Author: Elliot Eton
Due Date: December 17, 2022
Course: BTRY4840

Document title: README.md
Description: Here, I provide background information and instructions on how to recapitulate the findings of my report.

Background regarding the dataset: I performed all analysis on a dataset from a single patient (M12) with a JAK2-mutant-driven myeloproliferative neoplasm. To generate the dataset, the single cell method DOGMA-seq was performed. The ATAC (chromatin accessibility) and RNA (gene expression [GEX]) libraries were sequenced on an Illumina 10X machine and FASTQs generated. Note: The dataset is unfortunately restricted and not shared in the GitHub repository.

Below, I provide an outline or map to working through my repository. Key scripts are italicized. Scripts are defined by their input files, function(s), and output files.

Section 1: Sample processing

1. Since I started my analysis from the outputs of cellranger and velocyto, I do not provide the short scripts my colleague wrote to run cellranger count or velocyto. Tutorials are provided.[1,2]
    a. Briefly, ATAC+GEX FASTQs were loaded into the Cellranger pipeline, which processes cell-barcoded reads and performs STAR transcriptome alignment, unique molecular identifier (UMI) handling, and fragment/read counting. GEX FASTQs were additionally processed by velocyto to quantify unspliced (immature) and spliced (mature) mRNAs, which is necessary for multiVelo. Additional metadata was added to each cell in the Seurat object as required. For example, genotyping of the *JAK2* locus was obtained from a colleague who had written a classifier to genotype cells.[3] For another example, cell identities were obtained by mapping cellular gene expression data to an online bone marrow reference through Azimuth.[4] The ultimate Seurat object was used for calculations of chromatin potential.
2. *integrate.R*: I leveraged two Seurat vignettes[5,6] to process RNA+ATAC data and perform dimensionality reductions and clustering. I made critical adaptations to the script to analyze my dataset, although the general pseudocode is consistent with the vignettes.
    a. Input files: ATAC and RNA count matrices from cellranger
    b. Function:
        i. For ATAC and RNA, separately:
            1. Pre-process and filter data using several quality control measures.
            2. Normalize data.
            3. Identify highly variable features.
            4. Scale data.

[1] https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/tutorial_ct
[2] http://velocyto.org/velocyto.py/tutorial/cli.html#run10x-run-on-10x-chromium-samples
[3] Myers RM, Izzo, Fet al. Integrated Single-Cell Genotyping and Chromatin Accessibility Charts JAK2V617F Human Hematopoietic Differentiation. bioRxiv. 2022. doi:10.1101/2022.05.11.491515.
[4] https://app.azimuth.hubmapconsortium.org/app/azimuth-bone-marrow
[5] https://satijalab.org/seurat/articles/pbmc3k_tutorial.html
[6] https://satijalab.org/seurat/articles/weighted_nearest_neighbor_analysis.html

5. Perform linear dimensional reduction.
6. Determine dataset dimensionality.
7. Cluster cells through k-nearest neighbors.
8. Run non-linear dimensional reduction (UMAP).
9. Assign cell type identity to clusters.

ii. For ATAC and RNA, together:
1. Construct weighted-k-nearest-neighbors graph across ATAC and RNA.
2. Cluster cells based on output.

Section 2: Inferring latent time using multiVelo

The team behind multiVelo provides helpful tutorials to run a robust analysis.[7] I followed these tutorials and wrote critical adaptations to the code to support more robust data analysis.

1. *multivelo_run.py*
   a. Input files:
      i. M12_dogma_velocyto_output.loom
         1. Quantification of immature/mature mRNAs
      ii. 02_ee_azimuth_pred.tsv
         1. File containing cell barcodes and Azimuth cell annotations
      iii. Cellranger output folder
         1. Contains ATAC fragments file and peaks count file
      iv. Seurat folder
         1. nn_idx: coordinates of each cell on a k-weighted-nearest-neighbors graph
         2. nn_dist.txt: distance matrix
         3. nn_cells.txt: cell barcodes
   b. Function:
      i. Infer latent time and rate parameters as described in the methods section of my final project report.
   c. Output file:
      i. multivelo_result.h5ad
2. *multivelo_analysis.ipynb*
   a. Input files: multivelo_result.h5ad, umap coordinates
   b. Function: analyze multivelo output
      i. Identify genes under different regulatory regimes/modules
      ii. Quantify likelihoods of the model
      iii. Embed velocity vectors onto a UMAP
      iv. Explore gene dynamics through pseudotime
   c. Output files: figures

Section 3: Calculating chromatin potential

1. *01_link_peaks.R // 01_link_peaks.sh*
   a. Input files: M12_dogma Seurat object containing ATAC, RNA, and Peaks matrices (output of *integrate.R*)

[7] https://multivelo.readthedocs.io/en/latest/

   b. Function: Pearson correlate peaks with the expression of nearby genes
   c. Output files: M12_dogma Seurat object updated with links matrix and links dataframe alone
2. *02_id_dorcs.R*
   a. Input file: links dataframe
   b. Function: Identify genes that have the highest number of associated peaks ("domains of regulatory chromatin" or DORCs)
   c. Output files: csv identifying DORCs, figure ranking genes by number of peak associations and highlighting top DORCs
3. *03_dorc_scores.R*
   a. Input files: M12_dogma Seurat object with links matrix (output of *01_link_peaks.R*); dataframe and csv identifying DORCs (output of *02_id_dorcs.R*)
   b. Function: Calculate DORC scores by gene for each cell
   c. Output files: updated Seurat object with DORC scores and heatmaps of DORC scores (by cell and by cluster)
4. *04_knn.R // 04_knn.sh*
   a. Input files: M12_dogma Seurat object updated with links and DORC scores (output of *03_dorc_scores.R*)
   b. Function:
      i. Construct k-nearest neighbors graph in chromatin space
         1. Dimensional reduction: PCA
         2. Distance metric: Euclidean, cosine, Manhattan, hamming
      ii. Construct k-nearest neighbors graph in RNA space
         1. Dimensional reduction: PCA
         2. Distance metric: Euclidean, cosine, Manhattan, hamming
      iii. Compute weighted-k-nearest neighbors graph in RNA/chromatin shared space
   c. Output file: updated Seurat object with all (w)knn graphs, figures displaying separation of clusters by first two principal components
5. 05_distance.R
   a. Input files: Seurat object with completed knn graphs
   b. Function:
      i. Compute and average Euclidean distances between each cell and 10 nearest neighbors
      ii. Normalize distances for each cell
   c. Output files: figures displaying chromatin potential on UMAP and across clusters in violin plot