Jun 12, 2024

Transcriptome assembly

DOI
dx.doi.org/10.17504/protocols.io.4r3l2q9e4l1y/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.4r3l2q9e4l1y/v1

Protocol Citation: Rafael Rodrigues Ferrari, Thiago Mafra Batista 2024. Transcriptome assembly. **protocols.io https://dx.doi.org/10.17504/protocols.io.4r3l2q9e4l1y/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: 101667

Keywords: Trinity assembler, StringTie, Kallisto, bioinformatics, RNASeq

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

Abstract

This protocol provides detailed, step-by-step instructions for students and researchers to assemble transcriptomes from short reads generated by Illumina technology. In this tutorial, we will check the quality of the sequencing data and align the reads with a bacterial genome database to eliminate potential contamination. We will then use two approaches to assemble the transcripts: de novo assembly and alignment to the reference genome. The transcripts will be quantified and the assemblies evaluated.

SEQUENCING QUALITY CHECK

1 ****FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)****

```
$/programs/FastQC-v0.11.4/fastqc ../Illumina_shortreads.R* -t 64
```

CROSS-SPECIES CONTAMINATION FILTERING

2 ****Magic-BLAST (https://ncbi.github.io/magicblast/)****

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

```
magicblast -db
/databases/ref_prok_rep_genomes_out20/ref_prok_rep_genomes -query
/home/fafinha/collaris/reads/rnaseq_data/R2015100B_L3_139A39.R1.fa
stq \
-query_mate
/home/fafinha/collaris/reads/rnaseq_data/R2015100B_L3_139A39.R2.fa
stq -paired \
-no_discordant -num_threads 48 -infmt fastq -unaligned_fmt sam \
-out_unaligned
/home/fafinha/collaris/mafra/descontamination/fafinha/rnaseq/Illum
ina_rnaseq_unaligned_in_refseq_prok_collaris.sam \
-splice F
&>/home/fafinha/collaris/mafra/descontamination/fafinha/rnaseq/Ill
umina_rnaseq_aligned_in_refseq_prok_collaris.sam
```

Convert output file

```
$samtools view -Sb
Illumina_rnaseq_unaligned_in_refseq_prok_collaris.sam >
Illumina_rnaseq_unaligned_in_refseq_prok_collaris.bam

$samtools sort
Illumina_rnaseq_unaligned_in_refseq_prok_collaris.bam -o
Illumina_rnaseq_unaligned_in_refseq_prok_collaris_sorted.bam

$/programs/samtools-1.12/bin/samtools fastq -1 paired1.fq -2
paired2.fq -n
Illumina_rnaseq_unaligned_in_refseq_prok_collaris_sorted.bam -@24
```

ASSEMBLY

3 /////STRATEGY #1: DE NOVO ASSEMBLY\\\\\

****Trinity (https://github.com/trinityrnaseq/trinityrnaseq/wiki) (on Sagarana)****

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

```
/programs/trinityrnaseq-v2.10.0/Trinity --seqType fq --left
/home/fafinha/collaris/reads/rnaseq_data/R2015100B_L3_139A39.R1.fa
stq \
--right
/home/fafinha/collaris/reads/rnaseq_data/R2015100B_L3_139A39.R2.fa
stq --CPU 48 --max_memory 300G --trimmomatic --output
/tmp/trinity_collaris/
```

****Redundans (https://github.com/lpryszcz/redundans)(on Kiko)****

Filter out isoforms

```
$ ~/instaladores/redundans/redundans.py -r
/home/thiagomafra/collaris/genome.nextpolish.fasta -t 12 --
identity 0.90 \
-f ../Trinity.fasta -i
/data/kiko/thiagomafra/reads/R2015100B_L3_139A39.R1.fastq.gz
/data/kiko/thiagomafra/reads/R2015100B_L3_139A39.R2.fastq.gz 2>&1
\
| tee redundans.log
```

/////STRATEGY #2: REFERENCE-BASED ASSEMBLY (on Kiko)\\\\\\

****HISAT2 (https://github.com/DaehwanKimLab/hisat2)****

Mapping reads against the assembled genome

```
$hisat2-build -p 12
/home/thiagomafra/collaris/genome.nextpolish.fasta
genome.nextpolish.fasta

$hisat2 -t --new-summary -p 8 -x
/home/thiagomafra/collaris/genome.nextpolish.fasta -1
/data/kiko/thiagomafra/reads/R2015100B_L3_139A39.R1.fastq.gz \
-2 /data/kiko/thiagomafra/reads/R2015100B_L3_139A39.R2.fastq.gz -S
rnaseq_mapped_genome.sam 2>&1 | tee hisat2.log
```

****Samtools****

Convert the .SAM to .BAM

```
$samtools view -Sb ./rnaseq_mapped_genome.sam >
./rnaseq_mapped_genome.bam

$samtools sort ./rnaseq_mapped_genome.bam
./rnaseq_mapped_genome.sorted -@ 12
```

****Stringtie (https://github.com/gpertea/stringtie)****

Generate a .GTF from the .BAM

```
$~/instaladores/stringtie-2.2.1.Linux_x86_64/stringtie
/home/thiagomafra/collaris/hisat2_run/fafinha/rnaseq_mapped_genome
.sorted.bam \
-o ./stringtie_collaris_assembled.gtf -p 12 --conservative 2>&1 |
tee stringtie.log
```

****GffRead****

***Generate a transcript .FASTA from the .GTF

```
$gffread ./stringtie_collaris_assembled.gtf -g
/home/thiagomafra/collaris/genome.nextpolish.fasta -w
./stringtie_transcripts.fasta
```

****Redundans****

Filter out isoforms

```
$ ~/instaladores/redundans/redundans.py -r
/home/thiagomafra/collaris/genome.nextpolish.fasta -t 12 --
identity 0.90 \
-f ../stringtie_transcripts.fasta -i
/data/kiko/thiagomafra/reads/R2015100B_L3_139A39.R1.fastq.gz \
/data/kiko/thiagomafra/reads/R2015100B_L3_139A39.R2.fastq.gz 2>&1
| tee redundans.log
```

TRANSCRIPT QUANTIFICATION

4 ****Trinity (using Kallisto)****

Estimating transcript abundance (on headnode)

#Use trimmed reads (.fastq.P.qtrim) if --trimmomatic was used in the assembly (the 'index' and 'quantification' functions of Kallisto are performed here)

```
$/programs/trinityrnaseq-
v2.10.0/util/align_and_estimate_abundance.pl --transcripts
/home/fafinha/collaris/Trinity_run/assembly/Trinity.fasta --
seqType fq \
--est_method kallisto --left
/home/fafinha/collaris/Trinity_run/assembly/R2015100B_L3_139A39.R1
.fastq.P.qtrim \
--right
/home/fafinha/collaris/Trinity_run/assembly/R2015100B_L3_139A39.R2
.fastq.P.qtrim --thread_count 24 --trinity_mode --prep_reference \
--output_dir
/home/fafinha/collaris/Trinity_run/quantification/RSEM/transcript_
abundance_estimation/prep_quant
```

Generate a matrix of expression values (on headnode)

#A matrix of TMM-normalized expression values (*.TMM.EXPR.matrix) will not be generated if a single .tsv file (=single sample) is input

```
$/programs/trinityrnaseq-
v2.10.0/util/abundance_estimates_to_matrix.pl --est_method
kallisto --gene_trans_map none --cross_sample_norm TMM \
/home/fafinha/collaris/Trinity_run/quality_check/contig_ExN50/run2
/Kallisto_matrix/transcript_level/abundance.tsv
```

Counting expressed transcripts

Couting transcripts above a threshold (on headnode)

```
$/programs/trinityrnaseq-
v2.10.0/util/misc/count_matrix_features_given_MIN_TPM_threshold.pl
\
/home/fafinha/collaris/Trinity_run/quantification/Kallisto/express
ed_transcript_counting/run2/kallisto.isoform.TPM.not_cross_norm |
\
tee isoform_matrix.TPM.not_cross_norm.counts_by_min_TPM
```

Plotting the results graphically on R (on my pC)

```
$data =
read.table("/mnt/c/Users/Rafael/Documents/POSTDOCs/IZCAS/RESEARCH/
Genomics/Colletes_collaris/Trinity/isoform_matrix.TPM.not_cross_no
rm.counts_by_min_TPM", header=T)

$plot(data, xlim=c(-100,0), ylim=c(0,100000), t='b')
```

****Trinity (using RSEM)****

Estimating transcript abundance

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

```
$/programs/trinityrnaseq-
v2.10.0/util/align_and_estimate_abundance.pl \
--transcripts
/home/fafinha/collaris/Trinity_run/assembly_without_trimmomatic/Tr
inity.fasta --seqType fq --est_method RSEM --aln_method bowtie2 \
--trinity_mode --prep_reference --left
/home/fafinha/collaris/reads/rnaseq_data/R2015100B_L3_139A39.R1.fa
stq \
--right
/home/fafinha/collaris/reads/rnaseq_data/R2015100B_L3_139A39.R2.fa
stq --SS_lib_type RF --thread_count 64 --output_dir \
/home/fafinha/collaris/Trinity_run/quantification/RSEM/transcript_
abundance_estimation/prep_quant
```

Generate a matrix of expression values (on headnode)

```
$/programs/trinityrnaseq-
v2.10.0/util/abundance_estimates_to_matrix.pl --est_method RSEM --
gene_trans_map none --cross_sample_norm TMM \
/home/fafinha/collaris/Trinity_run/quantification/RSEM/matrix_gene
ration/RSEM.isoforms.results
```

Counting expressed transcripts

Couting transcripts above a threshold (on headnode)

```
$/programs/trinityrnaseq-
v2.10.0/util/misc/count_matrix_features_given_MIN_TPM_threshold.pl
\
/home/fafinha/collaris/Trinity_run/quantification/RSEM/expressed_t
ranscript_counting/RSEM.isoform.TPM.not_cross_norm | \
tee isoform_matrix.TPM.not_cross_norm.counts_by_min_TPM
```

Plotting the results graphically on R (on my PC)

```
$data =
read.table("/mnt/c/Users/Rafael/Documents/POSTDOCs/IZCAS/RESEARCH/
Genomics/Colletes_collaris/Trinity/RSEM/isoform_matrix.TPM.not_cro
ss_norm.counts_by_min_TPM", header=T)

$plot(data, xlim=c(-100,0), ylim=c(0,100000), t='b')
```

ASSEMBLY QUALITY ASSESSMENT

5 ****Contig Nx statistics****

Trinity

```
$/programs/trinityrnaseq-v2.10.0/util/TrinityStats.pl
/home/fafinha/collaris/Trinity_run/quality_check/Trinity.fasta >
Nx_stats.txt
```

Estimate transcript abundance (using a built-in Kallisto)

Prepare the reference + run the estimation using (on headnode)

#Use trimmed reads (.fastq.P.qtrim) if --trimmomatic was used in the assembly (the 'index' and 'quantification' functions of Kallisto are performed here)

```
$/programs/trinityrnaseq-
v2.10.0/util/align_and_estimate_abundance.pl --transcripts
/home/fafinha/collaris/Trinity_run/assembly/Trinity.fasta --
seqType fq \
--est_method kallisto --left
/home/fafinha/collaris/Trinity_run/assembly/R2015100B_L3_139A39.R1
.fastq.P.qtrim \
--right
/home/fafinha/collaris/Trinity_run/assembly/R2015100B_L3_139A39.R2
.fastq.P.qtrim --thread_count 24 --trinity_mode --prep_reference \
--output_dir
/home/fafinha/collaris/Trinity_run/quality_check/contig_ExN50/run2
/Kallisto_prep_quant/prep_quant
```

Generate a matrix of expression values (on headnode)

#A matrix of TMM-normalized expression values (*.TMM.EXPR.matrix) will not be generated if a single .tsv file (=single sample) is input

```
$/programs/trinityrnaseq-
v2.10.0/util/abundance_estimates_to_matrix.pl --est_method
kallisto --gene_trans_map none --cross_sample_norm TMM \
/home/fafinha/collaris/Trinity_run/quality_check/contig_ExN50/run2
/Kallisto_matrix/transcript_level/abundance.tsv
```

Calculate ExN50 stats (on headnode)

```
$/bkp/programs/trinityrnaseq-
v2.12.0/util/misc/contig_ExN50_statistic.pl \
/home/fafinha/collaris/Trinity_run/quality_check/contig_ExN50/run2
/Kallisto_stats/transcript_level/kallisto.isoform.TPM.not_cross_no
rm \
/home/fafinha/collaris/Trinity_run/assembly/Trinity.fasta | tee
ExN50.stats
```

Plot the Ex value against the ExN50 value (on my PC)

```
$/home/rafael/programs/trinityrnaseq/util/misc/plot_ExN50_statisti
c.Rscript  /home/rafael/collaris/ExN50.stats
```

****BUSCO (https://busco.ezlab.org/)****

Run BUSCO (using a docker)

```
$docker run --rm -e USERID=$UID -u $UID -v
/home/rferrari/:/home/rferrari/ -w
/home/rferrari/projetos/collaris/BUSCO_run/transcriptome/fafinha
ezlabgva/busco:v5.2.2_cv1 busco -i
/home/rferrari/projetos/collaris/data/Trinity_reduced.fsa -l
hymenoptera_odb10 --augustus_species Apis_mellifera -o run_final -
m tran -c 10
```

FILTERING

6 ****Trinity****

Kallisto's results

```
$/programs/trinityrnaseq-
v2.10.0/util/filter_low_expr_transcripts.pl --transcripts
/home/fafinha/collaris/Trinity_run/assembly/Trinity.fasta \
--matrix
/home/fafinha/collaris/Trinity_run/filtering/Kallisto/kallisto.iso
form.TPM.not_cross_norm --trinity_mode --min_expr_any 1 \
> filtered_transcripts_min_exp_1.fasta

$/programs/trinityrnaseq-
v2.10.0/util/filter_low_expr_transcripts.pl --transcripts
/home/fafinha/collaris/Trinity_run/assembly/Trinity.fasta \
--matrix
/home/fafinha/collaris/Trinity_run/filtering/Kallisto/kallisto.iso
form.TPM.not_cross_norm --trinity_mode --min_expr_any 1 --
highest_iso_only \
> filtered_transcripts_min_exp_1_highest.fasta
```

RSEM's results

```
$/programs/trinityrnaseq-
v2.10.0/util/filter_low_expr_transcripts.pl \
--transcripts
/home/fafinha/collaris/Trinity_run/assembly_without_trimmomatic/Tr
inity.fasta \
--matrix
/home/fafinha/collaris/Trinity_run/filtering/RSEM/RSEM.isoform.TPM
.not_cross_norm --trinity_mode --min_expr_any 1 /
> filtered_transcripts_min_exp_1.fasta

$/programs/trinityrnaseq-
v2.10.0/util/filter_low_expr_transcripts.pl \
--transcripts
/home/fafinha/collaris/Trinity_run/assembly_without_trimmomatic/Tr
inity.fasta \
--matrix
/home/fafinha/collaris/Trinity_run/filtering/RSEM/RSEM.isoform.TPM
.not_cross_norm --trinity_mode --min_expr_any 1 --highest_iso_only
/
> filtered_transcripts_min_exp_1_highest.fasta
```