Data Article

Leaf tissue specific transcriptome sequence and de novo assembly datasets of Asiatic mangrove *Rhizophora mucronata* Lam.

S.P. Meera, Anusha Sreeshan, Anu Augustine*

Department of Biotechnology and Microbiology, Kannur University, Thalassery Campus, Palayad P.O., Kannur-670661, Kerala, India

**Abstract**

Transcriptome data is beneficial to explore molecular mechanisms of extreme adaptations in non-model organisms like mangroves. In this data article, five major datasets and two data sub sets of a salt secreting mangrove, *Rhizophora mucronata* Lam., were described. A combination of Illumina HiSeq 2500, Trinity, BLAST X, Bowtie 2 and BLAST 2GO was used for RNA Seq, de novo assembly, transcript annotation, gene expression estimation and gene ontology annotation respectively. The RNA Sequence (Read 1 and Read 2) in Sequence Read Archive amounting to 46,366,348 paired end raw reads is the first data set made open for de novo or comparative transcript assembly. Assembled sequences of 93960 gene transcripts constitute the second data set in Transcriptome Shotgun Assembly. The gene/protein annotations to the assembled transcripts give two sub data sets containing 93960 each of GenBank and GenPept entries with comprehensive cDNA and translated protein sequences of genes. Of these, predicted proteins for 87768 coding sequences, mapped to UniProtKB serve as the third data set. The gene expression levels of the annotated transcripts comprise the fourth data set in Gene Expression Omnibus. The fifth data set in Figshare includes 44,028 gene ontology terms extracted for 21,073 confident transcripts. The data sets provide a valuable resource for further analyses including transcriptomic changes in response to environmental stresses.
Specifications table

| Subject       | Biology                  |
|---------------|--------------------------|
| Specific subject area | Plant Biotechnology; Genetics and Molecular Biology (General); Bioinformatics |
| Type of data  | RNA Seq- Illumina HiSeq 2500 System |
| How data were acquired | De novo transcriptome assembly- trinitynaseq_r20140717.tar.gz |
|               | Transcriptome annotation- BLAST X |
|               | Gene expression estimation-Bowtie 2 |
|               | Gene ontology annotation -BLAST 2GO |
| Data format   | RNA Seq Raw reads- FASTQ |
|               | Assembly and annotation- .sqn |
|               | Gene expression values- .txt |
|               | Gene ontology annotation- GAF version2. |
| Parameters for data collection | The leaf sample of *R. mucronata* were randomly collected from different trees during the summer season |
| Description of data collection | The randomly collected leaf tissues were pooled into two replicates and used for total RNA isolation using LiCl cold extraction method. RNA pool with RIN value 7.2 was processed for mRNA purification using magnetic beads to which poly-T oligomers were attached. High temperature was applied to fragment the pure form mRNA into small pieces in presence of divalent cations. The mRNA fragments are then processed for cDNA synthesis and subsequent sequencing on Illumina HiSeq 2500 platform. The RNA sequence data from the sequencer were converted to FASTQ files and de novo assembled using Trinity. BLAST X and BLAST 2GO were used for the assembled transcript annotation. |
| Data source location | 9°59′17.9″N 76°16′21.8″E |
| Data accessibility | India, Mangalavanam sanctuary, Ayyappankavu, Cochin, Kerala |

1. NCBI Sequence Read Archive SRR5012157. [https://www.ncbi.nlm.nih.gov/sra?term=SRR5012157](https://www.ncbi.nlm.nih.gov/sra?term=SRR5012157)
2. NCBI Transcriptome Shot gun Assembly GGEC00000000/ version GGEC00000000.1. [https://www.ncbi.nlm.nih.gov/nuccore/GGEC00000000](https://www.ncbi.nlm.nih.gov/nuccore/GGEC00000000)
   a) DDBJ/ENA/GenBank GGEC01093960. [https://www.ncbi.nlm.nih.gov/Traces/wgs/?val=GGEC01](https://www.ncbi.nlm.nih.gov/Traces/wgs/?val=GGEC01)
   b) GenPept MBW80484.1- MBX74444.1. [https://www.ncbi.nlm.nih.gov/Traces/wgs/?val=GGEC01](https://www.ncbi.nlm.nih.gov/Traces/wgs/?val=GGEC01)
3. UniProtKB/TrEMBL A0A2P2IGY3 - A0A2P2R510. [https://www.uniprot.org/uniprot/?query=organism%3A%22Rhizophora+ mucronata%22&species%3A%5B%5B61149%5D%22&sort=score](https://www.uniprot.org/uniprot/?query=organism%3A%22Rhizophora+mucronata%22&species%3A%5B%5B61149%5D%22&sort=score)
4. Gene Expression Omnibus GSE112162. [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3059021](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3059021)
5. Figshare 7016204. [https://doi.org/10.6084/m9.figshare.7016204](https://doi.org/10.6084/m9.figshare.7016204)

Value of the data

- The RNA seq raw datasets will be valuable to the mangrove research groups for identifying transcripts by either mapping reads to a related genome or re-assembling them de novo.
- The assembled unique coding sequences of transcripts involved in various plant cellular mechanisms can be used for designing probes and primers for gene expression/transformation studies.
- The uncharacterized gene sequences comprising 77.57% of the total transcriptome are open for molecular characterization by mangrove research community.
- The gene expression dataset can be used as a reference to choose genes involved in salt responsive as well as other metabolic pathways including the novel ones.
• The entire data set can be useful to explore evolutionary history of *R. mucronata* by comparing with other mangrove species.

1. Data description

The data reported here is a compilation of five major datasets and two data sub sets. The first one is the RNA-Seq raw reads of *R. mucronata* transcriptome for both read orientations (read 1 and read 2) of the paired end library in FASTQ format. The sequenced data were deposited at NCBI Sequence Read Archive (Accession SRR5012157) under the Bioproject accession PRJNA345155. Prior to submission, the identity of RNA-Seq raw reads for both the read orientations were validated with their md5checksum values 73b25f8e75ee37f334993e98b18ef063 and f8f0b96ccf0bb1ddd5f3bea87582fbb17 respectively. The *de novo* assembled RNA sequence reads generated by Trinity are contained in the second data set. This Transcriptome Shotgun Assembly project has been deposited under the accession GJEC00000000. The first version of the project, GJEC01000000 is described in this paper. Annotations to the transcript assembly give rise to two subsets of 93960 DDBJ/ENA/GenBank (Accession GJEC01000001-GJEC01093960) and 93960 GenPept entries (Accession MBW80484.1-MBX74444.1.). The file format used for assembly and annotation data was .sqn. The predicted proteins for 87768 coding sequences, mapped to UniprotKB together constitute the third data set (Accession A0A2P21GY3-A0A2P21R510). The fourth dataset in Gene Expression Omnibus [1] represents the gene expression levels (FPKM) of the annotated transcripts and can be downloaded in .txt format. The contig ID assigned for contigs by *de novo* assembly is listed in the first column of the .txt file. GenBank accessions of each contig IDs and the respective FPKM values of gene expression are listed in second and third column respectively (GEO Series accession GSE112162). The fifth dataset in Figshare is of 44,028 gene ontology terms extracted for 21,073 transcripts with significant uniprot hit (Accession 7016204). The file format of GO submission is GAF version 2.

2. Experimental Design, Materials and Methods

2.1. Plant material collection

The leaf sample of *R. mucronata* with TaxId 61149 has been deposited under NCBI BioSample accession SAMN05846347. The leaf tissues, randomly collected from different trees, of Mangalavanam sanctuary, Ayyappankavu, Cochin, Kerala (Coordinates: 9°59′17.9″N 76°16′21.8″E), during the summer season were used for transcript library preparation. The leaves were immediately carried to the research lab in sterile covers embedded in dry ice. To remove dust particles, the leaf surface was washed with distilled water and sterilized with 75% ethanol. Sterile leaves were pooled into two replicates (RNA 1 and RNA 2) and processed immediately for RNA extraction to avoid RNA degradation.

2.2. Transcript Library Construction and Illumina Sequencing

The experiments involved in transcriptome library preparation were conducted at Agrigenome Labs Pvt Ltd, Cochin, Kerala. Total RNA from the leaves were isolated by LiCl cold extraction method developed for mangrove tissues rich in polysaccharides and polyphenols [2]. The purity and quantity of total RNA were assessed with Agilent 2100 Bioanalyzer, and Nanodrop respectively. For the purification of mRNA carrying poly-A tail, magnetic beads attached to poly-T oligo nucleotides were used. Further fragmentation of purified mRNA was done with the application of high temperature in presence of divalent cations. The mRNA fragments served as the template for reverse transcriptase in a reaction mix containing random primers and thus
first strand cDNA were synthesized. DNA polymerase I together with RNase H was used to convert these single stranded cDNA into second strand cDNA. Prior to adapter ligation, every second strand cDNA molecule were modified with a single ‘A’ base. After this end repair process, purification of the final products was done and PCR amplified for the construction of the final cDNA library (Illumina’s TrueSeq RNA sample preparation kit). For ensuring mRNA quality, fragmentation sizes, enrichment success and final library sizes, Bioanalyzer plots were referred at each and every stage. Prior to sequencing on Illumina HiSeq 2500 platform, the library quantity was measured using Qubit and qPCR. The RNA-Seq data obtained after the sequencing run were processed to generate FASTQ files.

2.3. Data pre-processing and De novo transcriptome assembly

Quality parameters like average base content per read, base quality score distributions and GC distribution were checked for the fastq files obtained from sequencer. The adapter sequences added prior to sequencing were trimmed using Cutadapt v1.9 [3]. Based on quality report of fastq files, sickle v1.33 programme was used to filter out all low quality (Q<20) and low read length (Read length < 30 bases) data. The quality filtered data was then normalized using BB-NORM tool for assembly. We followed de novo strategy for the reconstruction of transcriptome of R. mucronata. The processed reads were assembled using the default Trinity programme [4, 5]. The transcriptome summary is given in Table 1.

2.4. Gene expression estimation

Bowtie2 program was used to align the quality filtered reads to the assembled transcriptome and the gene expression values were estimated [6]. All multiple mapped positions were recorded with a maximum of single mismatch in the seed region of length 31bp. About 95.14% of filtered reads were found in proper alignment to the assembled transcriptome. In brief, 52,153 unique transcripts observed with expression \(>1\) FPKM were processed for downstream annotation.

2.5. Transcriptome annotation

The de novo assembled transcripts were annotated using a series of bioinformatic applications. Briefly, the following steps were performed like BLASTX comparison with NCBI database, organism annotation and gene ontology (GO) annotation. First of all, BLASTX program was used to align the assembled transcripts with NCBI protein database. We then extracted the contigs without annotation and aligned to NCBI non-redundant reference sequence database. The transcripts having an E-value of \(<= 10^{-5}\) and similarity score of \(>=40\%\) to the matches found were further annotated. For each transcript, the most similar BLASTX and the corresponding organism name were retrieved. Once the protein was predicted using BLASTX, it was further annotated to public databases like UniProt Pathway, NCBI etc. Uniprot database could provide annotations for 21,073 BLASTX hit transcripts whereas NCBI protein predictions were used for the remaining.

### Table 1

| Paired end reads (No.) | 46,366,348 |
| Bases(MB) | 4,636,64 |
| Assembled transcripts (No.) | 93,960 |
| Length of the longest transcript (bp) | 17,650 |
| Transcript GC % (Mean value) | 41.22 |
Table 2
The softwares used for dataset preparation. For the mapping and annotation of GO, licensed version of Blast2GO by Agrigenome was used.

| Name     | Application                           | Version | Link                                      |
|----------|---------------------------------------|---------|-------------------------------------------|
| Cutadapt | Removal of Dimers/Adapters            | 1.9     | https://github.com/marcelm/cutadapt       |
| Sickle   | Quality Filtering                     | 1.33    | https://codeload.github.com/najoshi/sickle.tar.gz/v1.33 |
| BBnorm   | Normalization                         | 36.20   | https://github.com/BioInfoTools/BBMap/blob/master/sh/bbnorm.sh |
| Trinity  | Assembly                              | trinityrnaseq_r20140717 | http://trinityrnaseq.sourceforge.net |
| Bowtie   | Expression Estimation                 | 2.2.2   | http://bowtie-bio.sourceforge.net/index.shtml |
| BLASTX   | Homology based annotation             | 2.2.28  | http://www.ncbi.nlm.nih.gov/blast/blast.cgi?PROGRAM=blastx&PAGE_TYPE=blastSearch&LINK_LOC=blasthome |

2.6. Gene Ontology (GO) annotation

From gene ontology consortium, GO annotations for transcripts were collected based on UniProtKB keyword mapping using Blast2GO [7]. GO terms corresponding to biological process, molecular function and cellular localization were 1995, 2436 and 1435 respectively.

2.7. Code availability

Custom codes were used for all studies in the generation and processing of datasets using publically available bioinformatics tools (Table 2). For the final submissions, datasets were prepared in file formats recommended by the respective repositories. The transcriptome assembly data and the respective protein annotations were prepared in Sequin file format (.sqn) using tbl2asn command-line program so as to submit at Transcriptome Shotgun Assembly (TSA) of NCBI. Tbl2asn require the nucleotide sequence data in FASTA format (.fsa) and Genbank submission template with submitter's details (.sbt) to generate .sqn file. The command used for .sqn preparation was ‘tbl2asn -t E:\tsa2\input\tsa.sbt -p E:\tsa2\input\ -a s -V v -j ‘[organism=Rhizophora mucronata Lam]’ –kc’. Tbl2asn reads sequence annotations in the .sqn file and automates the submission of Genbank sequence records. The requisite file types for Gene Expression Omnibus (GEO; spreadsheet) and Gene Ontology (GO; GAF.2.) submissions were prepared by assembling data using custom codes.

3. Technical Validation

3.1. RNA and cDNA library quality control

Leaf tissues of *R. mucronata* were collected in dry ice containing boxes, transported to the laboratory and processed immediately to avoid RNA degradation. Good quality RNA was then isolated from the leaf sample. The quality was assessed using Agilent’s 2100 Bio analyzer. Nanodrop quantification was also performed on each RNA sample. The RNA integrity number (RIN) for RNA 1 and RNA 2 were 6.7 and 7.2 respectively. The RNA sample with RIN value >7 and concentration >50 ng/μl was processed for transcript library construction. Following amplification, cDNA quality was also assessed using Bioanalyzer DNA 21000 Chip. Bio analyzer plots for isolated RNA (Fig. 1) and amplified cDNA library (Fig. 2) were proper and ensured the quality of mRNA, proper enrichment, magnitude of fragments and final library sizes.
3.2. Sequencing quality control

For testing sequencing quality, multiple parameters were considered. Total read count and total base count of both the read orientation (R1 and R2) were compared and found to be of similar magnitude. The total number of reads was 23183174 and the total number of bases was 2318.32 MB for both R1 and R2. To ensure the quality of reads, phred score (mean read quality) for individual reads was calculated. R1 had a phred score value 36.67 and for R2, it is 35.91 which represents 99.99% base call accuracy [8]. We tested the sample to pass FastQC22 for basic statistics such as average base content per read, per base quality score distribution, length distribution, GC distribution in the reads and the quality of raw reads. The average base quality of ∼92% of bases in the paired-end read sequences are shown to be higher than Q30 with an error-probability <=0.001. The average GC content of the reads in the sample followed a normal distribution. The raw read quality measures are summarized in Table 3.
Fig. 2. Agilent 2100 Bioanalyzer tape station profile showing quality of library. The lower and upper peaks are markers, used as internal references to determine the molecular size of the sample. The peak 271 is the mean peak which represents the optimal mean value range between 225 to 280bp.

Table 3
Raw reads quality control. The quality percentage from %Q<10 to %Q>30 and the phred score quality indicators were calculated for each read orientation from the FastQC output.

| Read orientation       | R1                  | R2                  |
|------------------------|---------------------|---------------------|
| Mean read quality (Phred score) | 36.67               | 35.91               |
| Number of reads        | 23183174            | 23183174            |
| %GC                    | 42.72               | 42.43               |
| %Q<10                  | 0.91                | 2.18                |
| %Q 10-20               | 1.02                | 1.57                |
| %Q 20-30               | 3.71                | 4.62                |
| %Q>30                  | 94.35               | 91.64               |
| Number of bases(MB)    | 2318.32             | 2318.32             |
| Mean read length(bp)   | 100.0               | 100.0               |

3.3. De novo Assembly quality control

The assembled transcript length and GC distribution patterns prove that all assembled transcripts are of length more than 200bp with a normal GC content pattern (Fig. 3 a & b). The optimal length of transcripts ensures that they are not too short to be contaminated with adapter sequences nor too long to miss any shorter transcripts [9]. About 95.14% of reads from each sequencing orientation were properly aligned back and all multiple mapped positions were recorded with a maximum of single mismatch in the 31 bp long seed region.

3.4. Annotation quality control

So as to get quality annotations, the BLASTX search E-value threshold and the similarity score were set to <= 10^-5 and >= 40% respectively. Around 95.27% of assembled transcripts amounting to 49,688 were found to have significant similarity to NCBI entries. At least 1e^-5 confidence level was observed for 72% of BLASTX transcripts, which indicates high protein conservation. Near to 87% of the BLASTX transcripts have protein level similarity of more than 60% with NCBI database proteins (Fig. 3 c & d).
Fig. 3. Assembly and Annotation quality control parameters. The length distribution pattern and GC percentage of assembled transcripts is shown in (a) and (b) respectively. All the assembled transcripts were of length >200 bp and the GC percentage followed a normal distribution pattern which ensured the assembly quality. BLASTX E-value distribution pie chart (c) outlines the percentage of transcripts in different E- value ranges from $1 \times 10^{-5}$ to 0, significant enough to describe the similarity of assembled transcripts with the existing proteins. The percentage of transcripts and the corresponding BLASTX similarity score is depicted in the pie chart (d). (c) and (d) explains good quality annotations.
4. Usage Notes

The RNA-Seq raw reads for both Read 1 and Read 2 of paired end transcript library that have been deposited in Sequence read archive allows researchers, to identify transcripts by mapping reads to a related genome. They can also re-assemble them de novo. The assembly and annotation datasets can be utilized as a base to address a research problem. This novel transcriptome data explores unique coding sequences of transcripts involved in various plant cellular mechanisms which can serve as templates for designing probes and primers in advanced analytical studies. *R. mucronata* being a salt tolerant mangrove variety, the transcriptome will help in picking up genes involved in salt responsive pathways. Transcript sequences comprising around 77.57% of the total transcriptome being uncharacterized may hold novel gene candidates of *R. mucronata* which are open for further characterization in a broad future perspective. The data set of expression analysis can be used as a reference to choose the novel gene suspects with significant expression levels. Gene ontology annotations can be used to explain the molecular function, the cellular localization and the exclusive metabolic pathways of selected genes. These annotations can ease the process of gene characterization in mangrove research. Evolutionary history of mangroves can also be explored, by comparing the transcriptome data with that of other mangrove species.

Declaration of Competing Interest

S.P. Meera has received research fellowship from University Grants Commission (UGC). Anu Augustine and Anusha Sreeshan have received project grant and research fellowship respectively from Kerala State Council for Science Technology and Environment (KSCSTE) during the conduct of the study.

Acknowledgements

The authors would like to thank Mr. S.P. Midhun, technology lead; Infosys Ltd. for providing software services. The authors would also like to thank Dr. V.K. Abdul Kareem, Postdoctoral research associate, School of Life and Environmental Sciences, The University of Sydney for proof reading the article.

Author contributions

**S.P. Meera**: Conceptualization, Methodology, Software, Validation, Investigation, Writing - Original Draft, Data Curation **Anusha Sreeshan**: Investigation **Anu Augustine**: Conceptualization, Writing - Review & Editing, Supervision, Funding acquisition

References

[1] R. Edgar, M. Domrachev, A.E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, Nucleic. Acids. Res 30 (2002) 207–210 https://doi.org/10.1093/nar/30.1.207.
[2] J.A. Rubio pina, O. Zapata Perez. Isolation of total RNA from tissues rich in polyphenols and polysaccharides of mangrove plants, Electron J. Biotechn 14 (2011) 1–9, doi:10.2225/vol14-issue5-fulltext-10.
[3] M. Marcel, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMB net. J 17 (2011) 10–12 https://doi.org/10.14806/ej.17.1.200.
[4] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, et al., Full-length transcriptome assembly from RNA-seq data without a reference genome, Nat. Biotechnol 29 (2011) 644–652 https://doi.org/10.1038/nbt.1883.
[5] B.J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, et al., *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, Nat. Protoc 8 (2013) 1494–1512 https://doi.org/10.1038/nprot.2013.084.
[6] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol 10 (2009) R25 https://doi.org/10.1186/gb-2009-10-3-r25.

[7] S. Gotz, J.M. Garcia-Gomez, J. Terol, T.D. Williams, S.H. Nagaraj, M.J. Nueda, M. Robles, M. Talón, J. Dopazo, A. Conesa, High-throughput functional annotation and data mining with the Blast2GO suite, Nucleic Acids Res 36 (2008) 3420–3435 https://doi.org/10.1093/nar/gkn176.

[8] B. Ewing, P. Green, Base-calling of automated sequencer traces using phred. II. Error probabilities, Genome Res 8 (1998) 186–194 https://doi.org/10.1101/gr.8.3.186.

[9] J.B.W. Wolf, Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial, Mol. Ecol. Resour 13 (2013) 559–572 https://doi.org/10.1111/1755-0998.12109.