

README

Markos et al 2023: Cell type and regulatory analysis in amphioxus illuminates evolutionary origin of the vertebrate head

Introduction

This repository contains scripts and input data used for analysis of single cell RNA-Seq data presented in Markos et al 2023^[1]. Data consist of four 10X datasets, each representing selected stage of *Amphioxus* (*Branchiostoma floridae*) embryonic development. Aim of the analyses is to annotate the data and investigate developmental trajectories (transitions) across the identified celltypes and stages (timepoints) according to the hypotheses presented in the paper.

Important: Purpose of the repository is to serve as extended data accompanying the manuscript, we do not wish to update the code except for requirements raised during the review process.

Important: To reduce compressed repository size for submission, input files were omitted and labeled ON DEMAND in the content listing below. To retrieve these files, please contact corresponding author.

Software Prerequisites

Main body of the analyses is written in R with addition of some Python. The annotation part was run under R 4.2.1 and Seurat 4.3.0 The transitions part requires installation of some extra R and Python packages. We provide `renv.lock` file to install all necessary R packages and `environment.yml` file to recreate conda environment for the Python part. Please refer to [renv](#) and [conda](#) manuals. Code is tested under Ubuntu 20.04.5 LTS. Installation of R, all R packages and conda environment should take up to 60 minutes, depending on hardware performance and download speed.

Analyses description

Canonical steps of Seurat workflow were used to load, filter, normalize and cluster expression matrices of individual timepoints (provided here). Clusters were annotated in supervised manner based on known sets of markers, see the paper for details. Scripts describing timepoints analyses generate graphical output and are meant to be run interactively using e.g. RStudio or other R compatible IDE. They also output final Seurat objects in RDS format, which are used for downstream transitions analysis. We present the transitions analysis scripts in [R Markdown format](#) (with exception of some technical steps) for smoother readability and execution with e.g. `rmarkdown::render('01_Integration.Rmd')`.

All script files have their description in the header with some hints where appropriate. Paths are set as relative, meaning the code can be run from the downloaded repository directly, with R script's working directory being set to the same path as the source files.

To save computational time, we provide precomputed cell-by-cell transition probabilities matrix, which serves as an input to the last part of transitions workflow. This is used as default, to recompute the transition matrix de novo, please reset respective parameters in `04_Urd_transition_matrix.Rmd` and `05_Transition_graphs.Rmd` headers. This will increase running time substantially over 60 minutes.

Quick demonstration

We provide a wrapper script file `R_demo.r` which executes annotation analysis for all timepoints and outputs pdf images comprising main Figure 1 into the Results directory. This directory also contains respective original images, allowing confirmation of successful execution. Use `Rscript ./R_demo.r` (running time is around 6 minutes).

Content listing

The content listing is presented in order of the workflow logic: The individual timepoints first, then the transitions part. We provide also cell metadata table, resulting matrix of transition probabilities and gene id conversion table. Direct code output is not part of the repository.

- **10X_matrices**
 - **G4; N0; N2; N5** directories: outputs of 10X cellranger count pipeline (filtered expression matrices), inputs for the individual timepoints analyses *ON DEMAND*
 - **gene_id_conversion_table.csv**: gene id mapping between BraFlo100 and BraLan3 gene models, used to convert gene ids while loading the 10X data with Seurat
- **Individual_timepoints**
 - **Amphi*stage.R**: Seurat workflow used to process individual timepoints data separately
 - **EvidenceCrest_N5.R**: visualisation of Crest population markers in N5 stage as used in supplementary data
 - **PrechordalPlate_Vs_Notochord_N0.R**: investigation and visualisation of genes specific for Prechordal Plate and Notochord populations in N0 stage
 - **Zebrafish_Markers.R**: visualisation of Prechodral Plate and Notochord markers published in zebrafish
- **Transitions**
 - **software** directory
 - **renv.lock**: "lockfile" describing used R packages and their dependencies, to be used with `renv::restore()` within R in the repository directory
 - **environment.yml**: exported conda environment describing used Python modules and their dependencies, to be used with `conda env create -f environment.yml`
 - **01_Integration.Rmd**: MNN integration of individual timepoints

- **02_Convert_to_anndata.R**: conversion of integrated data from R native format to Python native format
 - **03_Cytotrace_pseudotime.py**: calculation of CytoTRACE pseudotime of integrated data
 - **04_Urd_transition_matrix.Rmd**: generation of transition matrix using modified URD approach
 - **05_Transition_graphs.Rmd**: processing of transition matrices, presenting them as directed graphs and exporting
 - **_functions.R**: special functions used throughout the transitions workflow
 - **Results**
 - **Fig1*_original.pdf**: output of `R_demo.sh`
 - **tm_original.RDS**: precomputed transition probabilities matrix *ON DEMAND*
 - **Export**
 - **Cell_metadata.csv**: celltype assignment of cells present in the analyses (all timepoints merged) with integrated UMAP coordinates used in the paper (not generated by the repository code)
 - **Transition_matrix_celltype_timepoint.xlsx**: resulting transition matrix in different transformations, each in own sheet with description (formatted in spreadsheet editor) *ON DEMAND*
-

1. Cell type and regulatory analysis in amphioxus illuminates evolutionary origin of the vertebrate head, In review [↩](#)