Chapter 2

RNA isolation and transcriptome sequencing for screening of genes involved in santalene biosynthesis
Chapter 2

RNA isolation and transcriptome sequencing for screening of genes involved in santalene biosynthesis

RNA sequencing is a quick and cost effective method to profile complete coding sequence of a genome due to its high throughput accuracy and reproducibility. Isolation of intact and functional RNA from interface of heartwood and sapwood from Indian Sandalwood *Santalum album* Linn is extremely difficult as it contains high percentage of polysaccharides, phenolics and other secondary metabolites. The presence of phenolics and polysaccharide compounds complicates the RNA isolation as they affect the purity as well the yield of the RNA. In this chapter, a protocol is developed for the isolation of high quality of total RNA from the interface of heartwood and sapwood from *S. album*. High quality total RNA obtained using modified protocol was used for RNA sequencing using Illumina GAII Analyzer and screening for genes involved in terpene biosynthesis in Indian Sandalwood.
2.1 Introduction

The endangered, Sandalwood or Chandan (*Santalum album* L.) belongs to the *Santalaceae* family, a medium-sized evergreen hemi-root parasitic tree, and is highly valued for its oil\(^1\), which is synthesized at the interface of heartwood and sapwood and subsequently stored in heartwood. It is believed that expression of genes involved in the biosynthesis of santalenes and their hydroxy derivatives santalols, occurs at interface of heartwood and sapwood. Isolation of total RNA from interface of heartwood and sapwood from Sandalwood is very challenging as it contains high percentage of polysaccharides, phenolics and other secondary metabolites\(^3\). Moreover, phenolic compounds are found in a wide range of polymerization states, including lignans, stilbenes, flavonoids and quinones, with the latter two associated with the colour change occurring during heartwood formation. There are several protocols\(^5\)-\(^7\) available for the isolation of total RNA from tissues rich in phenolic and polysaccharide compounds but most of them are tissue specific. However, the extraction of total RNA from plant tissues often requires modification in existing procedure or development of new procedure. The presence of phenolics and polysaccharides complicate RNA isolation from tissue rich with these molecules, as they tend to co-precipitate with nucleic acids giving a brownish colour to the RNA pellet, decreasing its solubility immensely and further leading to the degradation of RNA. The quality of RNA is very important for downstream processing because these phenolics and polysaccharides act as the inhibitors for reverse transcriptase, as well as they interfere in PCR reaction. In this study, a protocol was developed for the isolation of high quality total RNA from the interface of heartwood and sapwood from *S. album*, with extensive modifications of the protocol reported for isolation of high quality of RNA from gymnosperm and angiosperm tissue\(^7\).

RNA sequencing has become a powerful tool for the complete profiling of coding sequences of the genome\(^8\). Here we present the generation of massive amount of data from transcriptome sequencing using Next-generation sequencing (NGS) technology. Assembled data set was compared with NCBI database\(\text{\textregistered}\)Protein Data Bank (PDB) for screening for unigenes involved in terpenoid biosynthesis in Sandalwood, specifically prenyl transferase, terpene synthases and CYP450 systems.
2.2 Materials and Methods

2.2.1 Plant material

Sandalwood tissue samples were collected using wood-borer (5.15 mm diameter) at a height of about 1 meter from the ground from a 15 year old *Santalum album* tree and the tissues were immediately flash-frozen in liquid nitrogen and stored at -80 °C till further use.

2.2.2 Reagents

All the reagents and chemicals were procured from Sigma-Aldrich or Invitrogen. All the solutions were prepared in 0.1% diethylpyrocarbonate (DEPC) and sterilized in an autoclave, except Tris buffer, which was prepared in sterile DEPC treated water. All the plastic wares were soaked in freshly prepared DEPC water, dried, and sterilized in autoclave before use.

**Extraction buffer**

Lithium dodecyl sulphate (1.5 %), Lithium Chloride (300 mM), EDTA disodium salt (20 mM), Sodium Deoxycholate (1 % w/v), Tergitol nonidet NP-40 (1 % v/v). The buffer containing these constituents was autoclaved and the Tris-HCl (200 mM) pH 8.5 was added thereafter and following constituents were added just before use: Thiourea (5 mM), Aurintricarboxylic acid (1 mM), and DTT (10 mM).

PVPP (Polyvinylpolypyrrolidone)

Acetone (HPLC grade)

Chloroform/Isoamylalcohol (24:1; v/v)

3.3 M Sodium Acetate pH 6.1 (prepared in DEPC treated water and autoclaved)

5 M NaCl (prepared in DEPC treated water and autoclaved)

10 % CTAB (prepared in DEPC treated water and autoclaved)

Isopropanol

TE: 10 mM Tris-HCl; 1 mM EDTA; pH 8.0

10 M LiCl (prepared in DEPC treated water and autoclaved)

70 % Ethanol (prepared with DEPC treated autoclaved water)

DEPC treated water (DEPC from Sigma-Aldrich)

Centrifuges

Waterbath
2.2.3 Modified protocol for RNA isolation

- 1 g of plant tissue was ground in liquid nitrogen with 0.3 g of PVPP per gram of tissue in bead-beater at a frequency of 30 vibrations/second for 5 minutes.

**Critical step**

**Important for minimizing the oxidation of phenolic compounds**

- Finely crushed powder was transferred into tube and washed with acetone to remove phenolic contamination.

**Critical step**

**This step is very important to remove the phenolics (we can see the yellow colour of supernatant).**

- This was centrifuged at 3000 × g for 10 minutes at 4 °C.
- Pellet was dried and 20 mL extraction buffer/g of tissue was added.
- Vortexed at room temp for 10 minute.
- This was centrifuged at 3000 × g for 20 minutes at 4 °C.
- The supernatant was transferred into a fresh tube and kept on ice.
- To this, 2 mL 10 % CTAB solution was added at room temperature and incubated for 5 minutes at 60 °C, to remove the residual polysaccharides.

**Critical step**

**This step is very important to remove the polysaccharide to avoid the coprecipitation, which later leads to the degradation of RNA.**

- This mixture was extracted with Chloroform: Isoamyl alcohol (24:1) till the interface became clear and the supernatant was retrieved each time.

**Critical step**

**Extraction should be performed till the clear interface is observed to avoid the protein contamination.**

- The supernatant was transferred into another tube and to this, 1/9th volume of 3.3 M Sodium acetate and 0.6 volume of ice-cold isopropanol were added.
- This was incubated at -20 °C for 3 hrs and then centrifuged at 14000 × g for 20 minutes at 4 °C and the supernatant was discarded.
- To the pellet, 1 mL of TE Buffer (10 mM) and 1 mL of 5 M NaCl were added and incubated on ice for 30 minutes, with periodic vortexing.
- This mixture was extracted with chloroform: isoamyl alcohol (24:1) and the supernatant was retrieved each time. This was repeated till the interface became clear.
Critical step

Extraction should be performed till the clear interface is observed to avoid protein contamination

- To the supernatant, 2.5 M final concentration of LiCl was added and incubated for overnight at -20 °C.
- The RNA was pelleted down by centrifugation at 14000 × g for 30 minutes at 4 °C.
- The resultant pellet was washed with 70 % ethanol by centrifugation at 14,000 × g for 10 minutes at 4 °C.
- The pellet was air dried at room temperature and re-suspended in 50 μl DEPC treated water.

2.2.4 Quantification of total RNA

Total RNA was quantified using a spectrophotometer (NanoDrop, Thermo Scientific) by measuring optical density of isolated RNA in 10 mM TE buffer at a wavelengths of 230, 260, and 280 nm and purity was checked by comparing the ratio of 230/260 and 260/280. The integrity of total RNA was assessed by sharpness of rRNA (28S and 18S rRNA) on 1.5 % agarose gels and visualized by GelRed™ (Biotium).

2.2.5 RT-PCR

To test RNA quality, 1µg of total RNA was used for cDNA synthesis using RT kit from Promega according to manufacturer’s instructions. PCR Primers were designed for amplification of 18S rRNA gene (Sa18S-F 5’TGACGGAGAATTAGGGTTCG3’, and Sa18S-R 5’GTGCCAGCGGAGTCCTATAA3’) from 18S rRNA gene sequence (Accession no L24416) reported from Sandalwood. The PCR mixture was initially denatured at 94 °C for 5 minutes and then subjected to PCR condition at 20 seconds at 94 °C, 20 seconds at 55 °C and 1 minute at 72 °C for 32 cycles, and a final extension for 10 minutes at 72 °C. Reaction product was analyzed by electrophoresis on a 1 % agarose gel and visualized by GelRed™ (Biotium).

2.2.6 Transcriptome sequencing

Transcriptome library for RNA sequencing was constructed according to the Illumina TruSeq RNA library protocol outlined in “TruSeq RNA Sample Preparation Guide” (Part # 15008136; Rev. A; Nov 2010) from Genotypic Pvt. Ltd. In briefly, mRNA was purified from 1 μg of intact total RNA using oligoT beads (TruSeq RNA
Sample Preparation Kit, Illumina). The purified mRNA was fragmented for 2 minute at elevated temperature (94 °C) in the presence of divalent cations and reverse transcribed with Superscript II Reverse transcriptase by priming with Random Hexamers. Second strand cDNA was synthesized in the presence of DNA polymerase I and RnaseH. The cDNA was cleaned up using Agencourt Ampure XP SPRI beads (Beckman Coulter). Illumina adapters were ligated to the cDNA molecules after end repair and addition of A base and SPRI cleanup was performed after ligation. The library was amplified using 8 cycles of PCR for enrichment of adapter-ligated fragments. The prepared library was quantified using Nanodrop and validated for quality by running an aliquot on High Sensitivity Bioanalyzer Chip (Agilent).

2.2.7 De novo transcriptome assembly

Total 10.08 GB paired end raw data was generated with read length of 150 bp and primary QC check of the raw data was performed using the inbuilt tool SeqQC-V2.1. To obtain high quality clean read data for De novo assembly, the raw reads were filtered by discarding the reads containing adaptor sequence and poor quality raw reads (Phred score <20). The clean reads were first assembled into contigs using the Velvet_1.1.05 with an optimized hash length of 59. Assembled contigs were given as input for Oasis_0.2.01 to generate transcripts. The redundancy in the output transcripts of Oasis_0.2.01 was removed by using CD-HIT to generate unique unigenes.

2.2.8 Transcriptome annotation

To assign molecular function, biological processes and cellular components of unigenes, functional annotations were performed. ORFs were predicted in all six frames by Virtual Ribosome online program. The longest ORFs were selected for each unigenes and submitted to Pfam-A database to identify protein domain and architecture. Unigenes assigned with Pfam ID were used to perform BLAST2GO search against NCBI Nr database, SwissProt/Uniprot database, Protein Data Bank (PDB) with an E-value≤10^{-5}. The FASTA format of all the unigenes were submitted to KEGG database to assign KO (KEGG Orthology) number and generate KEGG pathways.
Chapter 2

2.3 Results and Discussion

2.3.1 Optimization of RNA isolation from the interface of heartwood and sapwood

Several standard protocols have been reported for isolation of total RNA from various plant tissue specialized for high phenolic and polysaccharide rich tissue, including CTAB-NaCl method\(^9\), modified SDS/Phenol method\(^10\). However, these protocols failed to yield high quality of total RNA from the interface of heartwood and sapwood from Sandalwood. Isolation of total RNA using CTAB-NaCl buffer, from the interface of heartwood and sapwood from Sandalwood was not efficient as the tissue is rich in polysaccharides and phenolics, SDS/Phenol method was also not effective to remove these molecules, which led to the degradation of RNA due to co-precipitation of polysaccharides and phenolics. High level of polyphenolic compounds interact irreversibly with DNA, RNA and proteins\(^11\), leading to RNA degradation\(^12\).

The RNA isolation protocol reported for RNA isolation from gymnosperm and angiosperm tissue\(^7\) was not efficient for isolation of high quality of RNA. This method resulted in very low yield with partially degraded RNA (indicated by strong smear) and also was found to be contaminated with genomic DNA. Moreover, the 28s rRNA and 18S rRNA band are not clear (Figure 2.3.1). The isolated pellet was also brown in colour and difficult to dissolve. This may be due to browning effect\(^11\), where phenolic compound, upon oxidation, develops a brown colour supernatant. The absorbance ratio \(A_{260/280}\) and \(A_{260/230}\) were found to be 1.12-1.52 and 0.42-0.65 respectively indicating contamination of protein, polyphenolics and polysaccharide. This quality of RNA was not suitable for the downstream processing for the discovery of genes involved in santalene biosynthesis.
Figure 2.3.1: Agarose gel electrophoresis of total RNA isolated using protocol reported for RNA isolation from xylem tissue from gymnosperm, **Lane 1 and 2:**
Total RNA $A_{260/280}$ ratio = 1.52, $A_{260/230}$ ratio = 0.65

To overcome these problems, the above protocol was extensively modified to isolate better quality of RNA. PVPP which is an inhibitor of polyphenol oxidase and can prevent browning effect\textsuperscript{13,14} was mixed with tissue before crushing, so that it could bind to phenolic compounds and prevent their oxidation, which in turn could not bind to nucleic acids. The ground tissue mixture was washed with acetone/methanol to remove soluble phenolic contaminants rendering a significant improvement in RNA quality. CTAB treatment before chloroform:isoamyl alcohol extraction was extremely helpful in removing the polysaccharide contamination and avoiding its co-precipitation with nucleic acid.

Total RNA extracted from interface of heartwood and sapwood of Indian Sandalwood produced two distinct rRNA bands (28S rRNA and 18S rRNA) on a 1.5 % agarose gel electrophoresis in the ratio of 2:1 without degradation and also showed no genomic DNA contamination (Figure 2.3.2).
The ratio of $A_{260/280}$ and $A_{260/230}$ were found to be 2.14 and 2.35 respectively, indicating no contamination of protein, polysaccharides, polyphenolics, salts, and solvents. The yield of total RNA ranges from 50-70 µg/g of tissue. To check the integrity of isolated total RNA, it was reverse transcribed as discussed earlier, and PCR amplification with 18 rRNA primer resulted in amplification of the expected 750 bp amplicon (Figure 2.3.3).

In summary, our modified protocol is simple and highly effective for extraction of high quality total RNA from the wood tissue which is rich in polysaccharides,
polyphenolics and other secondary metabolites. The use of PVPP along with tissue crushing and washing with solvent renders a significant improvement in the quality of total RNA as compared to the other protocols. The optimized protocol enabled us to isolate high quality total RNA without degradation and secondary metabolite contamination, which was used for downstream processing such as reverse transcription, EST library preparation and gene amplification.
2.3.2 Transcriptome sequencing and screening of genes involved in terpenoid biosynthesis in Indian Sandalwood

2.3.2.1 cDNA library preparation and sequencing

cDNA library was constructed from the mRNA purified from total RNA isolated from the interface of heartwood and sapwood of Indian Sandalwood, and was subjected to RNA sequencing using Illumina GAII Analyzer. cDNA library was constructed as discussed earlier and amplified by PCR to enrich the adaptor ligated fragments. The cDNA library was sequenced on one lane of the flow cell using paired end sequencing. A total of 33323756 raw reads were generated with a length of 150 bp corresponding to 10.08 GB. Primary QC check for raw reads was performed using the inbuilt tool SeqQC-V2.1. From the observation of QC report, adapter trimming and low quality trimming was performed throughout the sequence to get better quality of reads.

![Figure 2.3.2.1: Bioanalyzer profile of amplified adaptor ligated fragments](image)

2.3.2.2 De novo transcriptome assembly

Total 33323756 high quality paired end reads were generated with read length of 150bp. By discarding reads containing adaptor sequence, poor quality reads (Phred score <20), clean reads were preliminarily assembled into contigs by Velvet_1.1.05 with various hash lengths from 51 to 113. Total 58221 contigs were generated with optimised hash length of 59 with average contig length 571.6 bp and N50 value 863. These contigs were given as input for Oasis_0.2.01 to generate 90,478 transcripts having N50 value of 695 and average transcript length 474.18 bp. The total transcripts
were further subjected to cluster and assembly analysis using CD-HIT to remove the redundancy. Finally, it resulted in a total of 84,094 unique transcripts with an average size of 494.17 bp and N50 length 717, which contains 9136 transcripts (10.86 %) with length greater than 1 kb and 24912 transcripts (29.62 %) with length greater than 500 bp. The length distribution of transcripts is shown in fig 2.3.2.2. These results showed that the high throughput sequencing and assembly quality is good, which could be used for functional annotation.

**Transcripts Length Distribution**

![Transcript length distribution graph](image)

**Figure 2.3.2.2**: Transcript length distribution graph

**Table 2.1**: Transcriptome assembly statistics

<table>
<thead>
<tr>
<th>Velvet_1.1.05 (Oases 0.2.01) Transcripts Statistics</th>
<th>Contigs</th>
<th>Total transcript</th>
<th>Unigenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hash_length</td>
<td>59</td>
<td>59</td>
<td>55</td>
</tr>
<tr>
<td>Transcripts Generated :</td>
<td>58221</td>
<td>90478</td>
<td>84094</td>
</tr>
<tr>
<td>Maximum Transcript Length :</td>
<td>11726</td>
<td>12279</td>
<td>12279</td>
</tr>
<tr>
<td>Minimum Transcript Length :</td>
<td>117</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Average Transcript Length</td>
<td>571.671</td>
<td>474.18</td>
<td>494.17</td>
</tr>
<tr>
<td>Total Transcripts Length :</td>
<td>33283265</td>
<td>42902769</td>
<td>41557227</td>
</tr>
<tr>
<td>Total Number of Non-ATGC Characters :</td>
<td>20</td>
<td>120683</td>
<td>112096</td>
</tr>
<tr>
<td>Percentage of Non-ATGC Characters :</td>
<td>6.00903e-07</td>
<td>0.00281294</td>
<td>0.00269739</td>
</tr>
<tr>
<td>Transcripts &gt; 100 bp :</td>
<td>58221</td>
<td>90476</td>
<td>84092</td>
</tr>
</tbody>
</table>
2.3.2.3 Functional annotation of unigenes

Functional annotation of unigenes allows for insight into particular molecular function and biological processes in which the putative proteins are involved. We applied various approaches to functionally annotate the assembled transcripts.

2.3.2.3.1 KASS analysis

To identify the terpenoid biosynthetic pathways present at the interface of heartwood and sapwood of Indian Sandalwood, all the 84,094 unigenes were submitted to KEGG database for functional annotation of genes by bidirectional BLAST comparison against manually curated KEGG (Kyoto Encyclopedia of Genes and Genomes) database.\(^5\)

![KAAS analysis](image)

**Figure 2.3.2.3:** KAAS analysis of unigenes for KEGG pathway mapping

The results contains 4244 unigenes are assigned with KO number representing 298 KEGG pathways, from which 2702 unigenes represent unique KO number. The predicted pathways represented the majority of plant biochemical pathways including metabolism, cellular processes and genetic information processing. According to KEGG pathway mapping of our sequence dataset, 30 unigenes were found to be involved in terpenoid backbone biosynthesis and assigned with KO number.
Figure 2.3.2.4: KEGG pathways map for terpenoid biosynthesis, green colour highlighted IDs represent the presence of these pathway enzymes in our unigenes

2.3.2.3.2 Pfam analysis

Pfam is a database of manually curated protein families which contains information about protein domain and protein families represented as multiple sequence alignments and as profile hidden Markov models\textsuperscript{16}. For the functional annotation of unigenes, we made an Open Reading Frame (ORF) prediction analysis using online software Virtual Ribosome-V1.1. 84,094 unigenes were submitted to Virtual Ribosome to predict ORF of maximum length for each unigene in all six frames. Total 83,823 unigenes (99.66\%) were identified as having ORF starting at an ATG
codon, out of which 17,119 unigenes (20.35 %) were found to have the ORF of $\geq$100 amino acid length without redundancy, whereas 66697 (79.31 %) unigenes contains ORF $\leq$ 100 amino acid length and 278 (0.33 %) unigenes were found without having ORF in any of the frame (Figure 2.3.2.5).

**Virtual Ribosome**

![Virtual Ribosome Diagram](image)

**Figure 2.3.2.5:** ORF prediction using Virtual Ribosome-V-1.1

17,119 unigenes having ORF of $\geq$ 100 amino acid were submitted to Pfam-A database, which allows efficient and sustainable manual curation of alignments and annotation (Figure 2.3.2.6). From the Pfam analysis only 10,668 unigenes were assigned with Pfam ID out of 17,119 unigenes. Further analysis of 10,668 unigenes having Pfam ID revealed that 18 unigenes belonged to terpene biosynthesis representing Pfam ID: PF01397 (terpene synthase N-terminal domain), PF03936 (terpene synthase family metal binding domain), PF00432 (prenyl transferase and squalene oxidase repeat) and PF13243 (prenyl transferase-like) were selected for screening of terpene synthases.
Figure 2.3.2.6: Pfam analysis of transcript having ORF $\geq$100 amino acid

10,668 unigenes assigned with Pfam ID were subjected to Blast2Go analysis against NCBI Nr-database, Swissprot/ Uniprot database to screen the genes involved in terpenoid biosynthesis in Indian Sandalwood.

2.3.2.3.3 BLAST2GO analysis

BLAST2GO is a tool often used in functional genomics research. It works by blasting assembled transcript sequences against protein database in NCBI and assign GO terms for functional characterization of the unigenes.

Figure 2.3.2.7: BLAST2GO analysis of unigenes assigned with Pfam ID for the screening of terpenoid biosynthetic pathway genes
BLAST2GO analysis resulted in the screening of 18 unigenes (Figure 2.3.2.7), which have terpene synthase domain and may be involved in biosynthesis of sesquiterpenes present in Indian Sandalwood. Further screening for CYP450 system resulted in 72 unigenes containing the domain of CYP450, from which 10 unigenes were matching with terpene hydroxylase. These CYP450 systems require a CYP450 reductase, which helps in the hydroxylation. Transcriptome screening resulted in identification of 3 transcripts shown to have the domain of CYP450 reductase were selected. These transcripts will be used for establishing the biosynthesis of santalols in Indian Sandalwood *S. album*. 
2.4 Conclusion

In plant, high-throughput RNA sequencing accelerated the discovery of new genes, transcription pattern, and functional analysis. In the present work, we have optimized the protocol for the isolation of high quality of total RNA from the interface of heartwood and sapwood of Indian Sandalwood. RNA sequencing was performed on Illumina GAII Analyzer and 10.08 GB of sequence data file was generated. The raw data was assembled into 84,094 unique unigenes and screened against KEGG database, Pfam database, NCBI Nr-database, and Swissprot / Uniprot database for unigenes containing terpene synthase domain. These analyses resulted in identification of 18 unigenes related to terpene synthases, 72 unigenes for CYP450 mono-oxygenase and 3 unigenes representing CYP450 reductases. These unigenes were used for cloning and functional characterization of the genes involved in biosynthesis of santalene derivatives.
2.5 References