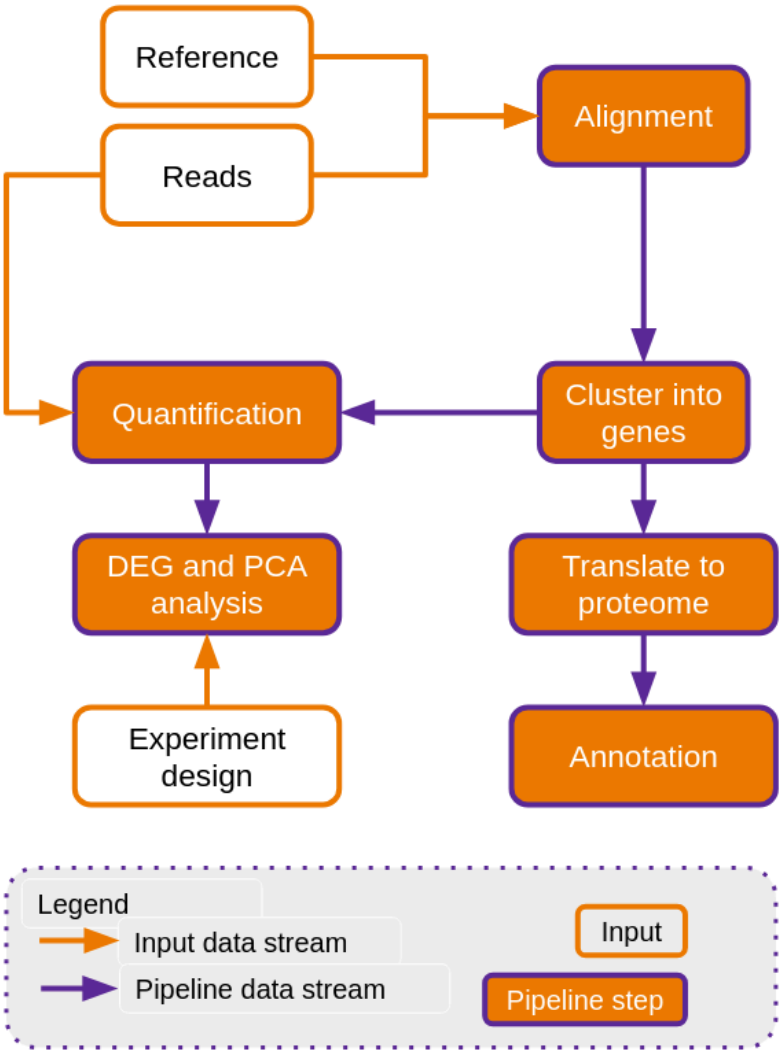

Manual Transcriptome Analysis Pipeline

Author: Kaiden R. Sewradj
Date of last update: 11/03/2025

Supervised by: Quentin Helleu, Ellen Carbo, Jean-Paul Concordet & Anne de Cian

0. Table of Contents

1. Requirements

This pipeline was designed to work with the Slurm workload manager and relies on this system to dispatch jobs. Other requirements are:

- eggnog-mapper v2.1.12 (Cantalapiedra *et al.*, 2021)
- R v4.4.1 (R Core Team, 2023)
    - DESeq2 v1.44.0 (Love *et al.*, 2014)
    - Factoextra v1.0.7 (Kassambara & Mundt, 2020)
    - ggplot2 v3.5.1 (Wickham, 2016)
    - stringr (Wickham, 2023)
    - tximport (Soneson *et al.*, 2015)
- Salmon v1.10.2 (Patro *et al.*, 2017)
- samtools v1.21 (Danecek *et al.*, 2021)
- Singularity v1.1.7 (Kurtzer *et al.*, 2017)
    - TransDecoder v5.7.1 (TransDecoder, n.d.)
        - Image at $DIR/bin/transdecoder_5.7.1.sif
    - Trinity v2.15.1 (Grabherr *et al.*, 2011)
        - Image at $DIR/bin/trinityrnaseq_2.15.1.sif
- STAR v2.7.11a (Dobin *et al.*, 2012)

Except for tools listed under singularity and R packages, all software is loaded with Modules v4.6.1 as follows: `module <tool>\<version number>`. All R packages are expected to be pre-installed.

2. Usage

2.1. Clone repo

```
git clone https://github.com/Kaiden-exe/transcriptome-analysis.git
```

2.1.1. What is in the repo?

- `scripts` [directory]
  - All scripts are here
- `config.sh`
  - Template configuration file.
- `transcriptome_analysis.sh`
  - Main script file that dispatches the Slurm jobs
- `summary_transcriptome_analysis.sh`
  - Run with: `bash summary_transcriptome_analysis.sh config.sh`
  - Will look at output directories in config.sh and give a summary of the results of the whole pipeline.
  - Run AFTER pipeline is done.
- Manual Transcriptome Analysis Pipeline.pdf
  - You're looking at it 😜

2.2. Prepare input files

2.2.1. manifest.tsv

This file has no header and consists of four columns in this order: *ID, reads1, reads2, speciesID*. For single-read samples, leave the reads2 column empty. The IDs needs to be unique per read (pair). The speciesID needs to be consistent across all input files and will be used to name files. Therefore it is advised to keep it short, and avoid special characters and spaces. It is also advised to give absolute paths for read files.

2.2.2. references.tsv

This file has no header and consists of two columns in the following order: *speciesID, reference*. The speciesID column must be in correspondence with the speciesID column in **manifest.tsv**, but the species do not need to appear in the same order. The speciesID column must have unique IDs. Lastly, it is advised to give absolute paths to reference files.

2.2.3. conditions.tsv

This file is WITH a header for three columns in the following order: *ID, condition, batch. The ID* column needs to be in correspondence with the ID column of **manifest.tsv**. Control condition must be 0Gy. If batches are unknown, fill in the same thing for each sample. This will cause batch effects to be ignored during the analysis.

2.3. Set up config file

2.3.1. Important variables to change as necessary

- `DIR`
  - Project directory, used to find the images of TransDecoder and Trinity.
- `manifest=manifest.tsv`
  - Change to your manifest.tsv file.
- `referenceTSV=references.tsv`
  - Change to your references.tsv file.
- `conditions=conditions.tsv`
  - Change to your conditions.tsv file.
- `DAT`
  - Change to directory with eggNOG databases.

2.3.2. [Optional] Change output directories

None of these directories need to be created before running the pipeline.

- `alignmentDir=alignment`
  - This is the output directory of STAR, meaning that alignments and the index of the reference genomes will be here.
- `trinityOut=trinity_output`
  - The transcriptomes created by trinity will end up here.
- `salmonOut=salmon_output`
  - The indexed transcriptomes and quantification per sample created by Salmon will be here.
- `DEGout=DEG_output`
  - Plots (PCA & volcano) and DEG tables will be here.
- `transdecoderOut=transdecoder_output`
  - Proteomes made by TransDecoder will be here.
- `eggnogOut=eggnog_output`
  - Annotation files created by eggNOG-mapper will be here.

2.4. Run pipeline

2.4.1. Running for the first time

After following previous steps, run the pipeline with:
`bash transcriptome_analysis.sh -c config.sh`

2.4.2. Running with new data

To add data of a new species, create a new manifest.tsv file and add necessary rows to conditions.tsv and references.tsv. Output directories can stay the same. Species not mentioned in manifest.tsv will NOT be overwritten, even if they are present in conditions.tsv or references.tsv.

3. Output

3.1. logfiles

All stdout and stderror logfiles are in here. Filenames include batch number and array number if applicable.

3.2. alignmentDir

This directory contains the indices of each reference, a sorted BAM file per ID, and a merged BAM file per species.

```
alignment
|--- reference_indices
|--- speciesID
|     |--- ID1Aligned.sortedByCoord.out.bam
|     …
|     |--- ID2Aligned.sortedByCoord.out.bam
|     …
|--- speciesID_merged.bam
```

3.3. trinityOut

The transcriptomes as created by Trinity are here. The original output is prefixed by trinity_. The files without this prefix are the ones used for creating the proteome. The unprefixed files have a prefix of speciesID| in front of all isoform names. This was done so these isoforms can not later be confused with isoforms from another species.

```
trinity_output
|--- speciesID
|     |--- speciesID.Trinity-GG.fasta
|     |--- speciesID.Trinity-GG.fasta.gene_trans_map
|     |--- trinity_speciesID.Trinity-GG.fasta
|     |--- trinity_speciesID.Trinity-GG.fasta.gene_trans_map
```

3.4. salmonOut

The indexation of the transcriptome can be found here along with the quantification per ID in manifest.tsv.

```
salmon_output
|--- ID1
|     ...
|     |--- quant.sf
|--- ID2
|     ...
|     |--- quant.sf
|--- speciesID_indices
|     ...
```

3.5. DEGout

This directory contains the DESeq2 results, a PCA plot and a volcano plot for each species. The analysis is saved in the .RData file so graphs can be more easily remade and tailored to taste.

```
DEG_output
|--- speciesID_DESeq_results.txt
|--- speciesID_PCA_Plot.png
|--- speciesID_results.RData
|--- speciesID_VolcanoPlot.png
```

3.6. transdecoderOut

The proteome of each species can be found here. All other TransDecoder files are also here.

```
transdecoder_output
|--- speciesID
|     ...
|     |--- speciesID.Trinity-GG.fasta.transdecoder.pep
```

3.7. eggnogOut

The annotations from eggNOG-mapper are here along with the other files produced by this tool.

```
eggnog_output
|--- speciesID
|     |--- speciesID.emapper.annotations
|     ...
```

4. Get a summary

The pipeline comes with a companion script that generates a summary file with the following information: amount of species analysed, number of genes and isoforms, mapping rates of quantification, amount of DEGs, and amount of annotated genes.

To get this summary run:

```
bash summary_transcriptome_analysis.sh config.sh
```

5. References

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution,* 38(12), 5825–5829. DOI: 10.1093/molbev/msab293

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2). DOI: 10.1093/gigascience/giab008

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15–21. DOI: 10.1093/bioinformatics/bts635

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman., N. & Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology,* 29(7), 644–652. DOI: 10.1038/nbt.1883

Kassambara, A. & Mundt, F. (2020). Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R Package Version 1.0.7. https://CRAN.R-project.org/package=factoextra

Kurtzer G.M., Sochat V. & Bauer M.W. (2017). Singularity: Scientific containers for mobility of compute. *PLoS ONE* 12(5): e0177459. DOI: 10.1371/journal.pone.0177459

Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550 (2014). DOI: 10.1186/s13059-014-0550-8

Patro R., Duggal G., Love M.I., Irizarry R.A. & Kingsford C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14, 417. DOI: 10.1038/nmeth.4197

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Soneson, C., Love, M.I. & Robinson, M.D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4, DOI: 10.12688/f1000research.7563.1

TransDecoder. (n.d.). GitHub - TransDecoder/TransDecoder: TransDecoder source. GitHub. https://github.com/TransDecoder/TransDecoder

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

Wickham, H. (2023). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.5.1, https://github.com/tidyverse/stringr, https://stringr.tidyverse.org.