Reproducibility in the Second Real Application (Prostate cancer single-cell RNA-seq data)

This file contains instructions for reproducing the results and figures contained in the second real application. The codes are in "code" folder. We assume the working directory has been appropriately set.

Step 1: "Step1_Preprocessing.R"

Preprocess raw data.

In this step, we focused on the scRNA-seq counts data. We removed genes with zero proportions greater than 95% across cells and external RNA controls consortium (ERCC) spike-in molecules. We then normalized the raw count data via logarithmic transformation accounting for library sizes. Next, we computed the expression variance for each gene, selected the top 100 highly expressed genes, and applied UMAP to reduce the data dimensionality to two.

The output of this step is given by "Real_data.RData".

In the "Real_data.RData":

- dataFile: Preprocessed scRNA-seq data with cell type names
- testData: Preprocessed scRNA-seq data (Input data of the TreeTC model)
- numOfSubjData: Number of cells in each patient
- annotation: Annotation file, including cell barcodes and the corresponding cell types

Step 2: "Step2_TreeTC_mcmc.R"

Apply the TreeTC model to the preprocessed data, conduct the MCMC posterior sampling, and select the iteration following the big node number rule. The output of this step is "Real_result_iteration.RData".

In the "Real_result_iteration.RData":

- q0: Root node of the tree (Normal Class)
- tssb: Tree structure (TssbMCMC Class)
- testData: Input data for TreeTC model
- numOfTrees: Maximal number of source groups ("L" in the model)
- numOfSubjData: Number of data for each source
- etaNormal: Shrinkage parameter for node variances
- tssbNum: Selected iteration index under the big node number rule
- finalSubjAssignments: Source group membership for the iteration "tssbNum"
- finalClIds: Observational-level clustering membership of all the data for the iteration "tssbNum"
- groups_subj: Group numbers (distinct values of the vector "finalSubjAssignments")
- ggFinal1 / ggFinal2: gg: Tree structures (igraph) of each source group for the iteration

"tssbNum"

- subFinalClIds1 / subFinalClIds2: Observational-level clustering membership of each source group for the iteration "tssbNum"
- subFinalCenters1 / subFinalCenters2: Observational-level cluster centers of each source group for the iteration "tssbNum"
- finalSigma: Variance of root node for the iteration "tssbNum"
- finalDrift: Drift parameter for the iteration "tssbNum"

Step 3: "Step3_Figure4.R"

Draw Figure 4 in the manuscript using the output of Step 2.

Step 4: "Step4_FigureS3.R"

Draw Figure S3 in the supplementary using the output of Step 1 and 2.