



Nov 27, 2024

Customisable differential TF activity estimation for *C. elegans* with the CelEst gene regulatory network and the decoupler package in R



DOI

<https://dx.doi.org/10.17504/protocols.io.81wgbze3ogpk/v1>

Marcos Francisco Perez¹

¹IBMB-CSIC



Marcos Francisco Perez

IBMB-CSIC

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.81wgbze3ogpk/v1>

Protocol Citation: Marcos Francisco Perez 2024. Customisable differential TF activity estimation for *C. elegans* with the CelEst gene regulatory network and the decoupler package in R. **protocols.io**

<https://dx.doi.org/10.17504/protocols.io.81wgbze3ogpk/v1>

Manuscript citation:

Perez MF, 2024. CelEst: a unified gene regulatory network for estimating transcription factor activities in *C. elegans*. GENETICS



License: This is an open access protocol distributed under the terms of the **[Creative Commons Attribution License](#)**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: September 24, 2024

Last Modified: November 27, 2024

Protocol Integer ID: 108233

Keywords: transcription factor, *C. elegans*, activity, TF activity estimation, RNA-seq, gene ontology, celest gene regulatory network, unified gene regulatory network, differential gene expression, differential gene expression of target, customisable differential tf activity estimation, transcription factor activity, gene expression, celest gene, regulatory network, unified gene, gene, decoupler package

Funders Acknowledgements:

Ministerio de Economía y Competividad (Spain)

Grant ID: RYC2021-034496-I

Abstract

Estimating differential TF activity from differential gene expression of targets can be performed in R using the decoupleR package and the CelEsT gene regulatory network. This protocol is a step-by-step guide to using the customisable R script 'File_S1.txt' found at github.com/IBMB-MFP/CelEsT-MS.

Perez MF, 2024. **CelEst: a unified gene regulatory network for estimating transcription factor activities in *C. elegans***. GENETICS

C. elegans TF activity estimation with customisable R script (File S1 - Customisable_R_script.R at github.com/IBMB-MFP/CelEsT-MS/Genetics)

- 1 Define your working directory by replacing the path in the `'dir.create("path/to/file")'` function in the script. If the desired directory does not exist, it will be created.
- 2 Run the code sections `##### DEFINE FUNCTIONS #####` and `##### LOAD PACKAGES & FUNCTIONS #####`. Packages will be installed from the appropriate locations if they are not found; the first time this may take a while if many packages are to be installed.
- 3 Download the CelEsT network Excel file (manuscript Table S3; or alternative networks orthCelEsT or maxCelEsT if desired) and TF annotations (Table S1) from github.com/CelEsT-MS/SUPP
- 4 Set the path to the file in the script and run the code to import the annotations and network files.
- 5 Choose your input data from your transcriptomic study:

If you choose to start your analysis from raw counts, 'uncomment' the relevant code by removing the `"#"` symbols at the start of the line, input your path and run the code to import your data.

If you choose to start your analysis from an existing differential expression (DE) analysis, import the file with the DE stats and gene identifiers in the same way. The first column should include gene identifiers, the second column should include the statistics.
- 6 It is critical that gene identifiers can be identified to match the network.

Run the code: `'print(head(wormRef::Cel_genes))'` to see different gene identifiers for *C. elegans* and their associated column heading.

If necessary, change the identifier to match the one used in your data with `'myGenelDformat <- "sequence_name"'`. Replace the string `"sequence_name"` with the column heading for the identifier used in your data.

e.g. if you have used WormBase Gene IDs in your data, change the code and run it as:

`'myGenelDformat <- "wb_id"'`
- 7 If you wish to perform DE analysis from raw counts, refer to the sections below as appropriate depending on whether you wish to correct for developmental age





differences (Step 8) or not (Step 12). If you are using pre-existing DE stats, go directly to Step 15.

PERFORM DE ANALYSIS WITH AGE CORRECTION WITH RAPToR

- 8 Uncomment the code section **#### PERFORM DE ANALYSIS WITH AGE CORRECTION WITH RAPToR ####**. In R studio, a whole section of code can be commented/uncommented with the Ctrl/Command + Shift + C shortcut.

Software

R Studio Desktop

NAME

The R Studio, Inc.

DEVELOPER

<https://www.rstudio.com/products/RStudio/>

SOURCE LINK

- 9 Developmental age correction relies on first estimating each sample's age by comparison to a reference time series. Several reference time series are compiled in the wormRef package but for good results you must choose one that covers an appropriate developmental window for your samples.

Run the code 'plot_refs("wormRef")' to see a graphic that shows the time window covered by each reference.

When you have chosen the appropriate reference for your data, modify and run the code below with the name of your chosen reference from the wormRef package; e.g.:

```
'myRef <- "Cel_larv_YA"
```

Following this, run the code including the 'prepare_refdata()' function.

- 10 To perform the DE analysis, you must specify which are your control and treatment samples in your study. These should match column names in your gene read count data. They must be inputted as a string vector, with each sample in quotation marks and separated by commas. e.g.:

```
'control_samples <- c("ctrl1", "ctrl2", "ctrl3")'
```

- 11 Run the remaining code in the section. The output object 'DE_stats' will serve downstream for TF activity estimation. Go to Step 15. If you wish to see the age estimates

for your samples, inspect the object 'temp_sample_ae', which includes a table of estimated ages with confidence intervals.

PERFORM DE ANALYSIS FROM RAW COUNTS

- 12 Uncomment the code section `#### PERFORM DE ANALYSIS FROM RAW COUNTS ####`. In R studio, a whole section of code can be commented/uncommented with the Ctrl/Command + Shift + C shortcut.
- 13 To perform the DE analysis, you must specify which are your control and treatment samples in your study as specified in Step 10.
- 14 Run the remaining code in the section. The output object 'DE_stats' will serve downstream for TF activity estimation. Go to Step 15.

FORMAT DIFFERENTIAL EXPRESSION STATS FOR DECOUPLER

- 15 Whether you have uploaded pre-existing DE stats or have just computed DE stats with or without developmental age correction, the code section `#### FORMAT DIFFERENTIAL EXPRESSION STATS FOR DECOUPLER ####` ensures that the format is correct to allow TF activity estimation.

Run the code in this section. This will remove the trailing point and number, if it exists, to convert transcript identifiers into gene sequence identifiers. It will also ensure that the gene identifiers used match the network.

ESTIMATE TF ACTIVITY WITH DECOUPLER

- 16 The code section `#### ESTIMATE TF ACTIVITY WITH DECOUPLER ####` will use the DE stats computed or provided and output an estimation of differential TF activity. The argument 'statistics' details the statistical method to be used. It is recommended to use the multivariate linear model ("mlm") method, as this displayed the best performance in Perez 2024 and also is the fastest method - if so, leave the code as it is.

However if you wish to use other statistical methods, modify the argument as appropriate. Multiple methods can be specified as a vector or strings. If you use multiple methods, you can calculate a consensus score which integrates the results by specifying 'consensus_score = TRUE'. E.g:

```
DEdata_decouple <- decoupleR::decouple(  
  mat = DE_stats[, 1, drop = FALSE],  
  network = CelEsT_GRN,  
  .source = "source",
```



```
.target = "target",  
statistics = c("mlm", "ulm", "wsum"),  
args = list(mlm = list(.mor = "weight")),  
consensus_score = TRUE  
)
```

- 17 The remaining code in this section annotates the output with additional information and saves the output as a white-space separated text file in an 'output' subdirectory of the folder specified. If you wish for the table to be saved elsewhere, modify the file path in the 'write.table()' function.

Expected result

The saved table will contain the TF activity estimates from your DE analysis, together with estimated p-values and useful annotations (other gene identifiers and putative TF mode-of-regulation).

Protocol references

If you use *CelEst* in a publication, **don't forget to cite:**

Perez MF, 2024. CelEst: a unified gene regulatory network for estimating transcription factor activities in *C. elegans*. *Genetics*

&

Badia-i-Mompel, P et al. (2022) decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances*, **doi: 10.1093/bioadv/vbac016**

If you use this script for your DE analysis, **cite:**

Love, MI et al. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **doi: 10.1186/s13059-014-0550-8**

If you choose to control for developmental age in your analysis, **cite:**

Bulteau, R & Francesconi, M. (2022) Real age prediction from the transcriptome with RAPToR. *Nature Methods*, **doi: 10.1038/s41592-022-01450-0**