# Overview

These are the analytic program and preprocessed data in the manuscript: Nguyen et al**. Single-cell RNA sequencing analysis identifies one subpopulation of endothelial cells that proliferates and another that undergoes the endothelial-mesenchymal transition in regenerating pig hearts**. *Frontiers in Bioengineering and Biotechnology*, Tissue Engineering and Regenerative Medicine. 2023. DOI: 10.3389/fbioe.2023.1257669. The analysis was done via Matlab in three steps:

- Step 1 (file *Step1*\_*Pre*\_*Clustering*\_*endothelial*\_*cells.m*) computes a cell-cycle-specific Autoencoder model, which embeds the endothelial cell data into just 10 dimensions, then performs 2D (UMAP) visualization and pre-clustering.

- Step 2 (file *Step2*\_*Final*\_*Clustering*\_*Marker*\_*analysis.m*) clusters and defines 5 endothelial cell subpopulations in the dataset. These include one subpopulation of proliferating endothelial cells (indication of angiogenesis) and one subpopulation of endothelial cell that expresses endo-mesenchymal cluster, which only appears in recovered hearts. Statistics of all genes in each subpopulation (cluster) are computed.

- Step 3 (file *Step3\_Key\_Results.m*) reproduces the key result figures from the manuscript.

Steps 2 &3 can be executed without rerunning Step1; also, Step3 can be executed without rerunning Step 1&2. Thus, expert bioinformaticians can execute all Steps 1, 2, and 3 to reproduce the results completely. Researchers with some bioinformatic or programming skills can execute Steps 2 and 3. Researchers with no bioinformatic/programming skills can just execute Step 3 to see the statistics and visualize the specific (interested) genes.

# Hardware & software

- Software: The entire analytic program is written and executed in Matlab. Successful execution was confirmed in Matlab 2020b, Matlab 2021b, and Matlab 2022b version.

The analytic program utilizes the third-party toolkits:

+ Stephen Meehan. Uniform Manifold Approximation and Projection (UMAP). Retrieved from <https://www.mathworks.com/matlabcentral/fileexchange/71902-uniform-manifold-approximation-and-projection-umap>.

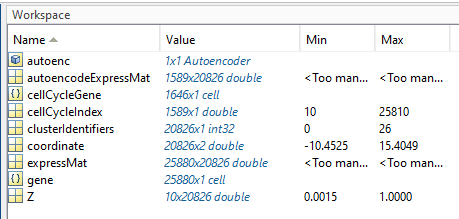
+ Miguel García et al. Violin Plots for Matlab. Retrieved from <https://github.com/bastibe/Violinplot-Matlab>.

- Hardware: A computer with NVIDIA Graphic Processing Unit (GPU) is recommended to speed up Step 1. The authors used a computer having a NVIDIA Quardo RTX 4000 GPU.

# Explanation of Matlab workspace variables in each Step

## Step1\_Pre\_Clustering\_endothelial\_cells.m

Below is the screenshot of Matlab workspace after finishing Step1:



*autoenc*: the Autoencoder model that embeds the endothelial cells’ gene expression into just 10 dimensions.

*autoencodeExpress*: expression of cell-cycle-specific genes, log2-scaled, which is used to computer *autoenc*.

*cellCycleGene*: the list of cell-cycle-specific genes, which is used to produce *autoencodeExpress*.

*cellCycleIndex*: associated with *autoencodeExpress* and *cellCycleGene*, for indexing purposes.

*clusterIdentifiers*: pre-clustering results of all cells after being embedded by *autoenc*, followed by the UMAP toolkit.

*coordinate*: 2D coordinates, computed via UMAP toolkit, for visualization. *coordinate* is associated with *clusterIdentifiers*.

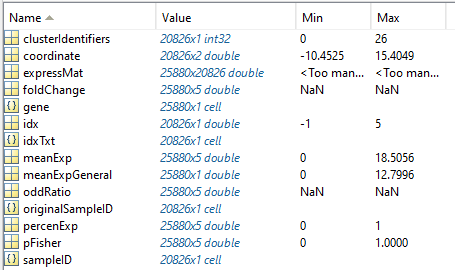
*expressMat*: expression of all genes in all endothelial cells from the single-cell RNA sequencing data. Each ‘row’ corresponds to a gene (a total of 25880 genes). Each of 20826 column corresponds to a cell. *autoencodeExpress* is deducted from *expressMat* by selecting just 1646 *cellCycleGene*.

*gene*: the list of entirely 25880 genes.

*Z*: the embedding (10 dimensions – row) of all 20826 cells by *autoenc*.

## Step2\_Final\_Clustering\_Marker\_analysis.m

Below is the screenshot of Matlab workspace after finishing Step2



*clusterIdentifiers*: pre-clustering results of all cells already done and stored from Step1.

*coordinate*: 2D coordinates, computed via UMAP toolkit, for visualization; already done and stored from Step1.

*expressMat*: same as Step1, expression of all genes in all endothelial cells from the single-cell RNA sequencing data. Each ‘row’ corresponds to a gene (a total of 25880 genes). Each of 20826 columns corresponds to a cell.

*gene*: the list of entirely 25880 genes.

*idx*: cluster number for each cell, from 1 to 5. Cells with *idx* < 1 do not correspond to any cluster and should be removed (lines 41-49)

*idxTxt*: name of each cluster; correspond to *idx* by:

*idx* == 1 - *idxTxt* = 'VEC1'

*idx* == 2 - *idxTxt* = 'VEC2'

*idx* == 3 - *idxTxt* = 'VEC3'

*idx* == 4 - *idxTxt* = 'LEC1'

*idx* == 5 - *idxTxt* = 'LEC2'

*foldchange*: the fold-change expression of each gene (25880 genes) in each cluster

*meanExp*:average expression of each gene in each cluster

*percenExp*: for each gene in each cluster, it is the ratio (from 0 to 1) of the cell cluster expressing the gene.

*pFisher*: statistical p-value of each gene in each cluster, computed by Fisher’s Exact Test.

*oddRatio*: odd-ratio of each gene in each cluster, computed by Fisher’s Exact Test (associated with *pFisher*)

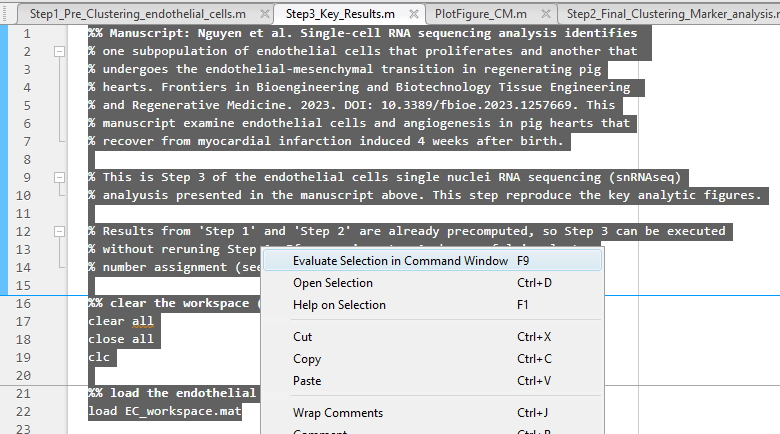
*sampleID*: the pig animal group for each cell.

*originalSampleID*: the pig individual animal (an animal group consists of 2-5 pigs) for each cell.

## Step3\_Key\_Results.m

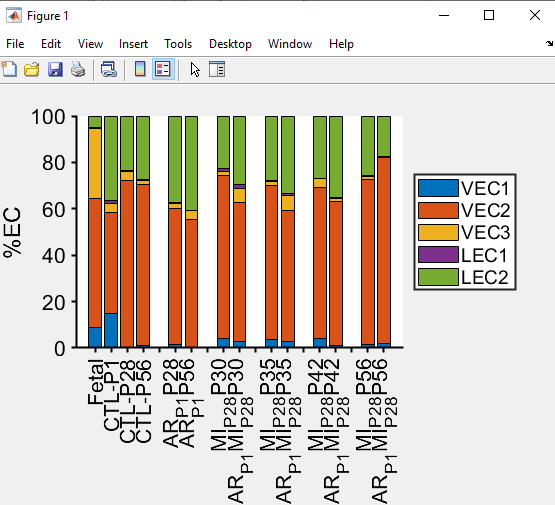
The Matlab workspace in Step3 is the same to Step2. Step3 is primarily for researchers with no bioinformatics/programming background to directly examine and reproduce the results as in the manuscript Nguyen et al**. Single-cell RNA sequencing analysis identifies one subpopulation of endothelial cells that proliferates and another that undergoes the endothelial-mesenchymal transition in regenerating pig hearts**. *Frontiers in Bioengineering and Biotechnology*, Tissue Engineering and Regenerative Medicine. 2023. DOI: 10.3389/fbioe.2023.1257669.

- Load the data: Select from line 1 to line 22, right click, and choose Evaluate Selection in Command Window



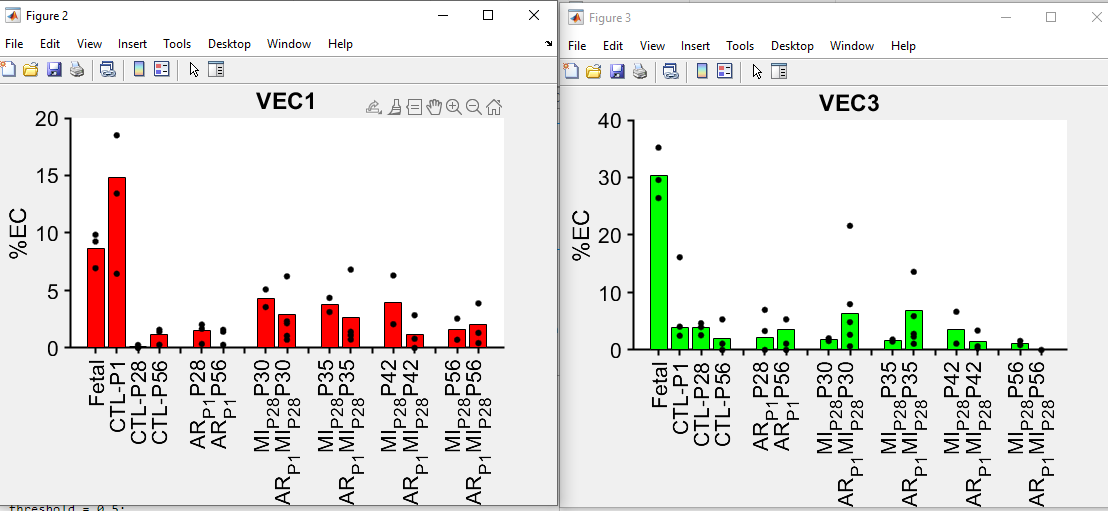
- Plot the endothelial cell cluster proportions (5 clusters) in the stack-bar chart (line 26, right click and choose Evaluate Selection in Command Window)



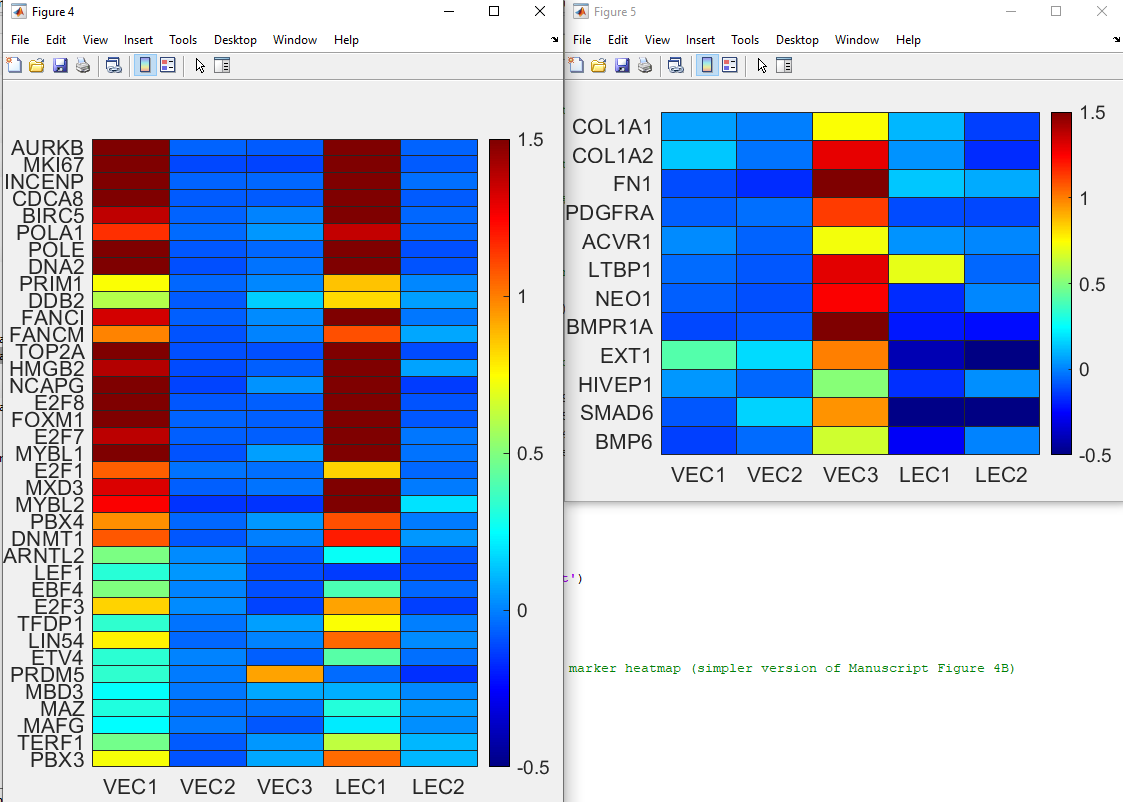


The figure can be manually changed (i.e. change color) by Matlab. See a tutorial at <https://www.youtube.com/watch?v=GJJVbKbfIpQ>.

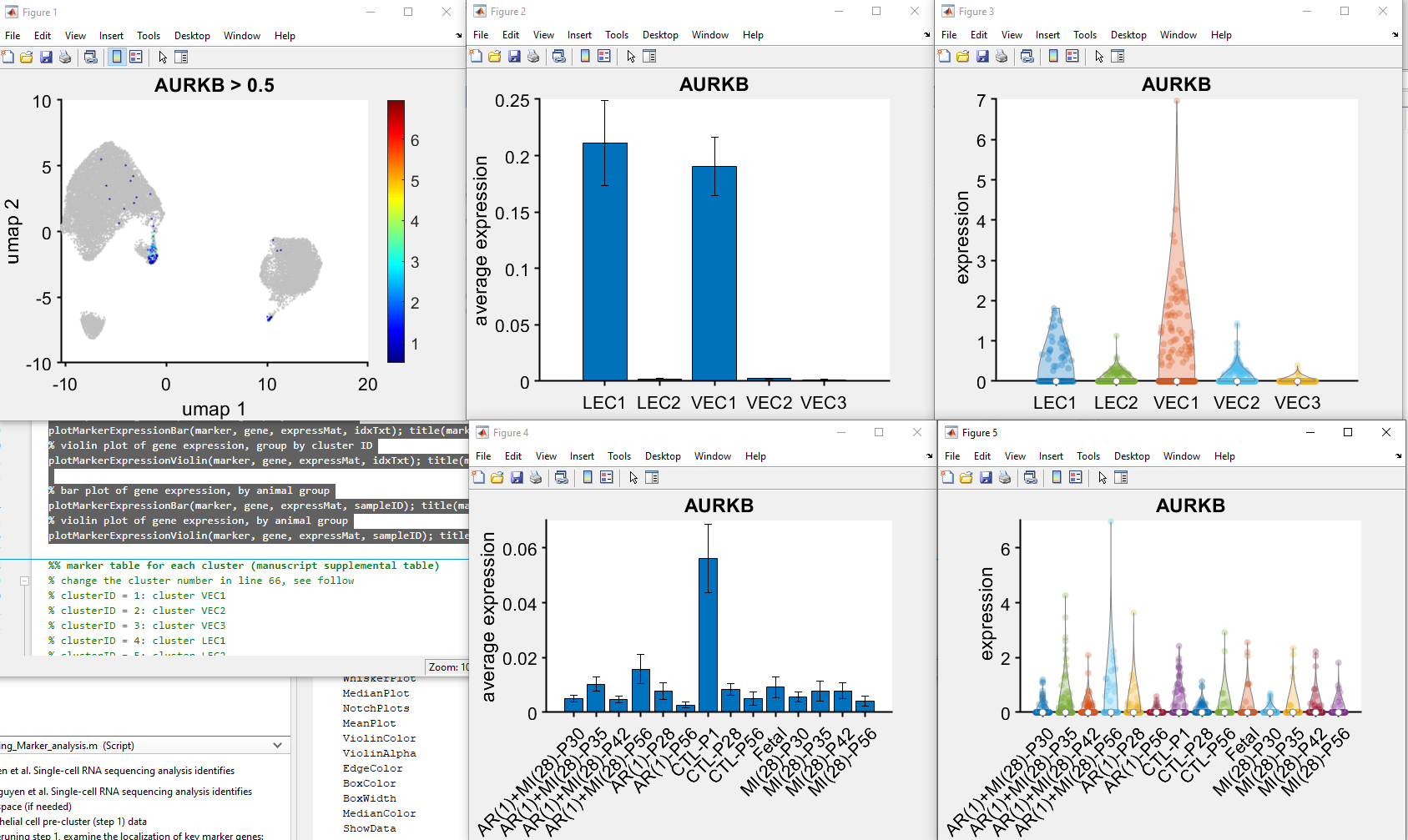
- Plot the proportions of cluster VEC1 and VEC3, which are the key results in the manuscript (line 36



- Plot the heatmap of cell-cycle marker genes (enriched in cluster VEC1 and LEC1) and endo-mesenchymal-transition genes (enriched in cluster VEC3), lines 33 and 36

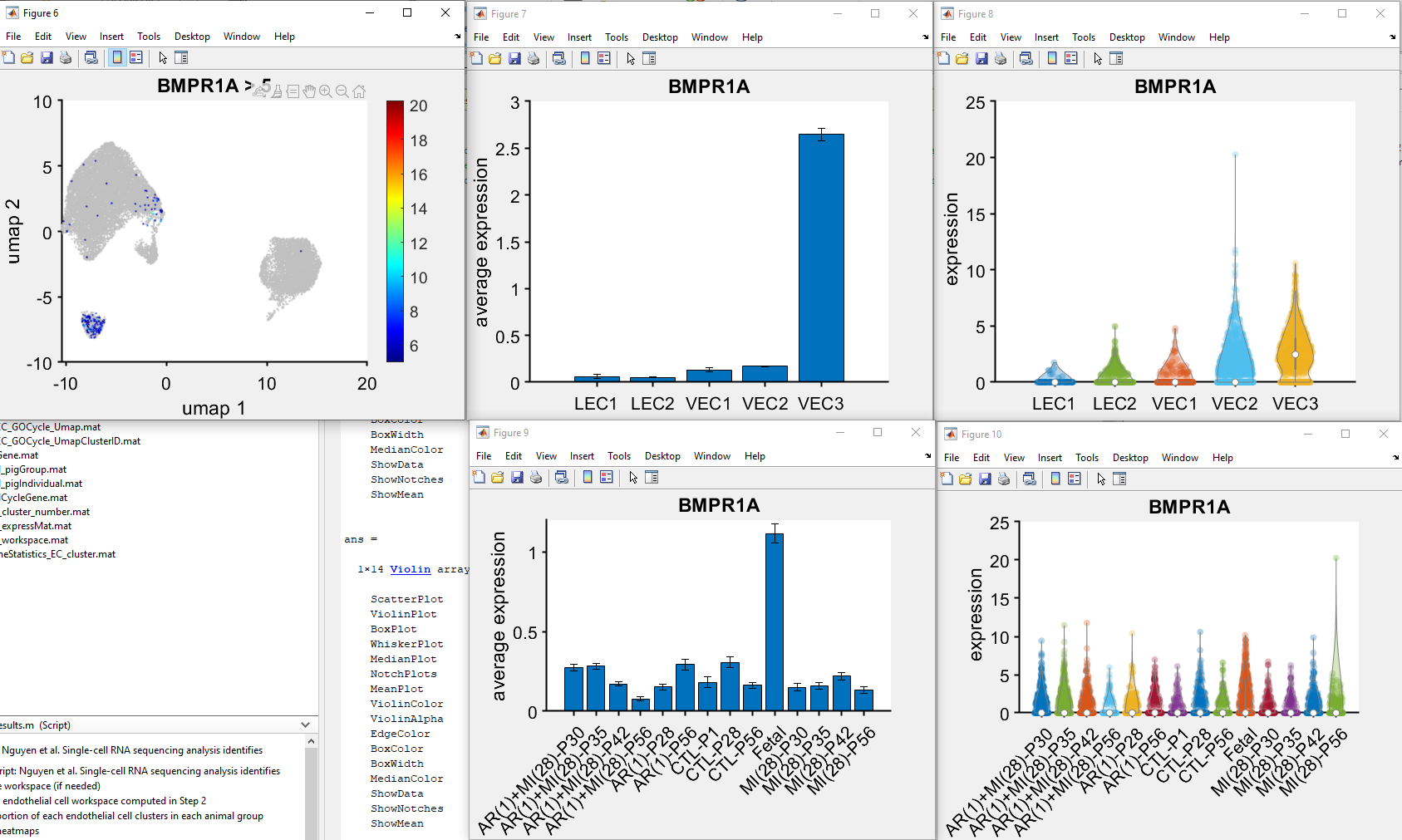


- Lines 44-56 are to plot figures of a specific gene, which is specified in line 44. The example uses AURKB.



Change the gene name in line 44 to plot other genes. For example, if changing to BMPR1A

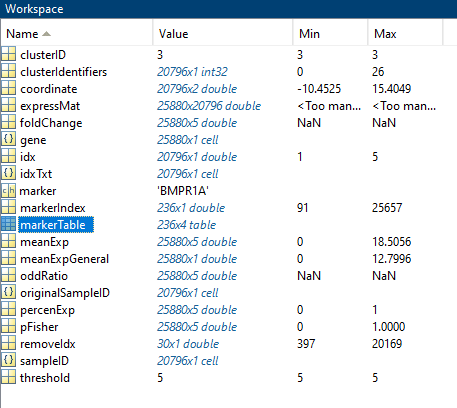
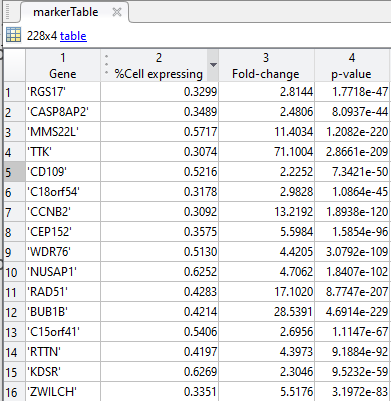




- Line 66-71 are to view the upregulated genes in each cluster. The cluster is specified in line 66. This example views the upregulated genes in cluster VEC1 (clusterID = 1).



Run line 66-71 (right click and choose Evaluate Selection in Command Window). Double-click in *markerTable* of the Matlab workspace.

To view upregulated genes in other clusters, change the line 66:

clusterID = 1: VEC1 cluster

clusterID = 2: VEC2 cluster

clusterID = 3: VEC3 cluster

clusterID = 4: LEC1 cluster

clusterID = 5: LEC2 cluster