Jun 12, 2024

Transcriptome annotation

DOI
dx.doi.org/10.17504/protocols.io.5qpvok92bl4o/v1

Rafael Rodrigues Ferrari[1], Thiago Mafra Batista[1]

[1]Universidade Federal do Sul da Bahia

bioinfo

Thiago Mafra Batista
Universidade Federal do Sul da Bahia

DOI: **dx.doi.org/10.17504/protocols.io.5qpvok92bl4o/v1**

Protocol Citation: Rafael Rodrigues Ferrari, Thiago Mafra Batista 2024. Transcriptome annotation. **protocols.io**
https://dx.doi.org/10.17504/protocols.io.5qpvok92bl4o/v1

Manuscript citation:
Ferrari RR, Batista TM, Zhou QS, Hilário HO, Orr MC, Luo A, Zhu CD. The Whole Genome of Colletes collaris (Hymenoptera: Colletidae): An Important Step in Comparative Genomics of Cellophane Bees. Genome Biol Evol. 2023 May 5;15(5):evad062. doi: 10.1093/gbe/evad062. PMID: 37075227; PMCID: PMC10159585.

Protocol status: Working
We use this protocol and it's working

Created: June 12, 2024

Last Modified: June 12, 2024

Protocol Integer ID: 101716

Keywords: transcriptome annotation, functional annotation, Trinotate

Funders Acknowledgement:

Suzano Papel e Celulose S/A
Grant ID: 23746.009802/2021-88
Veracel Celulose
Grant ID: 23746.001092/2022-30

Abstract

This protocol provides detailed, step-by-step instructions for students and researchers to annotate transcriptomes. In this tutorial, we will follow the Trinity -> TransDecoder -> Trinotate pipeline, using the SwissProt and Pfam databases for functional annotation of protein-coding transcripts.

FINDIG CODING REGIONS WITHIN TRANSCRIPTS

1 ****TransDecoder (https://github.com/TransDecoder/TransDecoder/wiki)****

Extracting the long open reading frames (ORFs)

Prepare a .pbs file to run the analysis remotely on Sagarana

```
/home/fafinha/bin/TransDecoder-TransDecoder-
v5.5.0/TransDecoder.LongOrfs -t
/home/fafinha/collaris/Trinity_run/assembly/Trinity.fasta
```

Including homology searches as ORF retention criteria

BlastP search

Prepare a .pbs file to run the analysis remotely on Sagarana

```
blastp -query
/home/fafinha/collaris/TransDecoder_run/2_homology_searches/blastp
/Trinity.fasta.transdecoder_dir/longest_orfs.pep \
-db /home/fafinha/collaris/TransDecoder_run/uniprot_sprot.fasta -
max_target_seqs 1 -outfmt 6 -evalue 1e-5 -num_threads 64 \
-out
/home/fafinha/collaris/TransDecoder_run/2_homology_searches/blastp
/blastp_output.fmt6
```

Pfam search

Download the Pfam database (Pfam-A.hmm)

$wget ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.hmm.gz

*Decompress the file

$gzip -d Pfam-A.hmm.gz

Index the database

```
$/programs/hmmer-3.3.2/bin/hmmpress Pfam-A.hmm
```

Prepare a .pbs file to run the analysis remotely on Sagarana

```
/home/fafinha/anaconda3/bin/hmmscan --cpu 64 --domtblout
/home/fafinha/collaris/TransDecoder_run/2_homology_searches/pfam/p
fam.domtblout \
/home/fafinha/bin/pfam/Pfam-A.hmm
/home/fafinha/collaris/TransDecoder_run/2_homology_searches/blastp
/Trinity.fasta.transdecoder_dir/longest_orfs.pep
```

Predicting the likely coding regions

#Run the 'TransDecoder.Predict' script in the same directory where the 'Trinity.fasta.transdecoder_dir' folder is located

Without homology

```
$/home/fafinha/bin/TransDecoder-TransDecoder-
v5.5.0/TransDecoder.Predict -t
/home/fafinha/collaris/Trinity_run/assembly/Trinity.fasta
```

With homology

BlastP

```
$/home/fafinha/bin/TransDecoder-TransDecoder-
v5.5.0/TransDecoder.Predict -t
/home/fafinha/collaris/Trinity_run/assembly/Trinity.fasta \
--retain_blastp_hits
/home/fafinha/collaris/TransDecoder_run/run2/homology/blast/blastp
_output.fmt6
```

Pfam

```
$/home/fafinha/bin/TransDecoder-TransDecoder-
v5.5.0/TransDecoder.Predict -t
/home/fafinha/collaris/Trinity_run/assembly/Trinity.fasta \
--retain_pfam_hits
/home/fafinha/collaris/TransDecoder_run/run2/homology/pfam/pfam.do
mtblout
```

BlastP + Pfam

```
$/home/fafinha/bin/TransDecoder-TransDecoder-
v5.5.0/TransDecoder.Predict -t
/home/fafinha/collaris/Trinity_run/assembly/Trinity.fasta \
--retain_blastp_hits
/home/fafinha/collaris/TransDecoder_run/run2/homology/blast/blastp
_output.fmt6 \
--retain_pfam_hits
/home/fafinha/collaris/TransDecoder_run/run2/homology/pfam/pfam.do
mtblout
```

FUNCTIONAL ANNOTATION

2 #Perform 'FINDIG CODING REGIONS WITHIN TRANSCRIPTS' first

****Trinotate (https://github.com/Trinotate/Trinotate.github.io/blob/master/index.asciidoc) (on kiko)****

Generate databases

```
$/home/thiagomafra/instaladores/Trinotate-Trinotate-
v3.2.2/admin/Build_Trinotate_Boilerplate_SQLite_db.pl Trinotate
```

****Blastn****

Prepare a .pbs file to run the analysis remotely on Sagarana

```
/programs/ncbi-blast-2.10.1+/bin/blastp -query
/home/fafinha/collaris/Trinotate_run/1st_step/Trinity_reduced.fast
a.transdecoder.pep -db \
/home/fafinha/collaris/Trinotate_run/1st_step/uniprot_sprot.fasta
-num_threads 64 -outfmt 6 -evalue 1e-6 \
-out /home/fafinha/collaris/Trinotate_run/2nd_step/blastp.tab
```

Keep best hits only

```
$cat blastp.tab | sort -k1,1 -k12,12nr -k11,11n | sort -k1,1 -u >
blastp_besthits.tab
```

****Blastx****

Prepare a .pbs file to run the analysis remotely on Sagarana

```
/programs/ncbi-blast-2.10.1+/bin/blastx -query
/home/fafinha/collaris/Trinotate_run/1st_step/Trinity_reduced.fast
a \                                                          -db
/home/fafinha/collaris/Trinotate_run/1st_step/uniprot_sprot.fasta
-num_threads 64 -outfmt 6 -evalue 1e-6 \
-out /home/fafinha/collaris/Trinotate_run/2nd_step/blastx.tab
```

Keep best hits only

```
$cat blastx.tab | sort -k1,1 -k12,12nr -k11,11n | sort -k1,1 -u >
blastx_besthits.tab
```

****TMHMM (on kiko)****

```
$/home/thiagomafra/instaladores/tmhmm-2.0c/bin/tmhmm --short <
/home/thiagomafra/collaris/trinotate_run/Trinity_reduced.fasta.tra
nsdecoder.pep \
> /home/thiagomafra/collaris/trinotate_run/fafinha/run2/tmhmm.out
```

****HMMER (on kiko)****

```
$/home/thiagomafra/instaladores/hmmer-3.1b2-linux-intel-
x86_64/binaries/hmmscan --cpu 64 \
--domtblout
/home/thiagomafra/collaris/trinotate_run/fafinha/run2/TrinotatePFA
M.out /home/thiagomafra/collaris/trinotate_run/fafinha/run2/Pfam-
A.hmm \
/home/thiagomafra/collaris/trinotate_run/Trinity_reduced.fasta.tra
nsdecoder.pep >
/home/thiagomafra/collaris/trinotate_run/fafinha/run2/pfam.log
```

****SignalP (on kiko)****

```
$/home/thiagomafra/instaladores/signalp-4.1/signalp -f short -n
/home/thiagomafra/collaris/trinotate_run/fafinha/run2/signalp.out
\
/home/thiagomafra/collaris/trinotate_run/Trinity_reduced.fasta.tra
nsdecoder.pep
```

****RNAmmer (on kiko)****

```
$/home/thiagomafra/instaladores/Trinotate-Trinotate-
v3.2.2/util/rnammer_support/RnammerTranscriptome.pl --
transcriptome
/home/thiagomafra/collaris/trinotate_run/Trinity_reduced.fasta --
path_to_rnammer /home/thiagomafra/instaladores/rnammer/rnammer
```

Generating a .gene_trans_map

```
$/home/thiagomafra/instaladores/trinityrnaseq-
v2.10.0/util/support_scripts/get_Trinity_gene_to_trans_map.pl \
/home/thiagomafra/collaris/trinotate_run/Trinity_reduced.fasta >
Trinity.fasta.gene_trans_map
```

****Populating the .sqlite file****

Loading transcripts and coding regions

```
$/home/thiagomafra/instaladores/Trinotate-Trinotate-
v3.2.2/Trinotate Trinotate.sqlite init --gene_trans_map
./Trinity.fasta.gene_trans_map \
--transcript_fasta
/home/thiagomafra/collaris/trinotate_run/Trinity_reduced.fasta \
--transdecoder_pep
/home/thiagomafra/collaris/trinotate_run/Trinity_reduced.fasta.tra
nsdecoder.pep
```

Loading BLAST homologies

```
$/home/thiagomafra/instaladores/Trinotate-Trinotate-
v3.2.2/Trinotate Trinotate.sqlite LOAD_swissprot_blastp
blastp_besthits.tab

$/home/thiagomafra/instaladores/Trinotate-Trinotate-
v3.2.2/Trinotate Trinotate.sqlite LOAD_swissprot_blastx
blastx_besthits.tab
```

Loading Pfam protein domains

```
$/home/thiagomafra/instaladores/Trinotate-Trinotate-
v3.2.2/Trinotate Trinotate.sqlite LOAD_pfam TrinotatePFAM.out
```

Loading transmembrane domains

```
$/home/thiagomafra/instaladores/Trinotate-Trinotate-v3.2.2/Trinotate Trinotate.sqlite LOAD_tmhmm tmhmm.out
```

Loading signal peptide predictions

```
$/home/thiagomafra/instaladores/Trinotate-Trinotate-v3.2.2/Trinotate Trinotate.sqlite LOAD_signalp signalp.out
```

Loading rRNA gene predictions

```
$/home/thiagomafra/instaladores/Trinotate-Trinotate-v3.2.2/Trinotate Trinotate.sqlite LOAD_rnammer
Trinity_reduced.fasta.rnammer.gff
```

****Generate an output of Trinotate annotation report****

```
$/home/thiagomafra/instaladores/Trinotate-Trinotate-v3.2.2/Trinotate Trinotate.sqlite report >
trinotate_annotation_report.xls
```