Feb 19, 2025

Bulk RNA sequencing analysis

DOI

dx.doi.org/10.17504/protocols.io.yxmvm2m55g3p/v1

Raquel Garza[1]

[1]Laboratory of Molecular Neurogenetics, Department of Experimental Medical Science, Wallenberg Neuroscience Center and Lund Stem Cell Center, BMC A11, Lund University, 221 84 Lund, Sweden.

ASAP Collaborative Rese…

Jakobsson

Raquel Garza
Lund University

DOI: dx.doi.org/10.17504/protocols.io.yxmvm2m55g3p/v1

Protocol Citation: Raquel Garza 2025. Bulk RNA sequencing analysis. **protocols.io**
https://dx.doi.org/10.17504/protocols.io.yxmvm2m55g3p/v1

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

Created: March 03, 2023

Last Modified: February 19, 2025

Protocol Integer ID: 78060

Keywords: ASAPCRN

Funders Acknowledgements:

Aligning Science Across Parkinson's through the Michael J. Fox Foundation for Parkinson's Research
Grant ID: ASAP-000520
Swedish Research Council
Grant ID: 2018-02694
Swedish Brain Foundation
Grant ID: FO2019-0098
Cancerfonden
Grant ID: 190326
Barncancerfonden
Grant ID: PR2017-0053
NIHR Cambridge Biomedical Research Centre
Grant ID: NIHR203312
Swedish Society for Medical Research
Grant ID: S19-0100
National Institutes of Health
Grant ID: HG002385
Swedish Research Council
Grant ID: 2021-03494
Swedish Research Council
Grant ID: 2020-01660

Abstract

This protocol describes the steps for the bioinformatical analysis of bulk RNA sequencing with a focus on evolutionary young L1s.

TE subfamily quantification

1 Reads were mapped using STAR aligner (version 2.6.0c; RRID:SCR_004463) with an hg38 index and gencode annotation as the guide GTF (--sjdbGTFfile), allowing for a maximum of 100 multi mapping loci (--outFilterMultimapNmax 100) and 200 anchors (--winAnchorMultimapNmax). The rest of the parameters affecting the mapping were left in default as for version 2.6.0c.

2 The TE subfamily quantification was performed using TEcount from the TEToolkit (version 2.0.3; RRID:SCR_023208) in mode multi (--mode). Gencode annotation was used as the input gene GTF (--GTF), and the provided hg38 GTF file from the author's web server as the TE GTF (--TE).

TE quantification

3 Reads were mapped using STAR aligner (version 2.6.0c; RRID:SCR_004463) with an hg38 index and gencode annotation as the guide GTF (--sjdbGTFfile). To quantify only confident alignments, we allowed a single mapping locus (--outFilterMultimapNmax 1) and a ratio of mismatches to the mapped length of 0.03 (--outFilterMismatchNoverLmax).

4 We used SAMtools view (version 1.9; RRID:SCR_002105) to keep only the alignments in forward transcription. We separated alignments of the second pair mate if they mapped to the forward strand (-f 128 -F 16) and alignments of the first pair mate if they map to the reverse strand (-f 80).

5 To keep the reverse transcription, we kept alignments of the second pair mate if they mapped to the reverse strand (-f 144) and alignments of the first pair mate if they mapped to the forward strand (-f 64 -F 16).

6 Both BAM files were then quantified using featureCounts from the subread package (version 1.6.3; RRID:SCR_012919) forcing strandness to the features being quantified (-s 2). For consistency (and to avoid quantifying over simple repeats, small RNAs and low-complexity regions) we input the same curated hg38 GTF file provided by the TEtranscripts authors.

Gene quantification

7 Reads were mapped using STAR aligner (version 2.6.0c; RRID:SCR_004463) with an hg38 index and gencode annotation as the guide GTF (--sjdbGTFfile), no other parameters were modified (default values for --outFilterMultimapNmax, --outFilterMismatchNoverLmax, and --winAnchorMultimapNmax).

8 Genes were quantified using featureCounts from the subread package (version 1.6.3; RRID:SCR_012919) forcing strandness (-s 2) to quantify by gene_id (-g) from the GTF of gencode annotation.

Differential expression analysis

9 Differential gene expression analysis was performed using DESeq2 (version 1.28.1; RRID:SCR_015687) with the read count matrix from featureCounts (Subread version 1.6.3; RRID:SCR_012919) as input. Fold changes were shrunk using DESeq2:: lfcShrink.

10 We performed differential-TE subfamily expression analysis using DESeq2 (version 1.28.1; RRID:SCR_015687) with the read count matrix from TEcount (version 2.0.3; RRID:SCR_023208) using only the TE subfamilies entries. Fold changes were shrinked using DESeq2:: lfcShrink.

11 For visualization, using the gene DESeq2 object (see step #9) we normalized the TE subfamily counts by dividing the read count matrix by the sample distances (sizeFactor) as calculated by DESeq2 with the quantification of genes without multimapping reads (see section "Gene quantification").

12 To normalize uniquely mapped read counts per strand (see section "TE quantification"), we divided the read count matrix by the sample distances (sizeFactor) as calculated by DESeq2 (version 1.28.1; RRID:SCR_015687) with the quantification of genes without multimapping reads (see section "Gene quantification").

Comparison between sense and antisense transcription over TEs (optional)

13 To normalize uniquely mapped read counts per strand (see section "TE quantification"), we divided the read count matrix by the sample distances (sizeFactor) as calculated by DESeq2 (version 1.28.1; RRID:SCR_015687) with the quantification of genes without multimapping reads (see section "Gene quantification").

Comparing the ratio of detected elements of all L1s (optional)

14 Once the counts of individual elements were normalized by the gene sizeFactors (see "Comparison between sense and antisense transcription over TEs"section), we defined a "detected" element as an element with a mean >10 normalized counts in the group of samples of interest. The total number of elements is the number of elements from a particular subfamily annotated in the GTF file that was input to featureCounts (version 1.6.3; RRID:SCR_012919).

Transcription over evolutionary young L1s elements in bulk datasets (optional)

15 The BED file version of TEcount's GTF file was used to create BED files containing all L1HS, L1PA2, L1PA3, and L1PA4 elements longer than 6 kbp (full length). These BED files were then split by the strand of the element.

16 Create bigwig files using deeptools BamCoverage (version 2.5.4; RRID:SCR_016366) (using --normalizeUsingRPKM) of the uniquely mapped BAM files

17 Create four matrices per dataset using the deeptools' (version 2.5.4; RRID:SCR_016366) computeMatrix function:
1. Elements annotated in the positive strand using only the bigwig files with forward transcription (transcription in sense of the element)
2. Elements annotated in the reverse strand using only bigwig files with reverse transcription (transcription in sense of the element)
3 and 4. Same two using the antisense transcription bigwig files (e.g. elements annotated in the positive strand using reverse transcription bigwig files).

18 Concatenate the matrices of transcription in sense of the elements together using rbind from computeMatrixOperations (Repeat this step with the antisense matrices)

19 Heatmaps were plotted using plotHeatmap, setting missing values to white (--missingDataColor white), and colorMap to Blues (sense) or Reds (antisense).

20 To investigate if the expressed elements contained an intact YY1 binding site, we extracted the relevant sequences using getfasta from bedtools (version 2.30.0; RRID:SCR_006646) using GRCh38.p13 as input fasta (-fi) and forcing strandness (-s).

21 We quantified the number of elements with an exact match to the YY1 binding motif (CAAGATGGCCG) in the first 100 bp of the element