Dec 02, 2021

🌐 Assembly, Annotation, Quantification, and Differential Expression Analysis of *Shorea* sp. Transcriptome

Ahmad Husaini AHS Suhaimi[1]
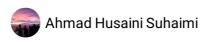
[1]Universiti Malaya

| 1 | ⤫ |

MP2Lab

Ahmad Husaini Suhaimi

Assembly, annotation, and quantification of transcripts from RNA-seq reads of *Shorea* sp. transcriptome followed by differential expression analysis using open source tools.

List of tools:
Trinity (Home · trinityrnaseq/trinityrnaseq Wiki · GitHub)
TransDecoder (Home · TransDecoder/TransDecoder Wiki · GitHub)
NCBI Blast suite (Download BLAST Software and Databases Documentation (nih.gov))
InterProScan (Download - InterPro (ebi.ac.uk))
CDHIT suite (CD-HIT Official Website (ucsd.edu))
Salmon (Salmon - Salmon 1.6.0 documentation)
DESeq2 (Bioconductor - DESeq2)

DOI

protocol ,

Dec 02, 2021

Dec 02, 2021

Transcript Assembly

1   Obtain the assembly statistics

```
$TRINITY_HOME/util/TrinityStats.pl Trinity.fasta > Assembly_stats.txt
```

2

Assemble the reads using Trinity assembler

```
Trinity --seqType fq --max_memory 245G --left reads.PAPER1.left.fq --
right reads.PAPER1.right.fq --CPU 20 --trimmomatic --monitoring >
runPAPER1.log 2>&1 &
```

Removal of Transcript Redundancy

3   Extract the longest open reading frame (ORF) using TransDecoder

```
TransDecoder.LongOrfs -t Trinity.fasta
```

4   Query the extracted ORFs against Uniprot database using blastp

```
blastp -query transdecoder_dir/longest_orfs.pep  -db
uniprot_sprot.fasta  -max_target_seqs 1 -outfmt 6 \
-evalue 1e-5 -num_threads 10 > blastp.outfmt6
```

5 Predict the single best ORF for each transcript

```
TransDecoder.Predict -t Trinity.fasta --single_best_only --
retain_blastp_hits blastp.outfmt6
```

6 Remove highly similar protein sequences using CD-HIT

```
cd-hit -i Trinity_paper1.pep.fasta -o Trinity_paper1_output -T 0 -M 0 -d
0 -c 1 -n 5
```

Transcript Annotation

7 Query the nonredundant nucleotide sequences against *A. thaliana* proteome

```
blastx -query Trinity_reduced_output.fasta -db
Araport11_genes.201606.pep.fasta -evalue 1e-10 -num_threads 33 -
max_taget_seqs 1 -outfmt 6 -out blastx_trasncript_aradb_e10.outfmt6
```

8 Query the nonredundant protein sequences against Pfam, PANTHER, GO, and KEGG databases

```
./interproscan.sh -i Trinity_reduced_output.pep.fasta -t n -appl
Pfam,PANTHER -goterms -iprlookup -pa --cpu 30
```

Transcript Quantification

9 Align and estimate transcript abundance using Salmon

```
$TRINITY_HOME/util/align_and_estimate_abundance.pl --transcript
Trinity.fasta --seqType fq --samples_file samples_description.txt --
est_method salmon --aln_method bowtie2 --thread_count 33 --
prep_reference --trinity_mode --coordsort_bam
```

10   Construct matrices of counts and normalized values

```
$TRINITY_HOME/util/abundance_estimates_to_matrix.pl --est_method
salmon --out_prefix Paper1_comORF_cdhit_only_c1 --
name_sample_by_basedir leaf_1312_rep1/quant.sf
leaf_1312_rep2/quant.sf leaf_1404_rep1/quant.sf
leaf_1404_rep2/quant.sf leaf_1406_rep1/quant.sf
leaf_1406_rep2/quant.sf --gene_trans_map
Paper1_comORF_cdhit_only_c1.fasta.gene_trans_map
```

Differential Expression Analysis

11   Perform pairwise comparison between the time points

```
$TRINITY_HOME/Analysis/DifferentialExpression/run_DE_analysis.pl --
matrix quant.sf/gene.counts.matrix --method DESeq2 --samples_file
samples_description_DESeq.txt --contrasts contrast_DESeq.txt
```

12   Extract the differentially expressed transcripts at $\log_2$ fold change ≥ 1, FDR < 0.05

```
$TRINITY_HOME/Analysis/DifferentialExpression/analyze_diff_expr.pl --
matrix ../ComORF_cdhit_only_c1.isoform.TMM.EXPR.matrix -C 1 -P 0.05
--samples ../samples_description.txt
```