This Matlab package is developed to detect and characterize spatial or developmental pseudotime trajectories from single-cell RNA-seq data. The trajectory (or branched trajectories) is detected as the probability density ridge of the data points in a 2D embedded space (e.g., UMAP). Following trajectory detection, genes that are differentially expressed along the trajectory (i.e., with a trajectory dependent expression pattern) are detected.

Note that such trajectory analysis is aimed at facilitating exploratory analysis such as detecting potential genes functioning during the analyzed process. Detection of a trajectory should only be used as supporting evidence (not proof) that the analyzed cells are from a temporally or spatially continuous process.

The method is also compatible with other multi-dimensional single-cell data, such as mass cytometry and multiplex microscopy data.

Input & Output

Input: A table describing gene (or marker) expression level in each single cell. Several manual inputs are requested, such as trajectory direction. Option of interactive manual input of trajectory is available (in the case of an unsatisfactory automatic trajectory detection).
Output: An inferred trajectory (or multiple trajectories/branches), coordinates of single cells on the trajectory (as well as deviations from the trajectory – used to identify and filter outliers), and top genes (markers) that follow a trajectory-dependent pattern (i.e., differentially expressed along the trajectory).

Data pre-processing & 2D embedding

Data were first pre-processed following the strategy of the Seurat single cell RNAseq analysis package (https://satijalab.org/seurat/). Cell by gene transcript count matrix were cell library size normalized, log-transformed, and scaled (see below diagram). Next, the program detects highly variable genes (dispersion of the gene, i.e., variance/mean > 0.5, as defined in the commonly used Seurat package). Note that in this step we used a moving average filter (smooth function in Matlab) instead of binning (which is implemented in Seurat) to improve continuity. To reduce computational intensity, only the highly variable genes were used in the downstream principal component analysis (PCA). After PCA, the program requests the user to carefully examine the top 20 principal components to select the high-quality principal components that are not driven by extreme outlier data points or immediate early genes. These top high quality principal components were then used as input to generate a 2D UMAP with cell-cell Euclidean distances as input using the umap Matlab package (https://www.mathworks.com/matlabcentral/fileexchange/71902). Other 2D embedding space can also be used such as tSNE. Cells that were sampled from a temporal (or spatial) continuous process are expected to form a visually detectable continuum in the embedded space, given that 1) overall gene expression changes gradually and continuously during the process and 2) sampling rate is high enough to capture the entire process.

Trajectory Detection

The trajectory of the cell continuum was detected as the probability density ridge of the data points in the 2D embedded space, using automated image processing (Matlab Image Processing Toolbox™); interruptions in the detected density ridge line can be connected manually following user direction, or automatically connected according to the shortest distance. If automatic detection fails to accurately capture the trajectory, the program also allows interactive manual assignment of a trajectory/branch. Direction of the trajectory was requested from user. Individual cells were then aligned to the trajectory by the shortest connecting point to the trajectory in the 2D embedded space; if the trajectory branched, cells were assigned to the closest branch. Individual cells that were too distant from the trajectory (adaptive thresholding along the trajectory) were deemed outliers and removed from further analysis.

Detecting genes differentially expressed along the trajectory

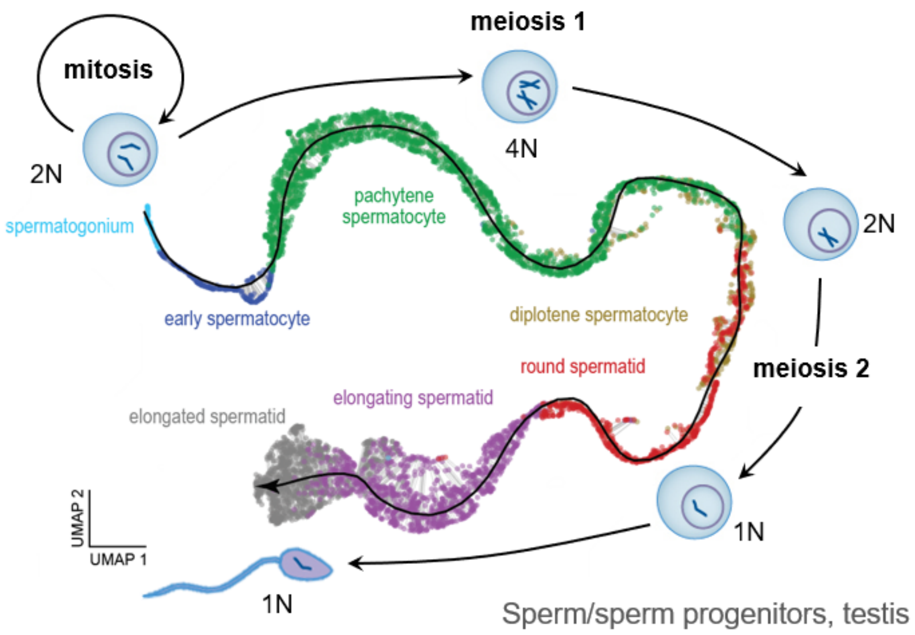
To detect genes whose expression followed the trajectory, we calculated the Spearman correlation coefficient and corresponding *p*-values (Bonferroni corrected) between the expression level of each gene and 20 preassigned unimodal patterns that smoothly change along the trajectory (with their single peaks uniformly distributed from the beginning of the trajectory to its end point). Expression patterns of the top ranking (top 1000 with *p*-value<0.01) and highly variable (dispersion > 0.5) genes were smoothed with a moving average filter and clustered by k-means clustering to detect the major trajectory-dependent expression patterns. The trajectory differentially-expressed genes were then ranked by the associated cluster (ranked by trajectory location of peak expression), and within the cluster by *p*-value from smallest to largest, and with the same *p*-value by mean expression level from highest to lowest.

Examples

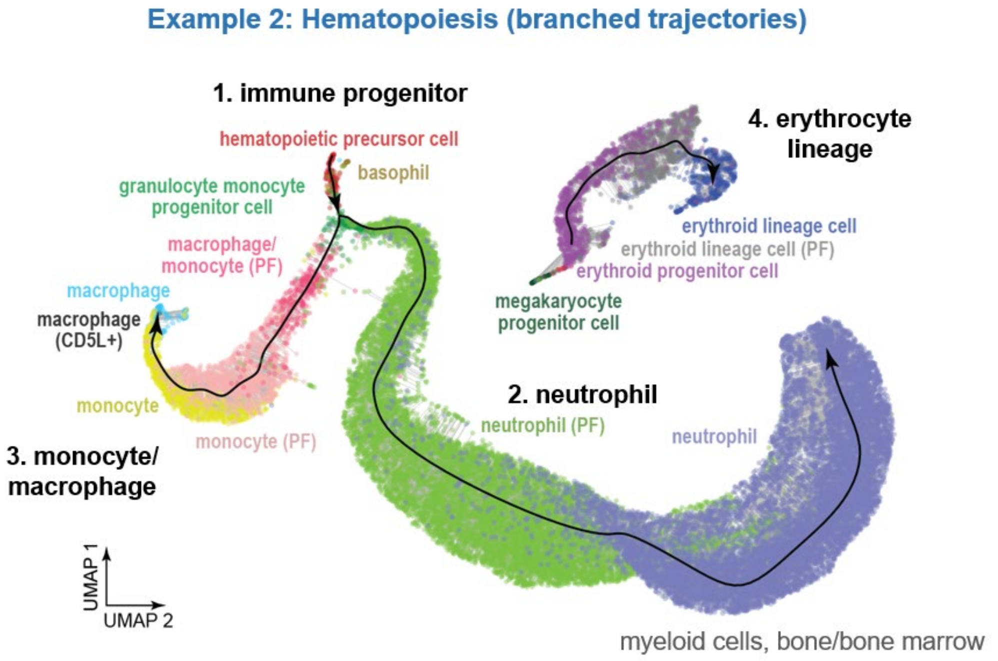
Two example datasets and code were provided:

1. Spermatogenesis (10x single-cell RNAseq of testis, sperms and sperm progenitors) – a single pseudotime trajectory

Example 1: Spermatogenesis (a single trajectory)



2. Hematopoiesis (10x single-cell RNAseq of the bone marrow, myeloid cells) – branched pseudotime trajectory


Example 2: Hematopoiesis (branched trajectories)

How to cite (& more examples)
Ezran C, Liu S, Chang S, Ming J, Botvinnik O, Penland L, Tarashansky A, Morree A de, Travaglini KJ, Hasegawa K, Sin H, Sit R, Okamoto J, Sinha R, Zhang Y, Karanewsky CJ, Pendleton JL, Morri M, Perret M, Aujard F, Stryer L, Artandi S, Fuller M, Weissman IL, Rando TA, Ferrell JE, Wang B, Vlaminck ID, Yang C, Casey KM, Albertelli MA, Pisco AO, Karkanias J, Neff N, Wu A, Quake SR, Krasnow MA. (The Tabula Microcebus Consortium). 2021 ***bioRxiv***. *Tabula Microcebus: A Transcriptomic Cell Atlas of Mouse Lemur, an Emerging Primate Model Organism* doi:10.1101/2021.12.12.469460

Also see:
Ezran C, Liu S, Ming J, Guethlein LA, Wang MFZ, Dehghannasiri R, Olivieri J, Frank HK, Tarashansky A, Koh W, Jing Q, Botvinnik O, Antony J, Chang S, Pisco AO, Karkanias J, Yang C, Ferrell JE, Boyd SD, Parham P, Long JZ, Wang B, Salzman J, Vlaminck ID, Wu A, Quake SR, Krasnow MA. (The Tabula Microcebus Consortium). 2022 ***bioRxiv***. *Mouse Lemur Transcriptomic Atlas Elucidates Primate Genes, Physiology, Disease, and Evolution.* doi:10.1101/2022.08.06.503035

Liu S, Ezran C, Wang MFZ, Li Z, The Tabula Microcebus Consortium, Long JZ, Vlaminck ID, Wang S, Kuo C, Epelbaum J, Terrien J, Krasnow MA, Ferrell JE. (2021) ***bioRxiv***. *An Organism-Wide Atlas of Hormonal Signaling Based on the Mouse Lemur Single-Cell Transcriptome.*; doi:10.1101/2021.12.13.472243