Downstream Exploration for Integrated Single Cell RNA Sequencing

Michiko Ryu, Dr. Yan Li (mentor)

University of Chicago, Master of Science in Physical Science Department

Abstract

Single-cell RNA sequencing(scRNA seq) is a powerful technique for scientists to discover and characterize cell types, populations, status, and their developmental trajectories. Due to the high level of variation between cells, it is challenging to develop a simple and efficient protocol to handle the noisy scRNA seq data.

In this thesis paper, I will explore a computational downstream for an integrated scRNA seq. The downstream has three major components:

(1) data preprocessing that removes unwanted data points, and cluster cells to find their features.
(2) pseudotime ordering that allows scientists to compare properties across clusters.
(3) an interactive web application to visualize gene level features of specific markers defined by users.

Data Used

To evaluate the entire downstream, we applied single cell data of ovary tissues obtained from 4 patients (#3041 #3061 #3203 #3296). Possible cell types within this dataset are myofibroblast, stromal, T/NK cells, B-cells, macrophage, smooth muscle cells, pericyte, endothelial, lymphatic endothelial, epithelial, non-ciliated epithelial, ciliated epithelial, and mast cells.

Marker	Gene	Cell Type
PDGFRA	PDGFRA	Myofibroblast
DCN	DCN	Stromal
RUNX3	RUNX3	T/NK cells
CD3E	CD3E	T/NK cells
JCHAIN	JCHAIN	B-cells
CD163	CD163	Macrophages
LYZ	LYZ	Macrophages
ACTA2	ACTA2	Smooth muscle
MYH11	MYH11	Smooth muscle
CSRP1	CSRP1	Smooth muscle
MCAM	MCAM	Pericyte
PECAM1	PECAM1	Endothelial
KDR	KDR	Endothelial
CD34	CD34	Endothelial
LYVE1	LYVE1	Lymphatic Endothelium
PROX1	PROX1	Lymphatic Endothelium
KRT7	KRT7	Epithelial
KRT8	KRT8	Epithelial
KRT18	KRT18	Epithelial
EPCAM	EPCAM	Epithelial
CDH1	CDH1	Epithelial
OVGP1	OVGP1	Non-ciliated epithelial
CAPS	CAPS	ciliated epithelial
TPSB2	TPSB2	Mast cells

Table 1. List of tested genes and their relative markers & tissue type

Method Development

Part 1: Pre-processing and Clustering

Removal of Unwanted Cells

1. Detect and remove cells that have high mitochondrial gene expression (>20)
2. Remove technical artifacts (doublet) based on both centroids and mediod algorithms of DoubletDecon from Seurat toolkit

Cluster Analysis

1. Use dimension reduction techniques (PCA, t-SNE, UMAP) to transfer a high dimensional data to 2D plots, Visually exploring.
2. Take the normalized read count data to discover differentially expressed genes.

Part 2. Visualization of Genes Markers on Shiny Application

Web Application Development (Shiny App)

Develop user interface with three reactive input components:

a. Text input for users to enter the full path of .rds file gotten from Part 1
b. Another text input to enter the path of an excel file that listed all gene/markers users are interested in
c. Selection box to choose one experimental condition

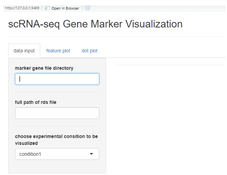

Figure 1. Actual Shiny App interface we developed

Visualization

1. Obtain "feature" like PC values or number of genes detected, so that we can color gene expression level changes across different clusters/ cells.
2. Display feature plot and dot plot separately on two more tabs.

Part 3. Pseudotime analysis

1. Label cell stage and separates them along differentiation trajectory by Principal Components(PCs), diffusion map, and Slingshot value.
2. Plot pseudotime population and compare values on each identities (clusters) to clarify separation of clusters in part 1.

Toolkit Used

1. Seurat: an R package originally developed as a clustering tool. It enables users to perform rapid mapping and integrative multimodal analysis for diverse types of single-cell data.
2. Shiny App: a new open-source R framework to build interactive web applications from RStudio.

Evaluation

Part 1: Pre-processing and Clustering

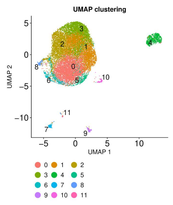

Figure 2. UMAP of ovary tissues

#sample 3041	No. of Original Cells	3514
	No. of Cells After Mt removal	1684
#sample 3061	No. of Original Cells	2546
	No. of Cells After Mt removal	2118
#sample 3203	No. of Original Cells	11492
	No. of Cells After Mt removal	7327
	No. of Cells After Mt & Doublet removal	7178
#sample 3296	No. of Original Cells	6516
	No. of Cells After Mt removal	5564
	No. of Cells After Mt & Doublet removal	4931

Table 2. No. of cells (before and after)

Part 3. Pseudotime analysis

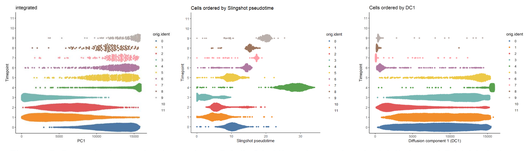

Figure 3. Pseudotime plot based on PC (left), Slingshot value (middle), and DM (right)

Dot Plot: Dot size reflects the number of cells expressed in that identity (cluster). A larger dot indicates a greater proportion of a gene in that cluster. Meanwhile, dot color is a reference of expression levels within a cluster. A darker dot means a higher average expression value of a specific gene detected from that type of cell.

Feature Plot: Redness indicate cells with a specific bio-marker. Because of the large population in clusters #0 #1 #2 and #3, there are several markers that have relatively high expressions in the clusters.

Pseudotime Plot: Pattern of x value describe properties of each identity (cluster). Difference between the patterns is a proof of cell separation.

Part 2. Visualization of Genes Markers on Shiny Application

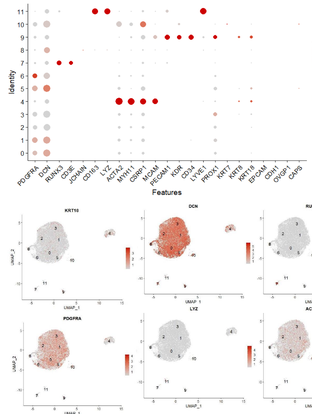

Figure 4. Dot plot of integrated data

Figure 5. Feature plots of each gene marker

Conclusion

Clusters in the UMAP of ovary tissues are gathering to form a large aggregation. However, none of them have similar properties according to pseudotime analysis. We were able to identify part of the cell types based on the dot plots and feature plots. Some cell properties are still ambiguous.

Future Work:
one future direction to improve the downstream is to add more visualization options like plots focusing on one specific cluster. Besides, we need to consider other experimental condition options; For example, we can separate metadata by tissue types, not patients' numbers.

Acknowledgements

My sincere thanks to:
- Dr. Yan Li
- Dr. Mengjie Chen
- Dr. Mark J. Oreglia