Draft Genome Sequence of Indian Sandalwood, *Santalum album* L.

Santalum album L. (Indian sandalwood) is renowned for its high quality fragrant oil extracted from the heartwood of stem and root. The species is indigenous to the tropical belt of the Indian peninsula, eastern Indonesia and northern Australia. The threat to this species in India has reached critical level due to over exploitation, illegal harvesting, monopoly in trading, lack of established plantations, habitat loss and lack of regeneration due to fire and grazing and incidence of spike disease. The species is considered to be approaching commercial extinction, widening the gap between supply and demand.

Motivation for whole genome sequencing of *S. album*:

Inspite of its commercial and cultural relevance to India, the sandalwood improvement program has been limited and research addressing the two commercially important traits including heartwood formation and essential oil yield has not generated significant output to support the breeding program. The genomic resources for this species are extremely scanty and limited to wood transcriptome data. No microsatellite markers are developed till data for population analysis of *S. album*.

A team of scientists including Dr. Modhumita Dasgupta, Prof. K. Ulaganathan and Dr. Suma Dev from Institute of Forest Genetics and Tree Breeding, Coimbatore; Centre for Plant Molecular Biology, Osmania University, Hyderabad and Kerala Forest Research Institute, Peechi, Kerala respectively are involved in the generating the first draft genome sequence of Indian sandalwood.

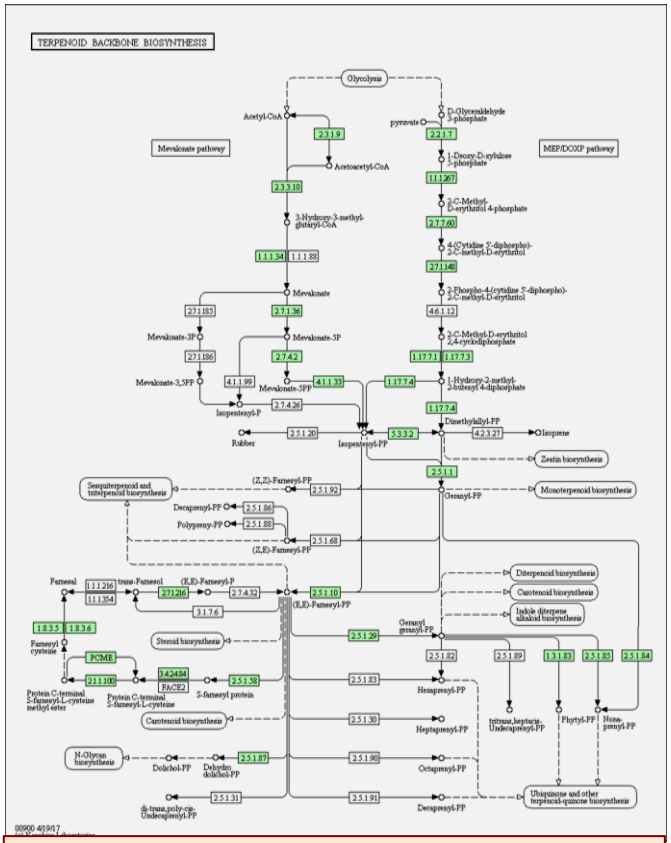
Biological material:

S. album is a diploid species with haploid chromosome number of n =11. The estimated genome size is ~ 281 Mbp (0.29 pg). Leaves of *S. album* were sourced from Kerala Forest Research Institute, Peechi, Kerala for whole genome sequencing.

Sequencing, de novo assembly and annotation:

Genomic DNA was isolated from leaves of *S. album* and whole genome sequencing was conducted using Illumina NextSeq 500 with 30X depth. Sequencing two libraries yielded a total of 33 million paired end reads of 150-270 nucleotide length. Subsequent to pre-processing, the reads were *de novo* assembled into contigs/scaffolds using Velvet genome assembler and CLC Genomic workbench. Gene prediction with Agustus tool identified ~ 40,000 protein coding genes, 180 known miRNA genes and many uncharacterized potential miRNA genes. The identified genes were annotated using homology based annotation tools. Comparative analysis of

the predicted genes with other plant genomes like *Arabidopsis thaliana*, *Populus trichocarpa* and *Oryza sativa* was conducted. Additionally, the genomic reads were used to reconstruct the mitochondrial and chloroplast genomes of Indian Sandalwood.

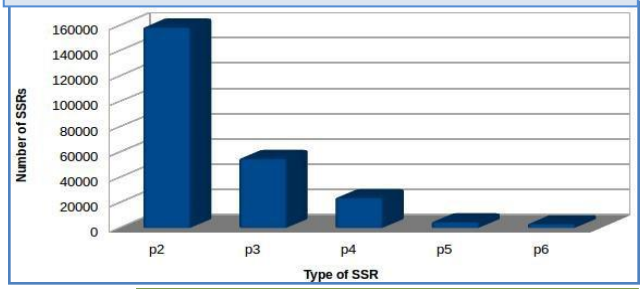


Green color blocks indicate the identification of the terpenoid biosynthesis pathway gene in *S. album* genome

The Sesquiterpene pathway

The essential oil of *S. album* contains predominantly the sesquiterpene alcohols (Z) – α -santalol, (Z) – β -santalol, (Z) –epi – β -santalol and (Z) – α –exo –bergamotol. α-santalol (40–55%) and β-santalol (12–27%) represents the major active components of sandalwood oil contributing to the fragrance. Santalol (α- and β-) is a 15C sesquiterpenoid synthesized using the universal building block Isopentenyl pyrophosphate (IPP) which is synthesized using two compartmentalized pathways, the cytosolic MVA pathway and the plastidic MEP pathway. The draft genome assembly predicted forty three genes associated with terpenoid biosynthesis along with several p450 monoxygenease genes. Further characterization of these genes and their possible role in sandalwood oil production is in progress.

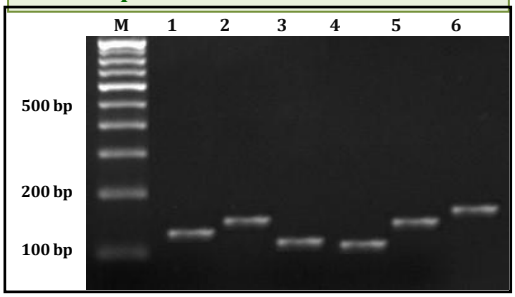
SSR distribution in *S. album* genome



Genome-wide mining for microsatellite repeats:

Trimmed reads were used for SSR prediction using MISA program. A total of 2,46,522 SSRs were predicted from the genome sequence and maximum representation was of di-nucleotide repeats (1,58,761) followed by tri nucleotide repeats (55,135). This is the first report on documentation of genome-wide SSRs in Indian sandalwood. The predicted SSRs were validated, wherein six primer pairs targeting di-nucleotide repeats were successfully amplified in two genotypes. All

Amplification of SSR markers



primers amplified in the size range from 100 – 200 bp.

Implications:

The draft genome of *S. album* will be the first reference for Santalales and will aid in

- Understanding the unique sesquiterpenoid biosynthetic pathway governing santalol production.
- Estimating genetic diversity and population structure using the genome-wide microsatellite markers.
- Accelerating breeding programs targeting heartwood formation and santalol content through genomic predictions and association genetics.
- Development of extended next generation DNA barcodes using organelle genome and nuclear ribosomal DNA sequences.