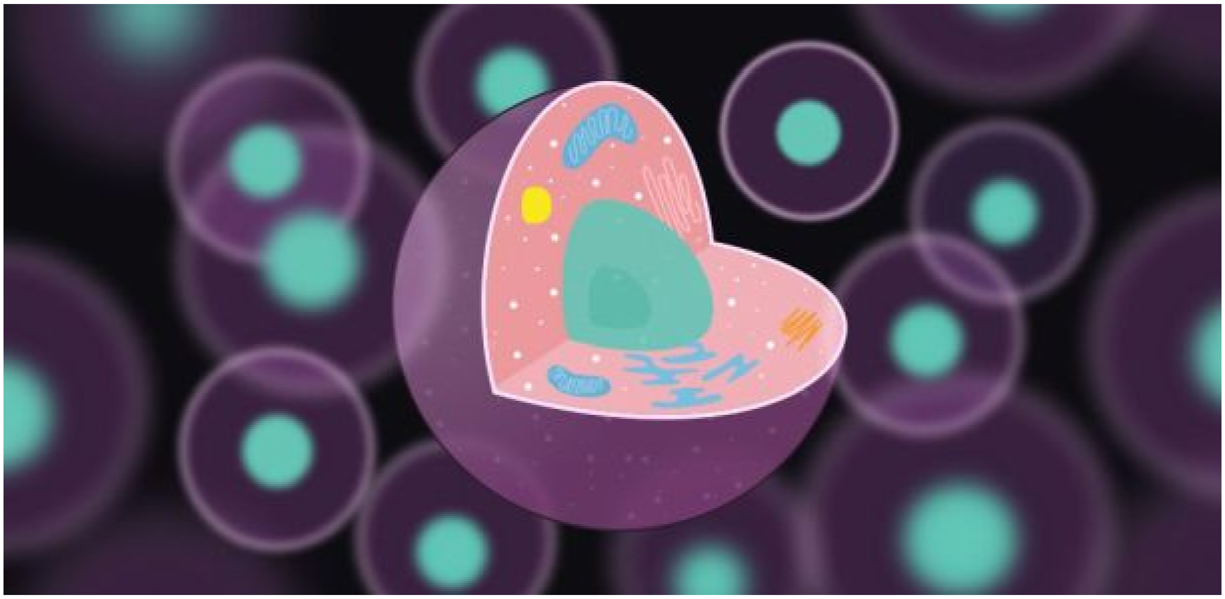
# Single Cell RNA- Seq Analysis of Pancreatic Cells



(Source:https://heise.cloudimg.io/width/2732/q50.png-lossy-50.webp-lossy-50.foil1/_www-heise-de_/imgs/18/1/7/6/6/1/1/0/googledocs-3a4a3a6a096b9117.jpeg)

**TA**

Kritika Karri

**PREPARED BY**

Garima Lohani

(Programmer and Analyst role)

# INTRODUCTION

The investigation of cellular diversity in Pancreas is important in the field of bioinformatics. The pancreas is involved in energy homeostasis (2). About 95 percent of the cells in the pancreas is composed of exocrine cell types. Acinar and duct cells make up the exocrine cell type. Acinar produces digestive enzymes (3),  and duct cells produce bicarbonates (4) and help in the transportation of digestive enzymes. On the other hand, the rest of the 5 percent of endocrine cells consist of alpha, beta, gamma, delta, and epsilon cells that secrete hormones for glucose regulation in humans (5). So dysfunction of the pancreas causes pancreatitis, type 1 diabetes, and type 2 diabetes and cancer (1). Currently, there is a lot of research-based on the replacement of beta cells in type 1 diabetes and to produce beta cells in vitro (6)(7). The most challenging aspect of the research is all cell types in the pancreas are interlinked functionally (1). Therefore, we need a technique to differentiate these different cell types at a molecular level to understand their roles in diseases.

Though RNA-sequencing is a powerful tool for the transcriptomic study, it cannot capture the heterogeneous population of cells at a high-throughput scale (8). However, droplet-based single-cell RNA-sequencing has emerged as a well-known method that captures thousands of cells using unique barcode sequences providing insights into cell types within a tissue (9). Baron et al used droplet-based RNA sequencing to classify the cell types within pancreatic cells of four humans and two mouse strains (1). The authors recognized all the previously known cell types in the samples. Also, they found distinct gene expression within ductal and beta cells. But in this project, we analyze data from a 51-year-old female human donor (1) and perform filtering, clustering, identification of novel markers, and labeling cell clusters to cell type to give biological significance to results.

# DATA:

The human single-cell RNA-Seq dataset samples in this study were downloaded from the NCBI GEO GSE84133 (1). The samples associated with the 51-year-old female donor in SRR files were  SRR3879604, SRR3879605 , and SRR3879606 (1). I used the pre-computed data for both the programmer and analyst roles.

# METHODS:

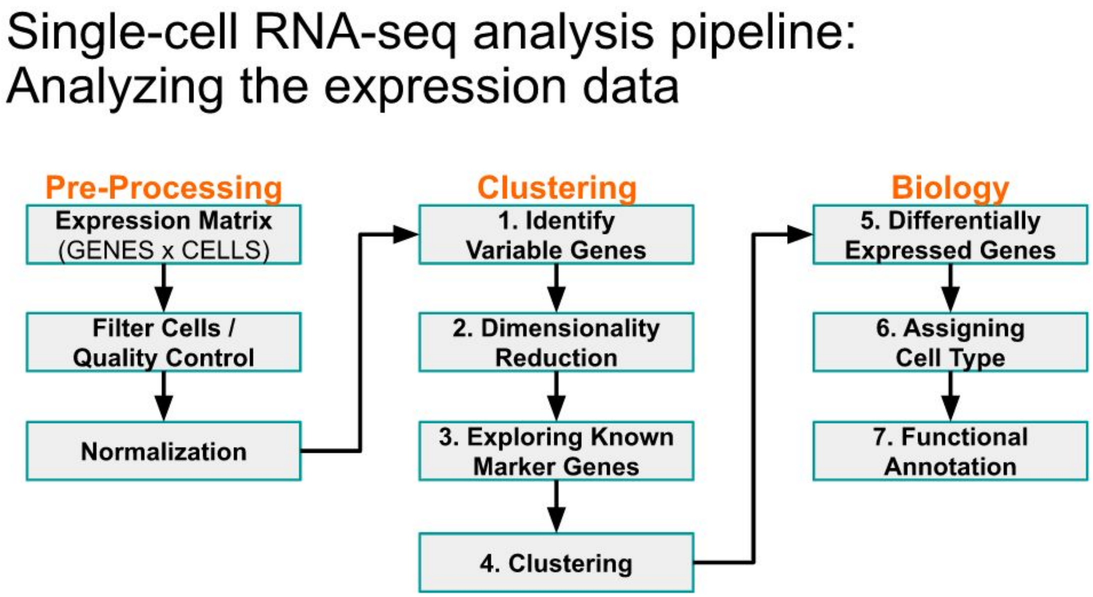

Single-cell RNA-seq analysis pipeline:
Analyzing the expression data
Pre-Processing
Expression Matrix (GENES x CELLS)
Filter Cells / Quality Control
Normalization
Clustering
1. Identify Variable Genes
2. Dimensionality Reduction
3. Exploring Known Marker Genes
4. Clustering
Biology
5. Differentially Expressed Genes
6. Assigning Cell Type
7. Functional Annotation

**Figure 1:** The schematic diagram of the single-cell sequencing pipeline. (10)

Overall, the procedure followed in this project can be summarized as given in figure 1. The major steps were filtering, clustering, identifying gene markers, labeling clusters, visualization of top markers, and finding novel markers.

**Filtering of cells and genes**

While preprocessing the data, the Alevin tool (11) generated an output file, the Unique Molecular Identifier (UMI) count matrix, that contains gene counts per cell. With the help of the tximport package, I successfully loaded the UMI count matrix into R and converted the UMI count matrix to the Seurat object with the help of the Seurat package (12). To remove low-quality cells, I performed two filterings. First of all, filtered out genes that are detected in less than 3 cells, and cells that are detected in less than 200 features. Second, filtered out cells with expressed RNA between 200 and 2500 at the mitochondrial percentage of 5 percent. To filter out low variance genes, I employed a log normalize method to keep the top 2000 high variance genes. Also, scaled all genes to have means of zero and variance of one. The Ensembl gene identifiers were also mapped to the gene symbol with the help of convertEnsembl2Symbol function in the SeqGsea package (13).

### Clustering of cells

During clustering, Principal Component Algorithm (PCA) was applied to gene data to perform linear dimensionality reduction. I produced various PCA plots, such as the Jackstraw plot, Elbow plot, and heatmaps. After the construction of a shared-nearest-neighbors graph, I identified the number of clusters in the gene data. Finally, I visualized the identified clusters in the t-SNE plot.

### Identification and labeling of the gene marker for each cluster

I identified gene markers using the function FindAllMarkers in the Seurat package (12). The three parameters taken into consideration were a limit of 0.25 min.pct, the threshold of 0.25 log fold change, and the true value set for the positive marker gene.

I used the gene markers present in the Baron et al's supplementary table 2 in the analyses. I searched Panglaodb (14) to determine the cell type for some of the cell clusters. To get more clarity, the gene set enrichment analysis was performed using Enrichr(15)(16) to determine the functional annotation of each cluster. Further, I did the literature search to label each cell cluster to different cell types.

### Visualization of clustered cells and top markers

The t-SNE plot was produced with a labeled cell type for each cluster. Using the Feature plot function in the Seurat package (12), I visualized top markers from each cluster. Besides, I produced a heatmap to show UMI counts for a gene across cells.

### Find novel markers gene

I used the threshold of 0.05 adjusted p-value to get differentially expressed genes, and then filter out genes that already exist in the gene marker list.

## RESULTS:

### The quality control of UMI count matrix:

There were originally 60233 genes with 108832 cells in an unfiltered dataset. After filtering out low-quality cells there were 26442 genes with 15147 cells. In order to ensure quality, we plotted the Violin plot and Scatter plot as shown in Figure 2A-2C. After variance filtering, there were 2000 genes with 14080 cells. I plotted average expression values vs standard variance as shown in Figure 2D.

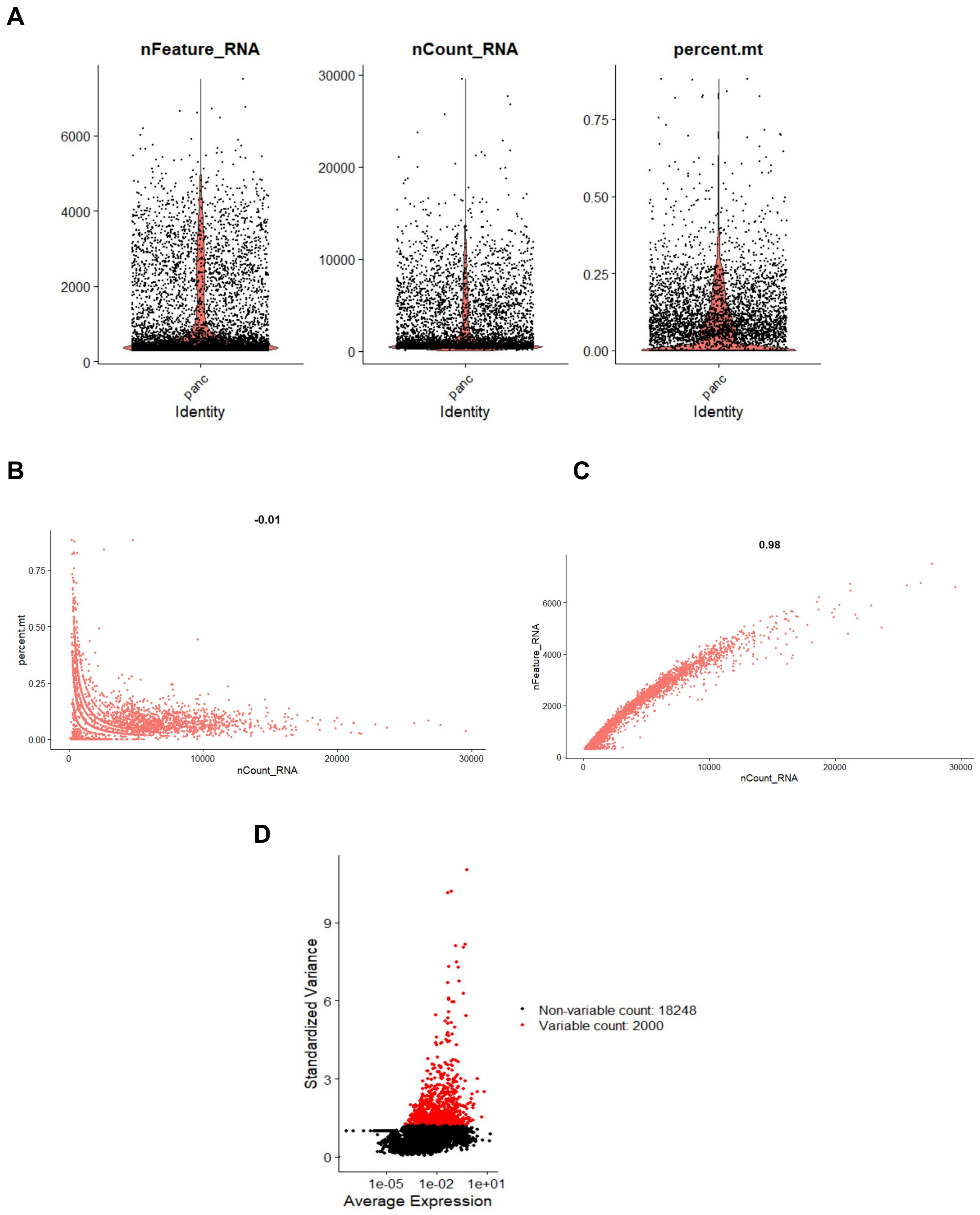


**Figure 2:** The quality control of the UMI matrix. A)Violin plots  B) Scatter plot for mitochondrial percent C)Scatter plot for RNA feature D)Scatter plot of 2000 variable features. The black dots indicate genes that are filtered out and red dots are top 2000 variance genes kept for the analyses.

**Identification of cell clusters:**

The number of clusters identified in the dataset was 13 and the proportion of cells in each cluster is shown in Figure 3A-3B. The PCA results were visualized using VizDimLoadings, Dimplot, and DimHeatmap as shown in Figure 4A,4B,4E. Further, to determine a significant principal component, JackStrawPlot was produced as shown in Figure 4C. The elbow plot was also used to determine the dimensionality of the dataset as shown in Figure 4D. Finally, the t-SNE plot was produced for the given clusters as shown in Figure 4F.

**A**

**Table 1:The number of cells in each cluster**

| **cluster id** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **number of cells** | 2859 | 2170 | 1376 | 1202 | 1136 | 1036 | 997 | 744 | 650 | 637 | 625 | 324 | 324 |

**B**

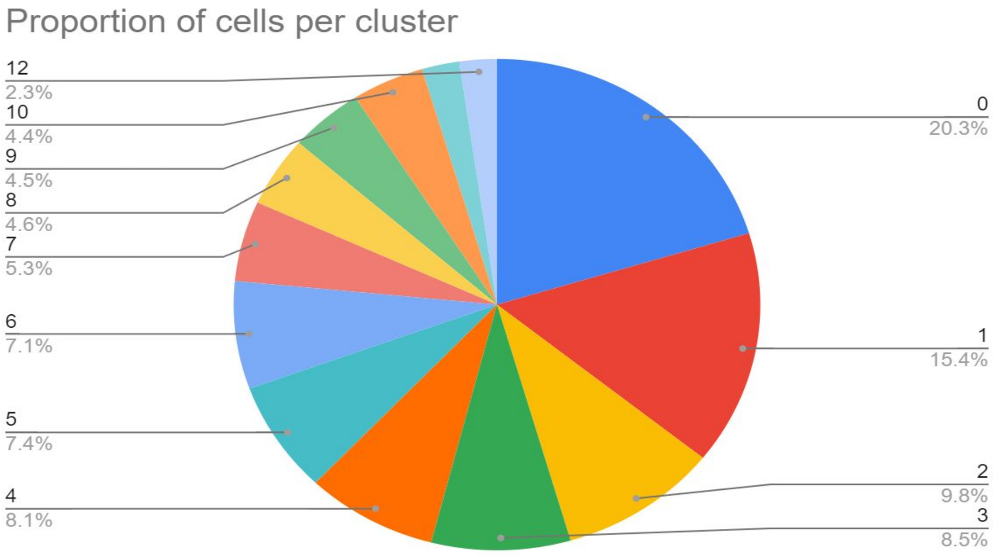


**Figure 3**: A) The number of cells in each cluster summarized in a table. B) Pie chart showing the proportion of cells in each cluster.

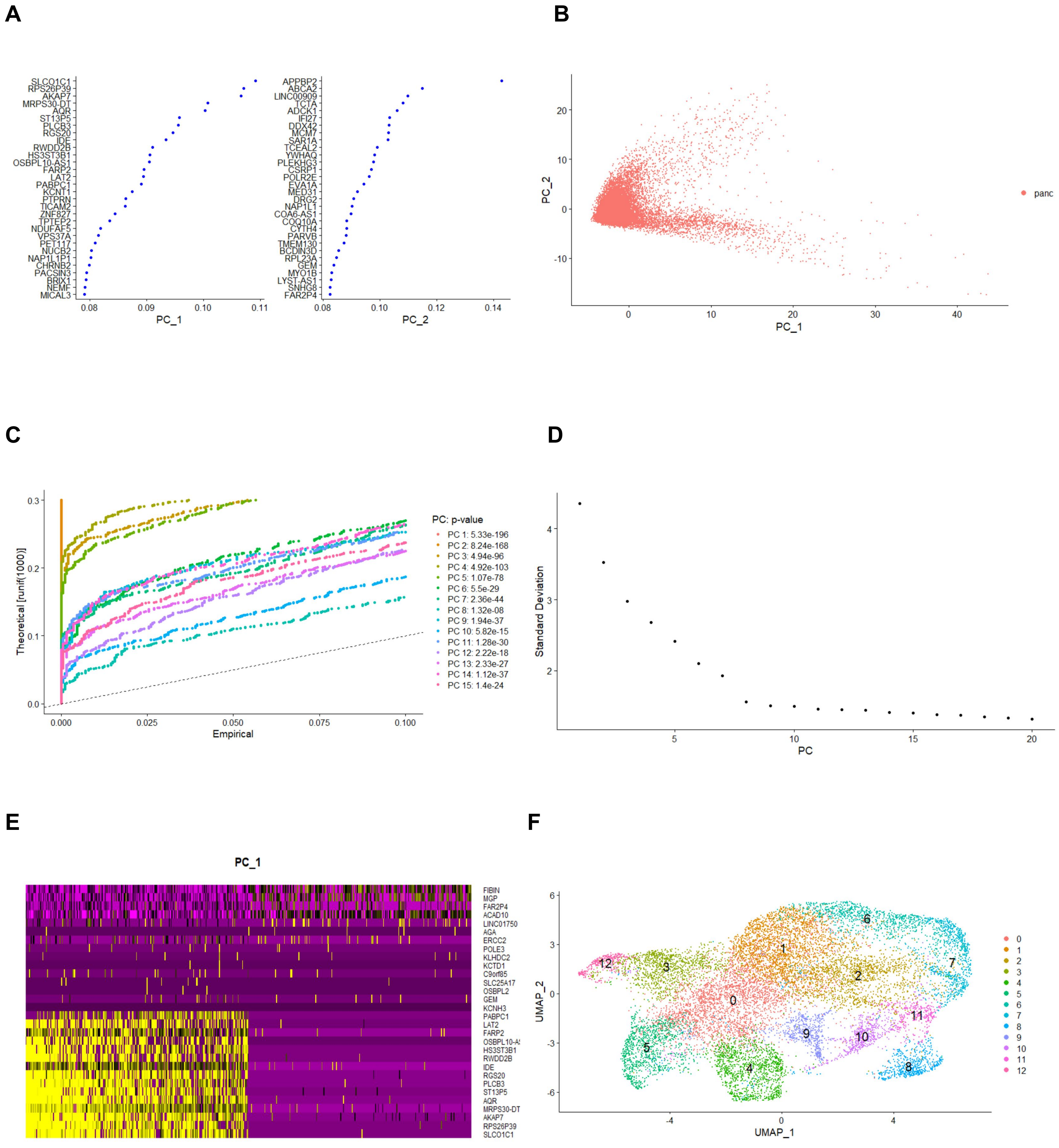


**Figure 4:** Visualization of PCA results and t-SNe plot.A)The scatter plot of the gene along PC1 and PC2. B)Scatter plot of PC1 and PC2. C)Jackstraw plot for each PCA. D)Elbow plot for each PCA flattens around PCA 10. E) The heatmap for the genes in the PC1.F) t-SNE plot for the given cluster

**Classification of each cell cluster to a cell type:**

The 13 cell clusters identified and labeled as shown in Table 2. Among the cluster-specific gene markers used in the paper, I identified gene markers- SST, INS, GCG, KRT19, PDGRFB, and CPA1 in my analyses. I could not find a known marker for cluster 7 and 12. I further did a literature search and used Panglaodb (14). Then I concluded clusters 7 and 12 to be cytotoxic T and mast cells respectively.

**Table2: Top differentially expressed genes and cell clusters assigned to their cell types**

| cluster id | cell type label | gene symbol | avg_logFC | pct.1 | pct.2 | p_val_adj |
|---|---|---|---|---|---|---|
| 0 | Delta | SP100 | 0.96 | 0.707 | 0.617 | 3.03E-118 |
| 1 | Beta | INS | 1.21 | 0.937 | 0.62 | 0 |
| 2 | Alpha | FN1 | 1.35 | 0.273 | 0.059 | 3.48E-219 |
| 3 | Acinar | REG1B | 2.47 | 0.712 | 0.12 | 0 |
| 4 | Alpha | TTR | 1.64 | 0.899 | 0.413 | 0 |
| 5 | Ductal | KRT19 | 1.89 | 0.44 | 0.096 | 3.51E-282 |
| 6 | Beta | EEF1A2 | 1.13 | 0.72 | 0.061 | 0 |
| 7 | Cytotoxic T | ACER3 | 1.80 | 0.563 | 0.17 | 1.84E-128 |
| 8 | Alpha | GC | 1.45 | 0.865 | 0.04 | 0 |
| 9 | Stellate | COL1A1 | 3.11 | 0.733 | 0.075 | 0 |
| 10 | Ductal | CRP | 2.06 | 0.502 | 0.018 | 0 |
| 11 | Acinar | ALDOB | 2.75 | 0.931 | 0.035 | 0 |
| 12 | Mast | ACP5 | 3.91 | 0.788 | 0.01 | 0 |

**Visualization of cell clusters and top gene marker in each cluster:**

The cell clusters were visualized in the t-SNE plot and FeaturePlot as shown in Figure 5. In comparison to Figure 1D of the paper, I was able to identify distinct 8 cell types namely acinar, ductal, alpha, beta, delta, stellate, cytotoxic T, and mast cells. The expression level in the heatmap associated with the top 2 differentially expressed genes is shown in Figure 5B. In comparison to Figure 1B of Baron et al, there is a clear expression pattern of these genes. GCG and GC genes are highly expressed for alpha cells. Similarly, KRT19 can be seen as highly expressed in ductal cells. Furthermore, the top markers from each cluster are plotted in FeaturePlot as shown in Figure 6.

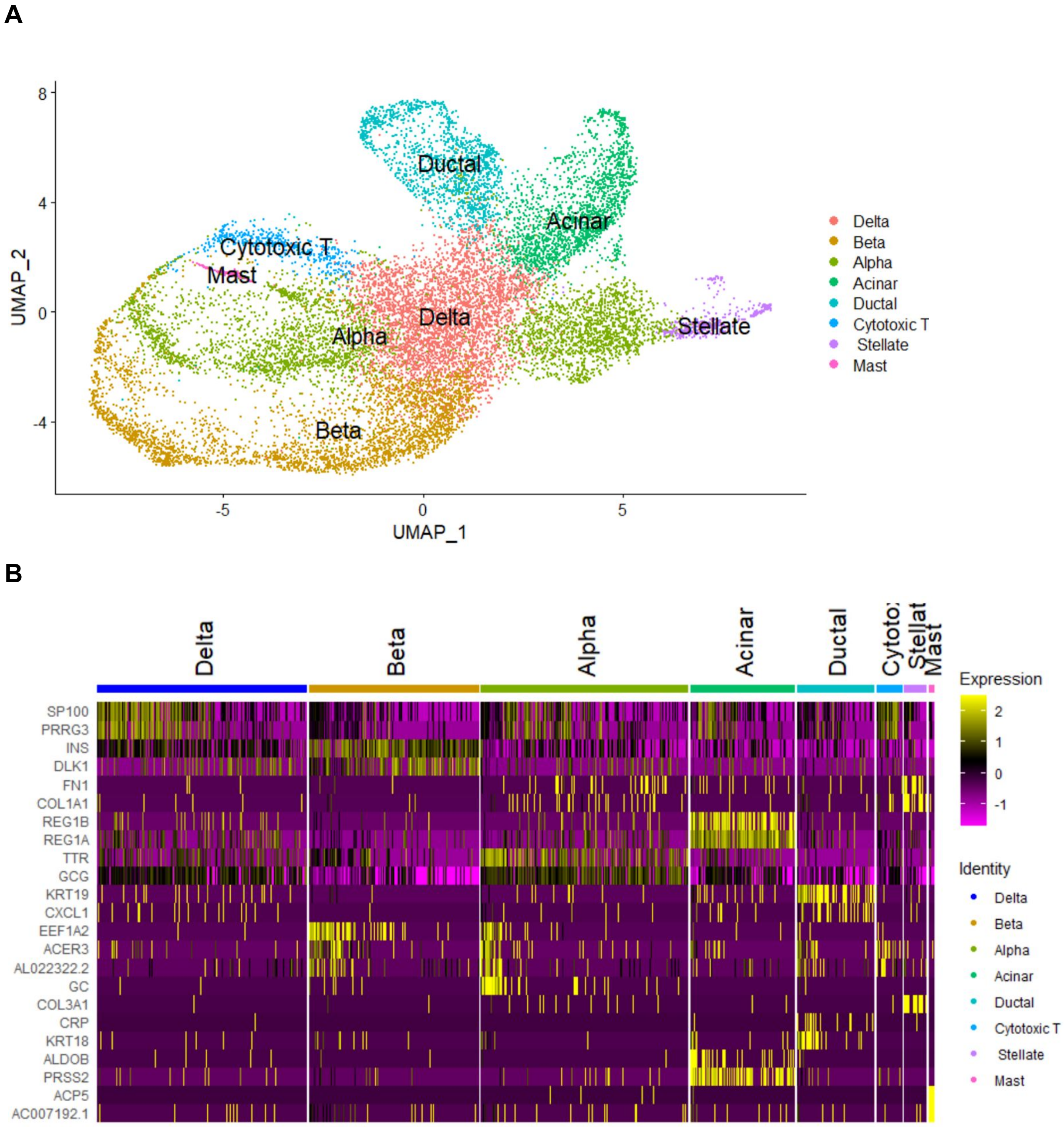


**Figure 5:** A)The t-SNE plot with labeled cell type. B) The heatmap associated with the top 2 genes. The yellow color indicates a highly expressed gene in the corresponding  cell

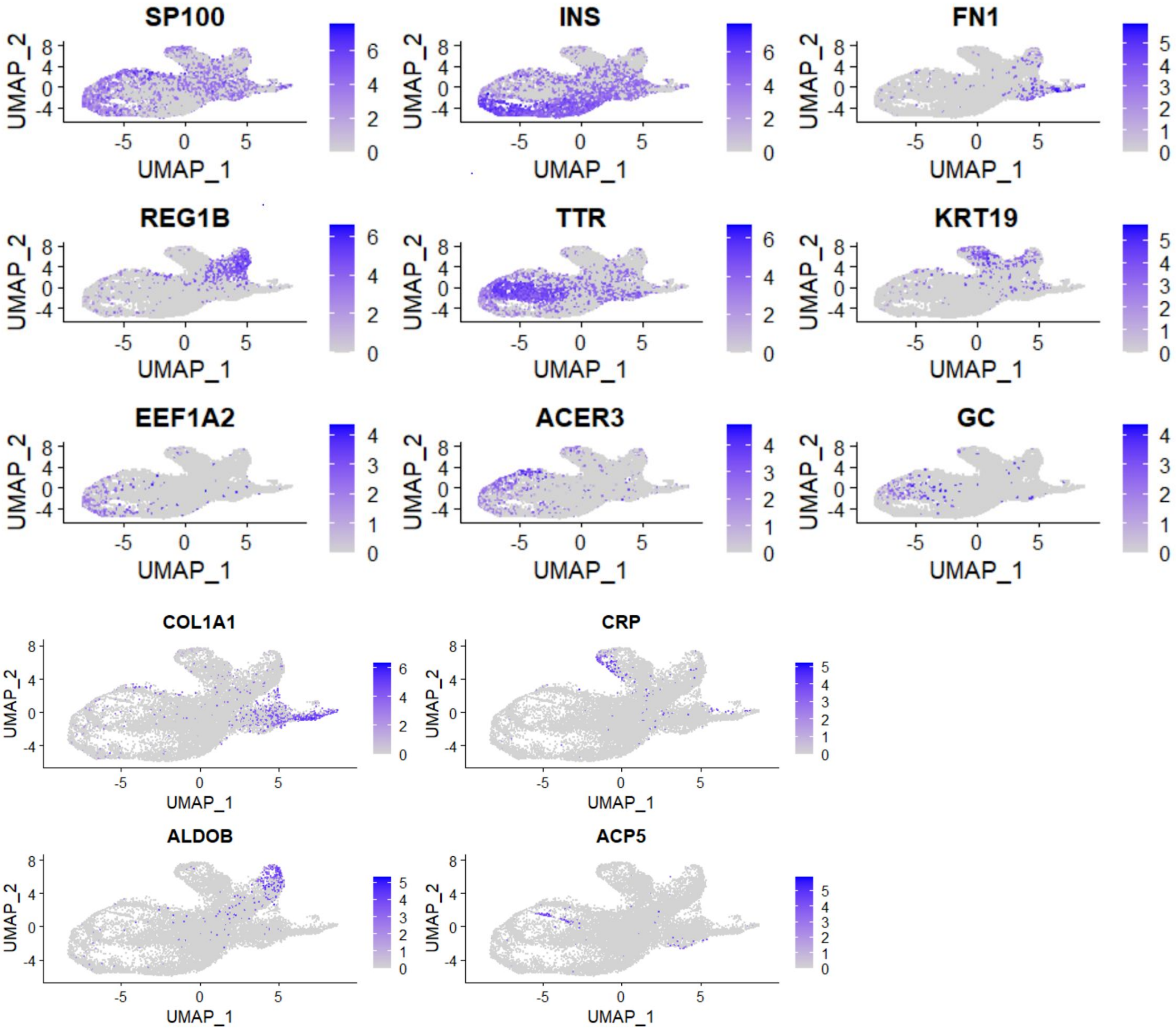


**Figure 6:** Top markers expressed in different regions in FeaturePlot in each cluster.

**Discovery and visualization of the novel marker in each cluster:**

The top novel gene markers discovered for each cluster are summarized in Table 3. Also, the top novel markers were visualized using FeaturePlot as shown in Figure 7.

**Table 3: The top novel markers in each cluster.**

| gene symbol | cluster id | cluster name | avg_logFC | pct.1 | pct.2 | p_val_adj |
|---|---|---|---|---|---|---|
| PHACTR1 | 0 | Delta | 0.78 | 0.22 | 0.166 | 3.08E-20 |
| G6PC2 | 1 | Beta | 1.06 | 0.23 | 0.105 | 1.20E-64 |
| COL1A2 | 2 | Alpha | 1.15 | 0.209 | 0.04 | 5.16E-177 |
| PLA2G1B | 3 | Acinar | 1.83 | 0.231 | 0.025 | 0 |
| IRX2 | 4 | Alpha | 1.23 | 0.199 | 0.058 | 6.12E-90 |
| KRT7 | 5 | Ductal | 1.37 | 0.238 | 0.054 | 4.39E-130 |
| APOBEC2 | 6 | Beta | 0.44 | 0.228 | 0.007 | 0 |
| SLC25A51P4 | 7 | Cytotoxic T | 1.10 | 0.245 | 0.046 | 3.67E-82 |
| AL606807.1 | 8 | Alpha | 0.52 | 0.214 | 0.006 | 0 |
| CRLF1 | 9 | Stellate | 1.36 | 0.232 | 0.004 | 0 |
| EPHA2 | 10 | Ductal | 0.43 | 0.236 | 0.003 | 0 |
| AC008760.2 | 11 | Acinar | 0.36 | 0.181 | 0.001 | 0 |
| FYB1 | 12 | Mast | 1.55 | 0.24 | 0 | 0 |

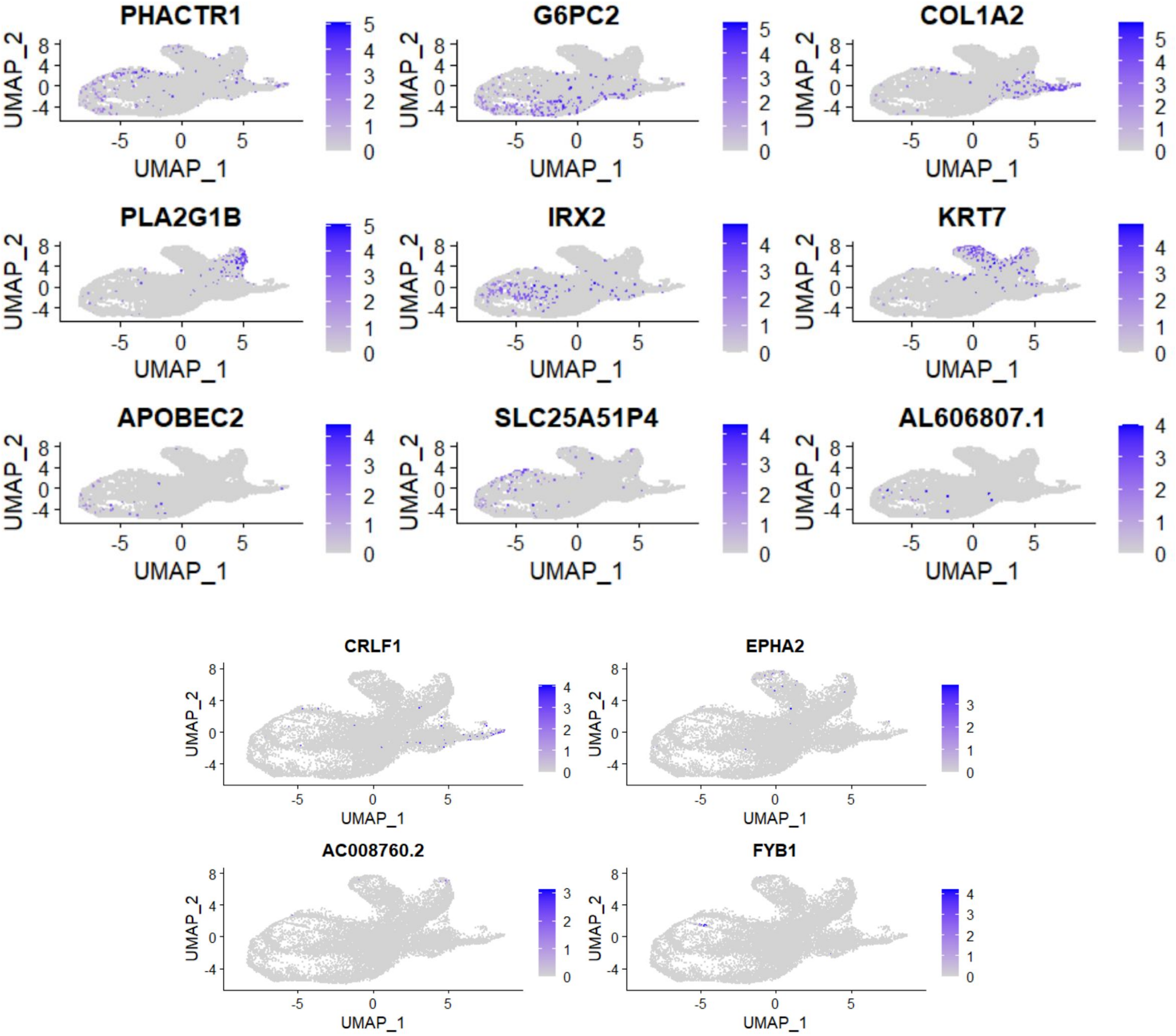


**Figure 7:** Visualization of top novel markers in FeaturePlot.

# DISCUSSION:

I found both exocrine cells: acinar and ductal cell types. The authors in this paper used CPA1 in their analyses for the identification of acinar cells.  CPA1 genes were significantly expressed in clusters 3 and 11. The enriched pathway associated with this gene was the pancreatic secretion pathway. This is in agreement with the function of acinar cells, which are known to secrete digestive enzymes (17). So clusters 3 and 11 strongly suggest being acinar cells. In the case of ductal cell type, the authors in the paper used KRT19 in their analyses. This gene was significantly expressed in clusters 5 and 10. Moreover, the CFTR gene, which functions in secretion of HCO3 (1),  was highly expressed in clusters 10. So clusters 5 and 10 strongly indicate being ductal cells.

Out of five endocrine cells, I found three cell types: alpha, beta, and delta cells. I have found three clusters 2,4, and 8 to be alpha cells. GCG was highly expressed in each of these clusters. The enriched pathways are glucagon signaling in metabolic regulation and ghrelin biosynthesis, secretion, and deacylation pathway.  These are in agreement with the function of the alpha cell to increase the glucose level in the human body (18). Hence it strongly implies these clusters be alpha cells. Continuing, I discovered clusters 1 and 6 to be beta cells because INS genes appeared as the top marker. The INS gene is known to provide instructions for producing the hormone insulin (20). The enriched pathway associated is gene expression regulation in the pancreatic beta cells. Also, enriched terms such as type 1 diabetes, appeared that can be related to the loss of beta cells. So it highly agrees with the function of beta cells, which is to secrete insulin and regulate blood glucose level (20). I discovered cluster zero as a delta cell. SST is highly expressed in this cluster with the p-adjusted value of 1.85E-296 and an average log fold change value of 0.885846912. SST is associated with pathway ghrelin-mediated regulation of food intake and energy homeostasis. It functions to inhibit insulin and glucagon secretion and thus regulates alpha-cell and beta-cell function (19). Therefore, cluster zero strongly implies to be a delta cell.

Further, I  found cluster 9 as a stellate cell type. In the paper, it was reported that they found stellate cells in each of the donors. The authors used the PDGFRB  gene in analysis for stellate cells. It was highly expressed in cluster 9. The authors in this paper state that activated stellate cells are known to produce large quantities of collagen, fibronectin, and extracellular matrix component (1).  The enriched pathways present are collagen biosynthesis and modifying enzymes, extracellular matrix organization, and PDGF genes and receptors. So it strongly implies this cluster is activated stellate cells as found in the paper.

Finally, I found two immune cell types: cytotoxic T and mast cells. CD226 is highly expressed in cluster 7 with a fold change of 1.42 and a p-adjusted value of 1.34E-100. It is known to function to regulate natural killer cell anti-tumor responses (21). The enriched terms appeared is the immune system. Thus, it is most likely to be cytotoxic T. Continuing, CD300A,  known to be significantly expressed in human mast cells (22), was present in cluster 12. So it is highly likely to be mast cells in the pancreas.

# CONCLUSION:

I successfully applied quality control metrics on a given UMI count matrix and reduced a large dataset of 60233 genes with 108832 cells to a smaller dataset of 2000 genes with 14080 cells. I found 13 numbers of cell clusters in the dataset. Then I identified the gene marker and assigned each cluster to a cell type. So, I found 8 different types of the cell type of pancreatic islet namely alpha, beta, delta, cytotoxic T, stellate, ductal, mast, and acinar. Further, I investigated the cell type of each cluster to provide biological significance. Overall, my results were synonymous with the paper.

Though I was able to analyze single-cell RNA sequencing data from a 51-year-old female donor, one major challenge that I encountered in the project was with the software installation of packages in the R environment. To overcome this challenge, I had to update R versions and all dependencies several times.

# REFERENCES:

1) Baron M, Veres A, Wolock SL, et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst . 2016;3(4):346–360.e4. doi:10.1016/j.cels.2016.08.011 .

2) Kimmel RA, Meyer D. Molecular regulation of pancreas development in zebrafish. Methods Cell Biol. 2010;100:261–280.

3) Whitcomb DC, Lowe ME. Human pancreatic digestive enzymes. Dig Dis Sci. 2007;52:1–17

4) Steward MC, Ishiguro H, Case RM. Mechanisms of bicarbonate secretion in the pancreatic duct. Annu Rev Physiol. 2005;67:377–409.

5) Mastracci TL, Sussel L. The endocrine pancreas: insights into development, differentiation, and diabetes. Wiley Interdiscip Rev Dev Biol. 2012;1:609–628.

6) Nostro MC, Sarangi F, Yang C, Holland A, Elefanty AG, Stanley EG, Greiner DL, Keller Stem Cell Reports. 2015 Apr 14; 4(4):591-604.

7) Pagliuca FW, Millman JR, Gürtler M, Segel M, Van Dervort A, Ryu JH, Peterson QP, Greiner D, Melton DA Cell. 2014 Oct 9; 159(2):428-39.

8) Blodgett DM, Nowosielska A, Afik S, Pechhold S, Cura AJ, Kennedy NJ, Kim S, Kucukural A, Davis RJ, Kent SC, Greiner DL, Garber MG, Harlan DM, diIorio P Diabetes. 2015 Sep; 64(9):3172-81.

9) Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW Cell. 2015 May 21; 161(5):1187-1201.

10)https://broadinstitute.github.io/2019_scWorkshop/identifying-cell-populations.html#google-slides

11) Srivastava, A., Malik, L., Smith, T. et al. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. Genome Biol 20, 65 (2019).

12) Butler, A., Hoffman, P., Smibert, P. et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36, 411–420 (2018).

13) Xi Wang, Murray J. Cairns, SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing, Bioinformatics, Volume 30, Issue 12, 15 June 2014, Pages 1777–1779.

14) Oscar Franzén, Li-Ming Gan, Johan L M Björkegren, PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data, Database, Volume 2019, 2019.

15) Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;128(14).

16) Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Research. 2016; gkw377 .

17) Williams JA. Regulation of acinar cell function in the pancreas. Curr Opin Gastroenterol . 2010;26(5):478-483. doi:10.1097/MOG.0b013e32833d11c6

18) https://www.uniprot.org/uniprot/P01275

19) Hauge-Evans AC, King AJ, Carmignac D, et al. Somatostatin secreted by islet delta-cells fulfills multiple roles as a paracrine regulator of islet function. Diabetes. 2009;58(2):403-411. doi:10.2337/db08-0792

20) Chen C, Cohrs CM, Stertmann J, Bozsak R, Speier S. Human beta cell mass and function in diabetes: Recent advances in knowledge and technologies to understand disease pathogenesis. Mol Metab . 2017;6(9):943-957. Published 2017 Jul 8. doi:10.1016/j.molmet.2017.06.019

21) Bottino C, Castriconi R, Pende D, Rivera P, Nanni M, Carnemolla B, Cantoni C, Grassi J, Marcenaro S, Reymond N, Vitale M, Moretta L, Lopez M, Moretta A (August 2003). "Identification of PVR (CD155) and Nectin-2 (CD112) as cell surface ligands for the human DNAM-1 (CD226) activating molecule". The Journal of Experimental Medicine. **198** (4): 557–67.

22) The Inhibitory Receptor IRp60 (CD300a) Is Expressed and Functional on Human Mast Cells
Ido Bachelet, Ariel Munitz, Alessandro Moretta, Lorenzo Moretta and Francesca Levi-Schaffer
J Immunol December 15, 2005, 175 (12) 7989-7995.

23)https://heise.cloudimg.io/width/2732/q50.png-lossy-50.webp-lossy-50.foil1/_www-heise-de_/imgs/18/1/7/6/6/1/1/0/googledocs-3a4a3a6a096b9117.jpeg