2021) MEGA11: molecular evolutionary genetics analysis version 11. Mol Biol Evol 38(7):3022-3027
50. Uchiyama K, Fujii S, Ishizuka W, Goto S, Tsumura Y (2013) Development of 32 EST-SSR markers for *Abies firma* (Pinaceae) and their transferability to related species. Appl Pl Sci 1(2):1200464
51. Usha R, Rani SJ, Prasuna TG (2013) Genetic relationship between quality and non quality wood of *Pterocarpus santalinus* L. (red sanders), an endemic tree species by using molecular markers. J Chem Pharm Sci 6(3):189–194
52. Vu DD, Shah SNM, Pham MP, Nguyen MT, Nguyen TPT (2020) *De novo* assembly and transcriptome characterization of an endemic species of Vietnam *Panax vietnamensis* Ha et Grushv. including the development of EST-SSR markers for population genetics. BMC Plant Biol 20:358
53. Wang H, Lei Y, Yan L, Wan L, Cai Y, Yang Z, Lv J, Zhang X, Xu C, Liao B (2018) Development and validation of simple sequence repeat markers from *Arachis hypogaea* transcript sequences. Crop J 6:172-180
54. Wang Z, Yu G, Shi B, Wang X, Qiang H, Gao H (2014) Development and characterization of simple sequence repeat (SSR) markers based on RNA-sequencing of *Medicago sativa* and in silico mapping onto the *M. truncatula* genome. PLoS ONE 9(3):e92029
55. Wu Q, Zang F, Xie X, Ma Y, Zheng Y, Zang D (2020) Full-length transcriptome sequencing analysis and development of EST-SSR markers for the endangered species *Populus wulianensis*. Sci Rep 10:1-11
56. Xie Z, Nolan TM, Jiang H, Yin Y (2019) AP2/ERF transcription factor regulatory networks in hormone and abiotic stress responses in *Arabidopsis*. Frontiers in plant science 10:228

Tables

Table 1 Details of *P. santalinus* plant materials used in this study

S. No.	Name of the population	Population Code	No of Individuals	Latitude (°N)	Longitude (°E)	Altitude (m)	Rainfall (mm)
1	Chitoor East	CE	4	13˚ 12′ 41.8"	079˚ 07′ 59.7"	357	809
2	Puttur	PU	2	13˚ 30′ 38.8"	079˚ 32′ 02.0"	221	920
3	Srikalahasti	SKH	4	13˚ 30′ 23.1"	079˚ 40′ 40.0"	111	900
4	Mamandur North	MN	4	13˚ 45′ 42.1"	079˚ 25′ 21.4"	214	890
5	Mamandur South	MS	4	13˚ 43′ 20.9"	079˚ 27′ 12.0"	192	890
6	Chamala	CH	11	13˚ 48′ 48.0"	079˚ 12′ 19.2"	626	850
7	Tirupathi	TR	9	13˚ 43′ 05.3"	079˚ 19′ 20.4"	964	890
8	Balapalli	BP	4	13˚ 44′ 12.1"	079˚ 18′ 44.0"	953	900

Table 2 Summary of *P. santalinus de novo* transcriptome assembly, EST-SSRs and functional analysis

Data Type	Category	Count
Raw data	Total Sequences	17930663
	GC%	46
Cleaned data	Total Sequences	16182076
	GC%	45
Unigene data	Total Sequences	25854
	Total size (bp)	20943459
	Average size (bp)	810.1
	N50 (bp)	1126
EST SSR data	Total number of identified SSRs	4675
	Number of SSR containing sequences	3564
	Number of sequences containing more than one SSR	746
	Number of SSR present in compound formation	398
SSR Category	Di- nucleotide repeat containing SSR	818 (26.15%)
	Tri- nucleotide containing SSR	1885 (60.26%)
	Tetra- nucleotide containing SSR	62 (1.98%)
	Penta- nucleotide containing SSR	105 (3.36%)
	Hexa- nucleotide containing SSR	258 (8.25%)
Meta- annotation data	Full length	4219
	Quasi full length	9493
	Partial	10204
	No information	1938
Functional annotation data	SwissProt	24110
	GO	20608
	Pfam	18476
	KEGG	8754
	KOG	8152
	COG	5163
	All annotated databases	24972

Table 3 Details of *P.santalinus* polymorphic primers identified in the study

S. No	Marker	Primer	Tm (°C)	Repeat type	Product Size (bp)	N	He	PIC
1	IFGTB1507	F-TCCGTAACCCTATTGTCTCT	54.97	(TCCCAA)4	146	8	0.5	0.8
		R-GTATGCAGTGAAGGTGGATT	55.42					
2	IFGTB4100	F-GACGATTTCTATGGGTTGAG	54.71	(GGT)7	180	20	0.8	0.9
		R-CACCATCACCTACCTTCATT	54.9					
3	IFGTB442	F-GTCATCGGAACCACTAACAT	54.89	(TGGGCC)4	237	17	0.5	0.9
		R-AGTAACAGAGAGCGAGTTGG	54.76					
4	IFGTB682	F-GCTCAAGATTTACCTCCAGA	54.57	(TGC)6	242	8	0.1	0.9
		R-CATTAGATCCAGAACCAGGA	55.13					
5	IFGTB4918	F-CCTTTTCTTTTCTCTCGCTC	54.36	(ATG)5	199	20	0.6	0.9
		R-AATTCTCCGACTTTCTGGTG	55.38					
6	IFGTB5113	F-GTGGATTTTGATGAGGAGGT	55.02	(TGCCAG)6	268	8	0.2	0.9
		R-AATGTAGGGTTATGGCCAAG	55.07					
7	IFGTB1732	F-GTTTTCAGACGATGCCATTT	54.87	(ACACAG)4	232	15	0.3	0.9
		R-TCCATGTTAAACAACCCCAA	55.03					
8	IFGTB7621	F-AAGATGCCACTTCATCATTC	55.1	(CTT)7	124	8	0.2	0.8
		R-GTTGAGAGGGAAGAAAAGGT	55					
9	IFGTB8938	F-CACCCAATTCCTCAATCTCA	55.02	(CAA)5	211	11	0.3	0.9
		R-ACCTTCCTCTTGTTGTTGTT	54.99					
10	IFGTB1649	F-CACCAACAGAAGAAACGAAG	54.87	(CTG)6	225	7	0.1	0.7

		R-CGGTGATATTGTTGAGTGGA	55.18					
11	IFGTB7867	F-GTGGTTTTTCTTCCACTTGA	55.25	(GGA)5	285	16	0.5	0.9
		R-CAATTTTACCCCTCACACTC	54.62					
12	IFGTB7376	F-GAAATCAATGGCTCTCAAAG	55	(CGG)5	244	5	0.1	0.8
		R-AAGCTCCTTCTTCACCTCTT	54.9					
13	IFGTB35	F-TCCAAACCCTAAAGACTCAA	54.94	(ACC)5	229	8	0.1	0.9
		R-GAAGGTGGATACAGCAGAAA	55.37					

Figures

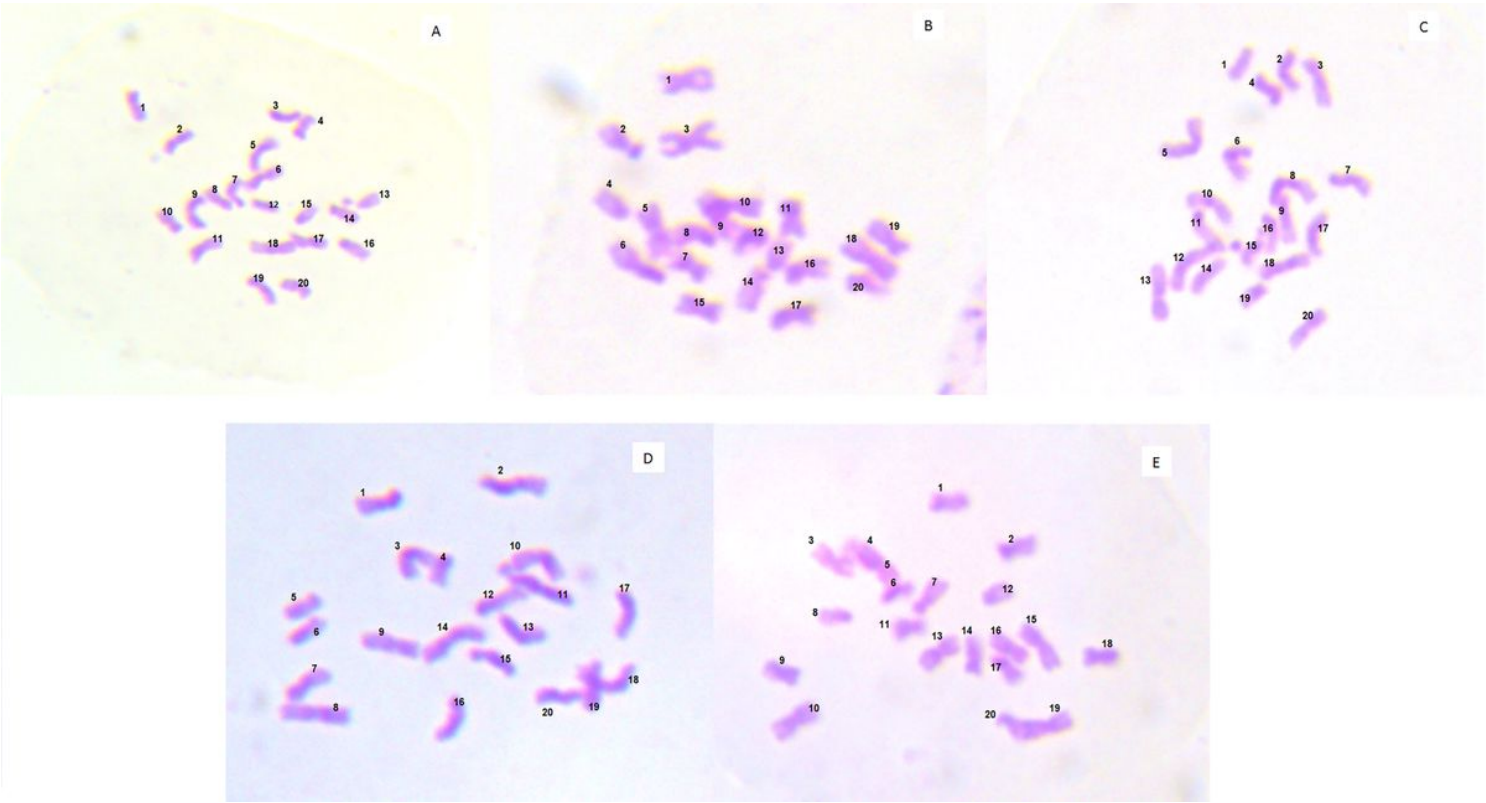

Figure 1

Somatic chromosomes in root meristem cells of *P. santalinus* (A - E). One genet was selected from each of the five natural populations.

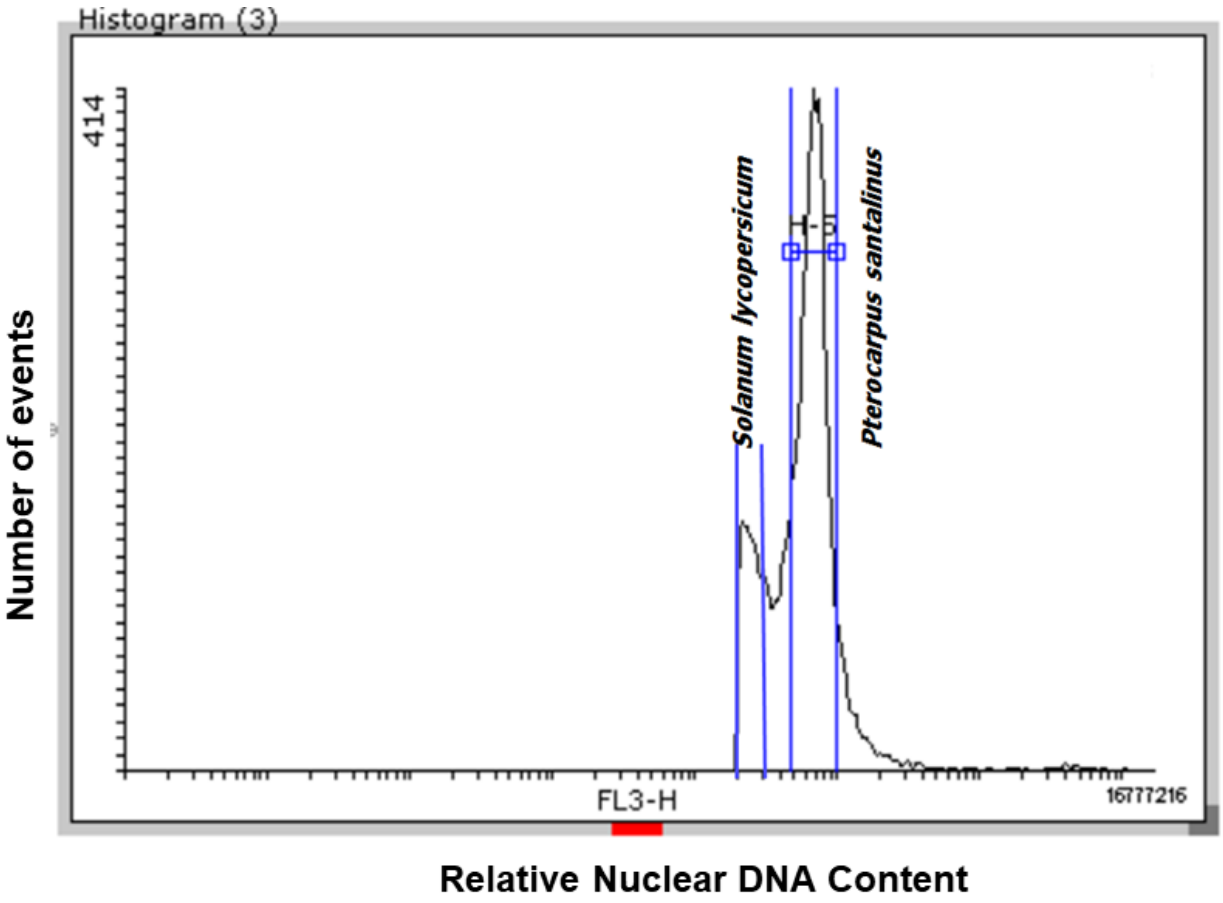

Figure 2

Histogram of PI fluorescence intensity in the nuclei of young leaves of *P. santalinus* (sample) and *Solanum lycopersicum* (internal standard, 2C = 2.0 pg)

Image not available with this version

Figure 3

Functional classification of assembled unigenes from *P. santalinus*. (a) Gene Ontology (GO) classification based on cellular component, molecular function and biological process (b) KEGG classification based on metabolism, genetic information processing and signalling and cellular processing.

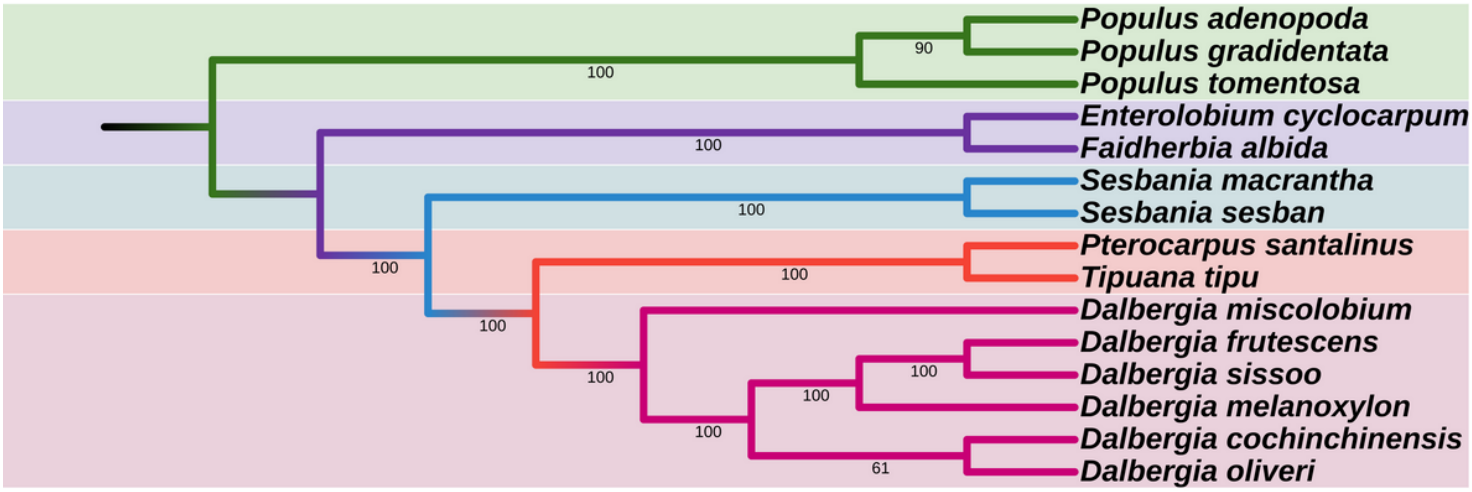

Figure 4

Molecular phylogeny of *P. santalinus* generated with single copy orthologs showing the relationship with Dalbergieae and Fabaceae members. The numbers in the node indicate bootstrap values with 1000 replications.

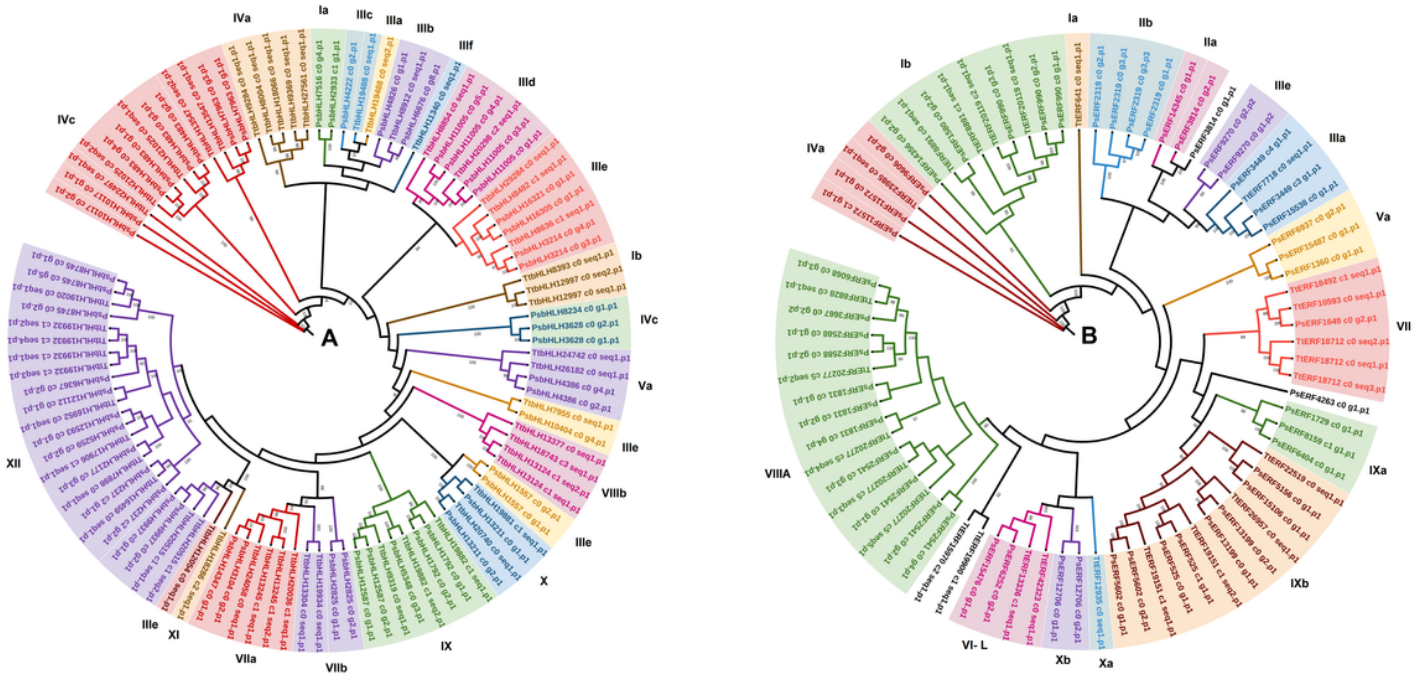

Figure 5

Phylogenetic relationships among the identified (a) bHLH and (b) ERF transcription factor family proteins in *P. santalinus* and *T. tipu*. The colored regions indicate different subfamilies and were named according to

Arabidopsis thaliana. *P. santalinus* and *T. tipu* genes are indicated at the end of the branches as Ps and Tt respectively. The numbers in the node indicate bootstrap values with 1000 replications.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile1.docx
- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx
- SupplementaryTable3.xlsx