Nitsueh Kebere - Programmer and Analyst
TA: Dakota Hawkins

Single-Cell RNA-seq Analysis of Pancreatic Cells

Introduction

The pancreas is a vertebrate organ that plays a role in digesting the food we eat and regulating energy homeostasis. Particularly, the pancreas is known to produce insulin to regulate the body's glucose or sugar level, and its dysfunction can lead to diseases such as type 1 and type 2 diabetes. In the past, several studies have characterized the gene expression profile of cells in the pancreas, and have identified different cell types. However, most of these studies have been limited by the difficulty of analyzing human cells at a high-throughput scale. To provide a more comprehensive gene expression profile of pancreatic cells, Baron et al. (2016) used the inDrop single-cell RNA-seq (sc-RNA-seq) platform, which uses high-throughput droplet microfluidics to allow the capturing of thousands of cells without presorting. The authors studied the gene expression profile of over 12,000 pancreatic cells from four human donors and two mice strains, and found 14 cell clusters and 13 cell clusters that matched previously characterized cell type in human and mice samples respectively. Furthermore, the authors also found subpopulations of ductal and beta cells with distinct expression profiles (Baron et al., 2016).

This particular project used the UMI counts matrix generated (using salmon alevin 1.2.0 (Srivastava, A. et al., 2019)) from the sc-RNA-seq data associated with a 51-year-old female donor in an attempt to replicate the primary findings of Baron et al. (2016). The UMI count matrix was processed, and K-nearest neighbor (KNN) graph-based clustering was implemented using the Seurat package (Stuart et al., 2018). Additionally, marker genes were identified and assigned to a cell type.

Methods and Results

Quality Control and Selecting Cells for Further Analysis

Before performing any sort of quality control, the Ensembl gene identifiers in the count matrix were mapped to gene symbols using the biomaRt package, and values of genes that were duplicated were summed before removing the gene duplicates.

Using the Seurat package, the Seurat object was initialized with the count matrix that contained genes that are expressed in 3 or more cells, and cells (barcodes) with 200 or more genes (Stuart et al., 2018). The number of counts per barcode (nCount_RNA), number of genes per barcode (nFeature_RNA), and the fraction of counts from mitochondrial genes per barcode (percent.mt) were visualized using a violin plot (Fig 1A). Furthermore, scatter plots were generated (Fig 1B) to visualize relationships between nCount_RNA and nFeature_RNA, and nCount_RNA and percent.mt. Based on the plots in Figure 1, low quality cells were filtered by removing cells that have unique feature counts over 2,500 or less than 200, and cells that have >10% mitochondrial counts. The filtering method was implemented to reduce low-quality/empty droplets, cell doublets or multiplets, and low-quality/dying cells, which are associated with extensive mitochondrial contamination.

Number of cells and genes in different filtering stages are shown in Table 1.

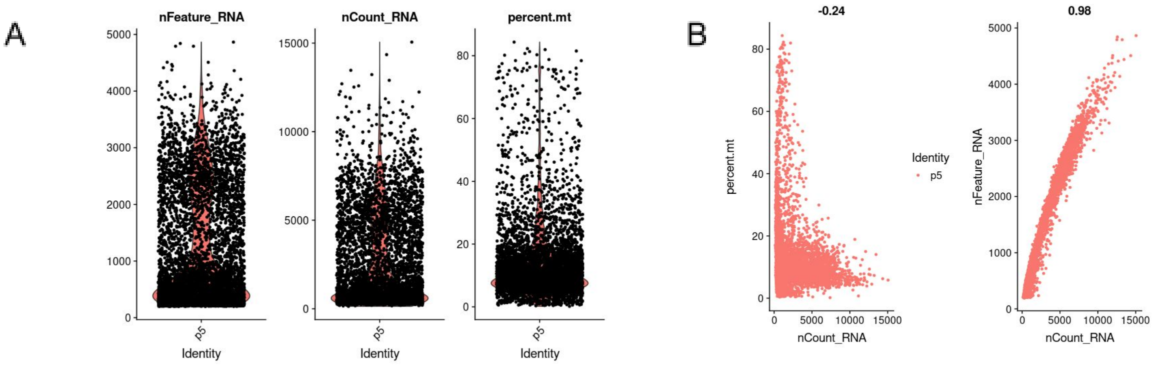


Figure 1. Quality control metrics as a Violin plot and Feature-Feature relationships as a scatter plot. A) Violin plot for the number of counts per barcode, number of genes per barcode, and percent mitochondria. B) Scatter plot showing relationship between features. The proportion of mitochondrial genes per barcode and number of genes per barcode is shown.

Table 1. Reports of the number of cells and genes in different filtering stages.

Data	Number of Genes	Number of Cells
Count Matrix	60233	8241
Count Matrix (after gene mapping)	60220	8241
Filtered Count Matrix	20438	1881

Normalization and Feature Selection

After removing low quality cells, the expression measurements for each cell was normalized by the total expression and log transformed. Low variance genes from the Seurat object were then filtered using “FindVariablesFeatures” function from Seurat, where 2000 variable genes were returned to use for downstream analysis (Fig 2).

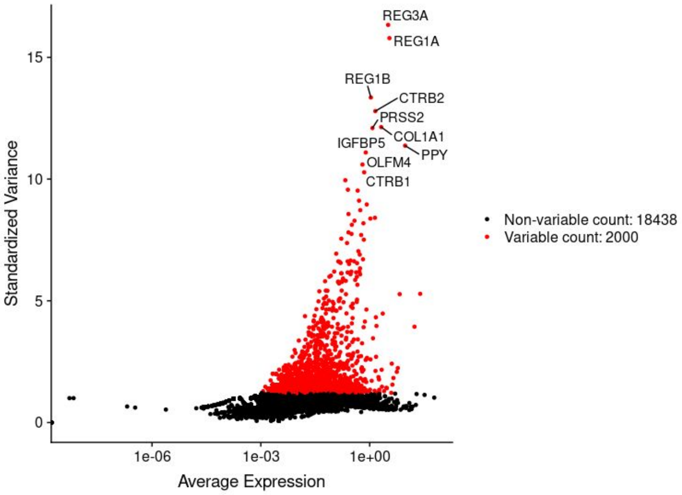


Figure 2. Scatter plot of highly variable genes. The top 10 most highly variable genes are labelled.

Linear Dimensional Reduction and Cell Clustering

A linear transformation (scaling) was applied to the 2000 variable features returned after filtering for low variance genes, and Principal Component Analysis (PCA) was performed on the scaled data. JackStraw and Elbow plots (Fig 3A&B) were generated to determine the number of components to include for downstream analysis.

The JackStraw plot generated shows that 15 PCs have low p-values and are significant, however the Elbow plot shows an ‘elbow’ around PC10-12, suggesting that the majority of true signal is captured in the first 10 PCs. Therefore, the first 10 PCs were used for clustering.

To cluster the cells, the Seurat package constructed a K-nearest neighbor (KNN) graph based on euclidean distance in the PCA space, and applied the Louvain algorithm to iteratively group cells together (Stuart et al., 2018). 11 clusters were obtained, and the resulting cell clusters were visualized using tSNE (Fig 3C). The relative proportions of cell numbers for each cluster are shown in a bar chart (Fig 3D).

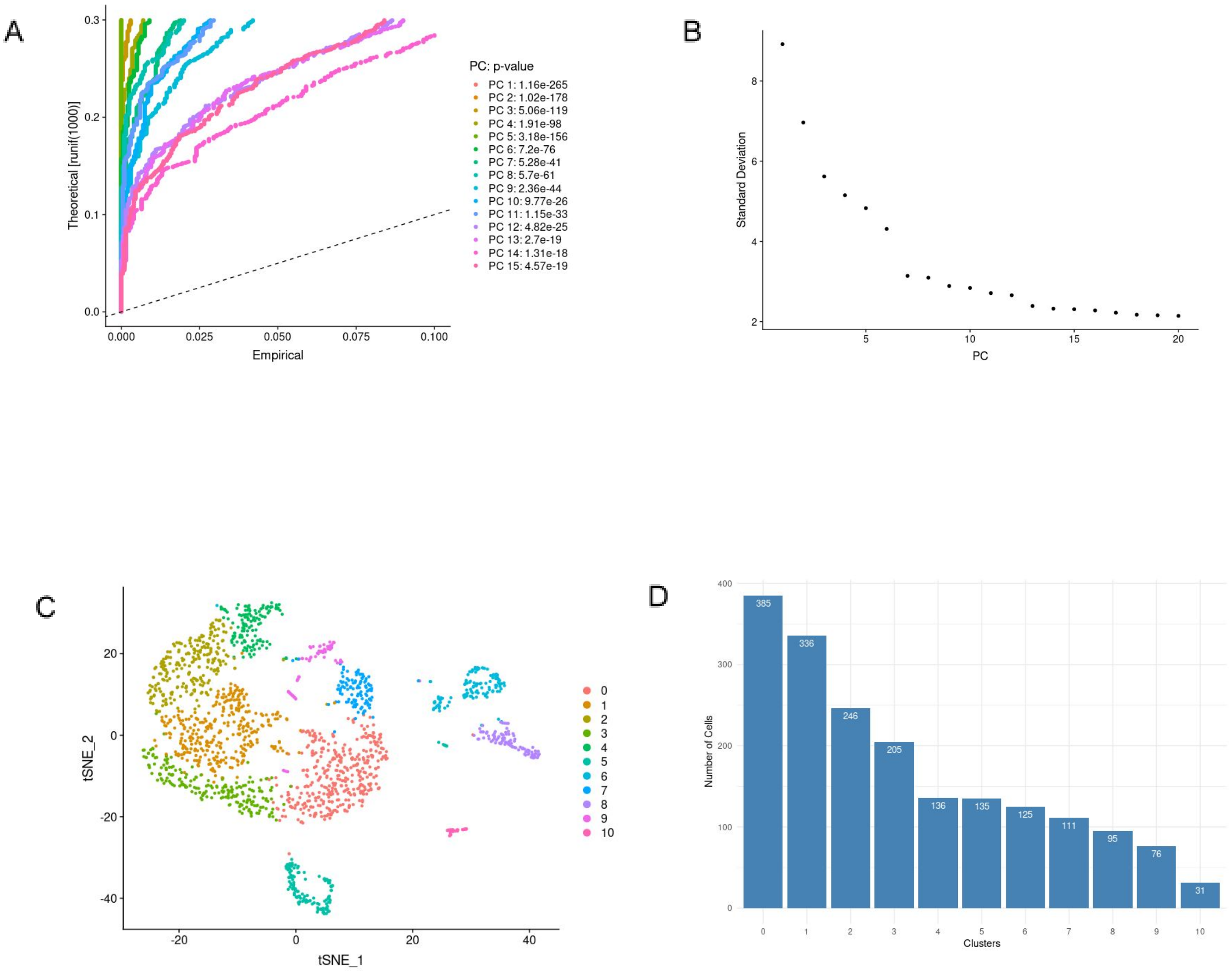


Figure 3. Linear dimensional reduction and cell clustering results. A) Jackstrow plot that shows the significant PCs. B) Elbow plot shows an elbow around PC10-12. C) tSNE plot that projects 11 different cell clusters. D) The number of cells in clusters 0-11.

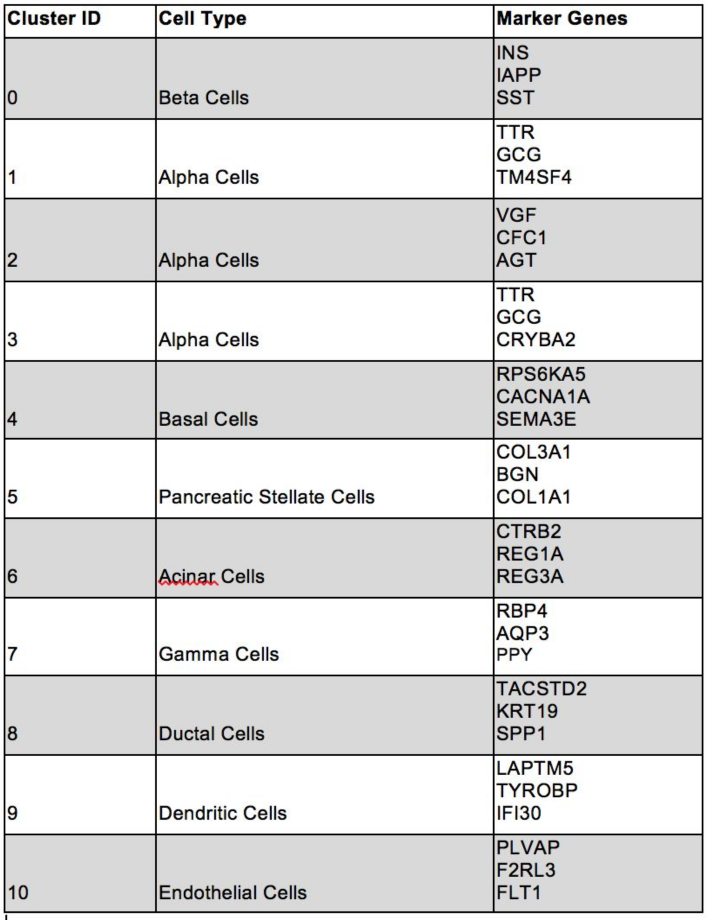
Identifying Cluster Marker Genes and Assigning Cell Type

After obtaining the cell clusters, marker genes were identified for all clusters using Seurat (Stuart et al., 2018). The marker genes were then sorted by adjusted p-value, and the top three marker genes for each cluster were used to assign cell types to the clusters.

From the marker genes discovered by Baron et al. (2016); INS, SST, GCG, PPY, and KRT19 were the only genes that appeared in the cell cluster generated in this project. In cases where these marker genes from Baron et al. (2016) were absent, PanglaoDB–a database that provides easy exploration of single cell RNA sequencing experiments–was used to assign cell types based on the top differentially expressed gene from the cluster (Franzén et al., 2019).

Table 2 shows the top three marker genes with the smallest adjusted p-values for each cluster along with the associated cell type.

Table 2. Top three marker genes and their associated cell type. INS, SST, GCG, PPY, and KRT19 were the only marker genes that appeared in cell clusters identified by Baron et al. (2016).



Cluster ID	Cell Type	Marker Genes
0	Beta Cells	INS IAPP SST
1	Alpha Cells	TTR GCG TM4SF4
2	Alpha Cells	VGF CFC1 AGT
3	Alpha Cells	TTR GCG CRYBA2
4	Basal Cells	RPS6KA5 CACNA1A SEMA3E
5	Pancreatic Stellate Cells	COL3A1 BGN COL1A1
6	Acinar Cells	CTRB2 REG1A REG3A
7	Gamma Cells	RBP4 AQP3 PPY
8	Ductal Cells	TACSTD2 KRT19 SPP1
9	Dendritic Cells	LAPTM5 TYROBP IFI30
10	Endothelial Cells	PLVAP F2RL3 FLT1

A heatmap was used to display and further evaluate the top three marker genes for each cluster (Fig 4). The heatmap generated in this project has a similar trend to the one generated by Baron et. al, where marker genes are highly expressed in their associated cluster type.

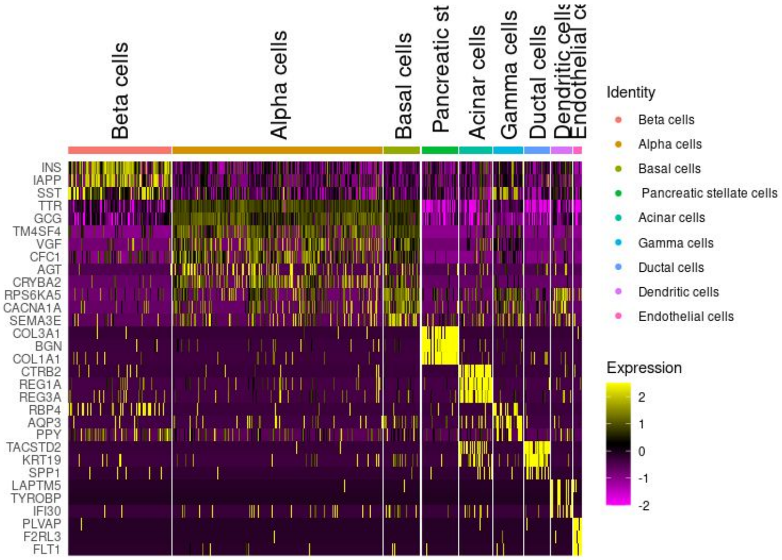


Figure 4. Heat map of log normalized counts for the top three marker genes. High expression is indicated by yellow and low expression is indicated by purple/magenta. Marker genes are highly expressed in their associated cell cluster

To visualize the cell clusters, a tSNE plot was generated (Fig 5). Of the nine major clusters identified by Baron et al., this project reproduced 7 of the clusters. Similar to Baron et al. (2016), the tSNE plot showed clear clustering of cell types indicating cell-to-cell differences. Particularly, Acinar, Ductal, Endothelial, and Stellate cells formed isolated clusters similar to Baron et al. (2016). However, Alpha, Beta, Basal, Gamma, and Dendritic cells showed less separation. This difference could be due the sample size used for this project. Baron et al. (2016), identified nine clusters using four human samples, while this project only used data from one human sample.

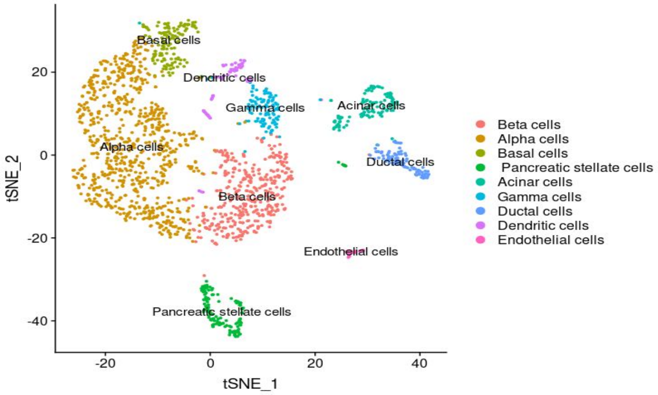


Figure 5. tSNE projection of clustered cells. Clusters are labelled by a specific cell type. Clear separation can be seen in Acinar, Ductal, and Endothelial cells.

Discussion

This project attempted to implement a single-cell analysis pipeline using the sc-RNA-seq data associated with a 51-year-old female donor from Baron et. al. In general, this project was successful in identifying distinct cells clusters and identifying cell types associated with those clusters. Compared to the 14 cell clusters identified by Baron et al., this project was able to identify 11 cell clusters. Furthermore, from the nine major cell clusters identified by Baron et al., seven of the nine cell clusters were similar to the major clusters identified in this project. Overall, this project was partly successful in reproducing the primary findings of Baron et al. (2016). The primary findings of Baron et al. (2016) could potentially be fully reproducible if analysis is performed with complete dataset.

References

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B.K., Shen-Orr, S. S., Klein, A. M., Melton, D. A., & Yanai, I. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell systems, 3(4):346–360.e4. https://doi.org/10.1016/j.cels.2016.08.011

Franzén, O., Gan, L.M., Björkegren, J.L.M. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data, Database, 2019:baz046. https://doi:10.1093/database/baz046

Srivastava, A., Malik, L., Smith, T. et al. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. Genome Biol 20, 65 (2019). https://doi.org/10.1186/s13059-019-1670-y

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W.M., Stoeckius, M., Smibert, P., Satija R. (2018). Comprehensive integration of single cell data. bioRxiv 460147. doi: https://doi.org/10.1101/460147