Project 4: Single Cell RNA-Seq Analysis of Pancreatic Cells

Authors: Monica Roberts (Data Curator), Italo Duran (Programmer), Preshita Dave (Analyst)

Introduction

The mammalian pancreas performs essential functions in the body for energy homeostasis and depends on interactions between diverse cell types. Given the importance of the pancreas in diseases such as diabetes and cancer, it is vital to study these complex interactions. A large majority of the cells are of two different types, acinar and duct cells. A minority are islet cells, which have five distinct subtypes. Certain cell types are implicated in specific diseases, such as insulin-secreting beta cells in diabetes. Gene expression profiles have been studied for the pancreas; however, they are usually done with bulk RNA-seq. Characterizing these interactions and determining how certain cell types are changed at the molecular level in diseased states requires single-cell RNA sequencing.

Baron et al. [1] created a transcriptional profile and map of the human and mouse pancreas with a droplet-based, single-cell RNA-seq method. They analyzed 12,000 pancreatic cells from four post-mortem human donors and two mice. They were able to cluster the cells into 15 previously identified cell types and detected novel subpopulations. This project attempts to replicate the study's results using one human donor's cells. We will process the reads and barcodes and then align them to a reference transcriptome to perform the counting of the unique molecular identifiers (UMI). Clustering will then be performed to determine the number of cell types found based on marker genes.

Data

The human donor cells were from the National Disease Research Interchange (NDRI) and kept in CMRLS for 24-48 hours upon collection. The two mouse donors, strains ICR and C57BL/6, were from Jackson Laboratories and followed all appropriate protocols for animal research. The islets were isolated and pooled from 5 mice. The cells were collected from perfusion of the common bile duct, digestion of the pancreata, and purification by centrifugation. Purification was then followed by several cycles of centrifugation, incubation, and addition of various media. They were finally re-suspended in PBS with Optiprep at a 75,000 cells/mL concentration before being added to the inDrop device. The cells were encapsulated, and barcodes were successfully added to 80% of the cells.

The inDrop encapsulation and reverse transcription was carried out according to the protocol in Klein et al. [3]. This protocol uses hydrogel microspheres to introduce the barcodes. Each barcode is an oligonucleotide. Each cell is encapsulated into a single droplet with lysis buffer, reagents to carry out reverse transcription, and the barcode primers. There is a 1:1 ratio between cells and droplets. Once the cell is isolated into the droplet, the mRNA is released from the lysed cell, and each molecule is tagged with a barcode during the synthesis of cDNA. Slight modifications were made to eliminate the necessity of primer ligation, and the number of PCR cycles during library enrichment was 9-12. Each mRNA molecule also contained a unique molecular identifier (UMI) to correct PCR amplification bias. The RNA was then sequenced on Illumina Hiseq 2500 machine. The first read for every molecule contained the barcode and UMI. Cells that did not contain a known barcode or UMI were removed, and the remaining reads were trimmed. Before beginning our analysis, all barcodes were padded to be 19 + 6 UMI bases.

Methods

Processing Single Cell Sequencing Reads and Choosing Barcodes

The metadata from the study was downloaded from the GEO accession page to identify the samples to be used in the analysis correctly. There were 13 samples in the library from four individuals. Only the samples from the 51-year-old female donor were used for further analysis, which included SRR3879604, SRR3879605, and SRR3879606. The read one fasta files for each sample were used to count the number of reads per distinct barcode. A whitelist of informative barcodes was then created. Only barcodes with more than 1000 reads were included in this list. The objective of this step was to eliminate barcodes that were too infrequent to provide any information to the analysis. If the read counts were too low, something could have gone wrong in creating the library for that cell and should not be considered in the subsequent analysis. Next, the UMI matrix was created using the alevin command within the salmon software package. [4] All fasta files from each sample were used as input, and the whitelist of barcodes from all three samples. Read one file containing the barcodes and UMI, while read two contained the transcript. Alevin is a tool used to quantify reads in single-cell sequencing data. Salmon tools have a mapping-based mode, allowing counting and mapping to coincide via a built-in alignment algorithm. The reads are mapped to a reference transcriptome downloaded from Gencode. [5] An index was created from the human transcriptome using the salmon index tool to use this mode. A transcript ID to the gene mapping file was provided for the alevin tool to collapse from transcript level to gene level. This allows the tool to quantify the counts per gene. The output of the alevin tool includes a UMI matrix, which contains the counts per gene, and was used in the subsequent steps of the analysis.

Processing the UMI counts matrix

Instead of trying to implement the methods as in Baron et al. study [1]. The Bioconductor package, Seurat, was used to compute the UMI counts. After receiving the precomputed UMI count matrix files, the alevin_output quants_mat.gz file; was importing the file using the tximport function. After formatting the genes, Ensembl identifiers to gene symbols that we can use in Seurat objects. Using the EnsDb.Hsapiens.v79 library. The percentage of counts genes (MT-) was processed by calculating the PercentageFeatureSet() function. The UMI count matrix's quality control was visualized using the Violin Plot Fig. 4 for nFeature_RNA, nCount_RNA, and percent.mt. Cells with unique feature count over 2,500 or less than 200 and >5% mitochondrial counts were filtered. The FeatureScatter() function was used to visualize two distinct scatter plots shown in Fig. 5.

After filtering out low-quality cells, it was followed by filtering out low variance genes by normalizing the data using NormalizeData() function that normalizes the feature expression measurements for each cell by the total expression using a scale factor of 10,000 parameters.

To Identify clusters of cell type subpopulations, detection of variable genes for downstream analysis was performed using the FindVariableGene() function in Seurat. This function assists in estimating the relationship between average expression and variability. Scaling the data can help remove unwanted sources of variation; the data could have technical noise from batch handling errors or biological noise caused due to variation in the cell. Data scaling was performed using Seurat's ScaleData function that regresses the total number of RNA molecules detected within a cell; nCounts_RNA; and mitochondrial percentage.

The PCA linear reduction analysis was performed on the scaled data using the RunPCA() and VariableFeature() function. The JackStraw function was used to find significant PCs with strong enrichment of low p-value genes. To visualize, the JackStrawPlot function was used in Figure 9. Graph-based clustering is the default clustering method in Seurat for identifying the cell clusters. For the UMAP in figure 11, the KNN(k-nearest neighbor graph)graphs were generated based on euclidean distance in PCA space, and the edge weights among cells were refined based on the shared overlap. The default Louvain algorithm was utilized to cluster the cells, and two functions from Seurat were applied: FindNeighbors() function and FindClusters().

Identifying Marker Genes for Each Cluster

After normalization, the data was scaled and clustered into groups. The data was clustered into 13 groups, and marker genes for each group were found through the Seurat's FindAllMarkers() function, which uses a Wilcoxon Rank Sum test by default. Corresponding log2 Fold-Change and p-value statistics were also included—only markers with a positive log2 Fold-Change and an adjusted p-value < 0.05 were selected. To identify the type of pancreatic cell that each cluster represented, the top 10 significant genes as measured by log2 Fold-Change and p-value were searched for, along with identifying the clusters which expressed a particular marker as highlighted in the supplementary material in Baron et al. [1]. PanglaoDB [2] was also used to identify clusters through the list of marker genes expressed by a cluster. The top marker genes were visualized across all clusters in a Violin Plot using R's VlnPlot() and a feature plot using R's FeaturePlot() commands. A non-linear dimensional reduction plot was rendered using Seurat's RunUMAP() function, which helped visualize how the different clusters separated into space. The 5 top marker genes per cluster were visualized through the DoHeatmap() function and novel markers for each cluster were found through the FindMarkers() function, which had strict thresholds set for different parameters.

Results

Reads per Distinct Barcode

The cumulative distribution of reads per distinct barcode was visualized to understand how the number of reads was distributed across the barcodes. The cumulative distribution was plotted for each sample and is shown below in figures 1, 2, and 3.

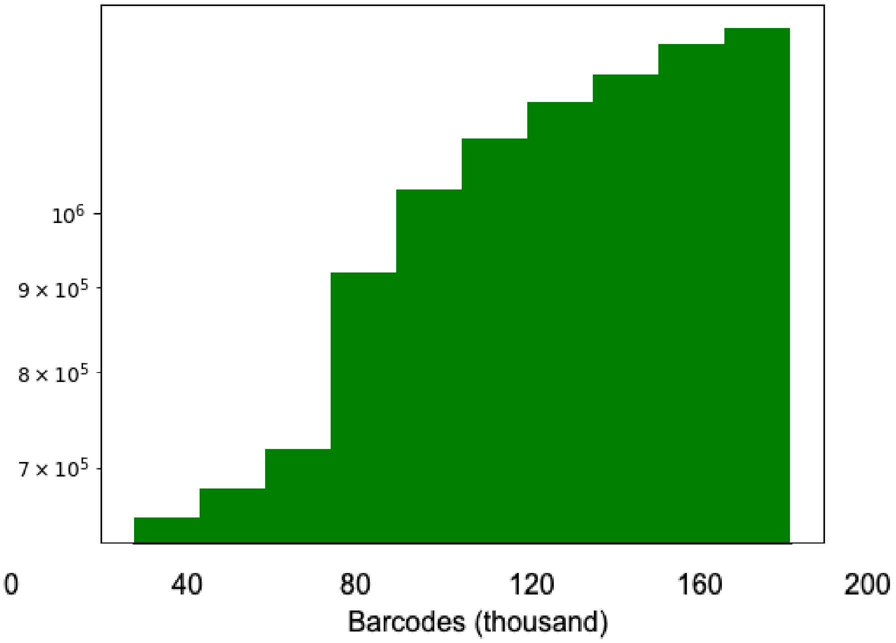

Figure 1: *Cumulative distribution of the number of reads per barcode for sample SRR3879604.*

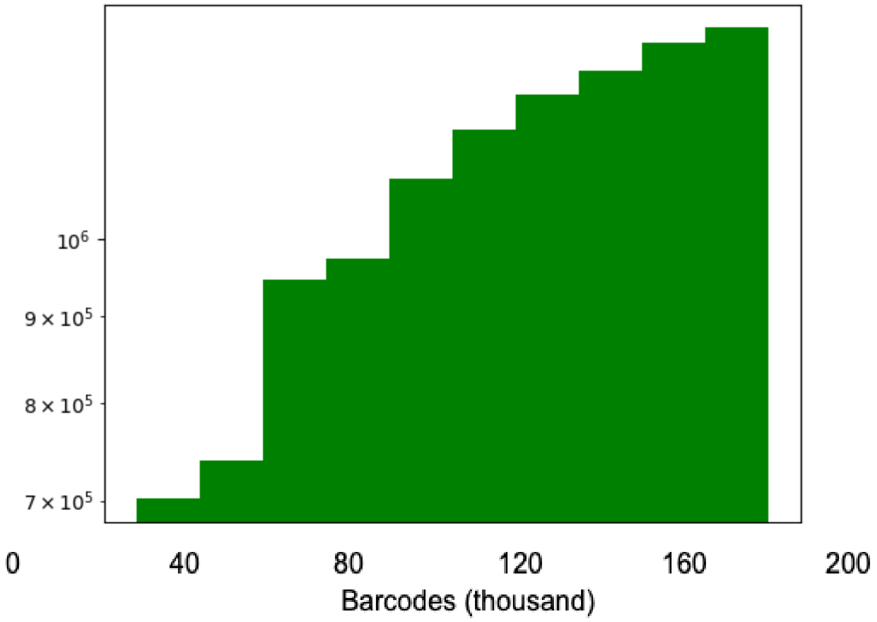

Figure 2: *Cumulative distribution of the number of reads per barcode for sample SRR3879605.*

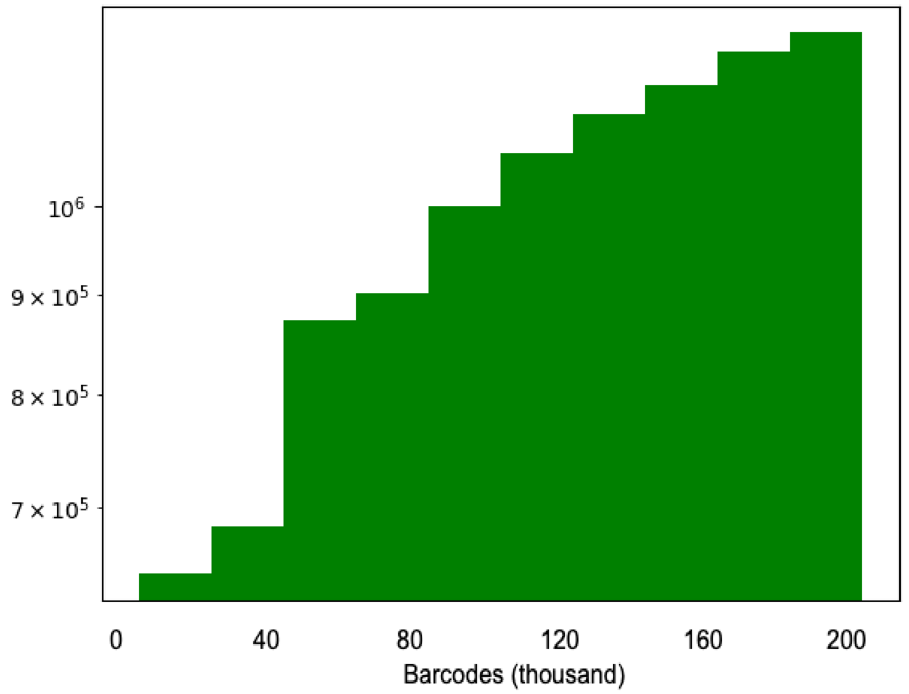

Figure 3: *Cumulative distribution of the number of reads per barcode for sample SRR3879606.*

The number of distinct barcodes per sample before filtering is shown in Table 1. The number after barcodes with less than 1000 reads were filtered out is shown in Table 2.

Sample	# of Distinct Barcodes
SRR3879604	1,293,792
SRR3879605	1,333,842
SRR3879606	1,227,152

Table 1: *Number of distinct barcodes per sample.*

Sample	# of Distinct Barcodes
SRR3879604	25,591
SRR3879605	20,447
SRR3879606	20,815

Table 2: *Number of distinct barcodes per sample after barcodes with less than 1000 reads were filtered out.*

Salmon Alevin Mapping Statistics: The salmon alevin tool's output included mapping statistics. The summary statistics are shown below in Table 3.

Mapping Statistic	Number
Total Reads	1,324,837,961
Total Reads with ≥1 Nucleotide in UMI	67,930
Mapping Rate	40.88
Total # Barcodes Observed	4,251,176
Total # Barcodes Used	201,913
Final Total # Barcodes	51,015
Total # UMIs Post-Deduplication	20959656
Mean # UMIs per Cell	410
Mean # Genes per Cell	257

Table 3: *Mapping statistics of alignment performed by salmon alevin tool.*

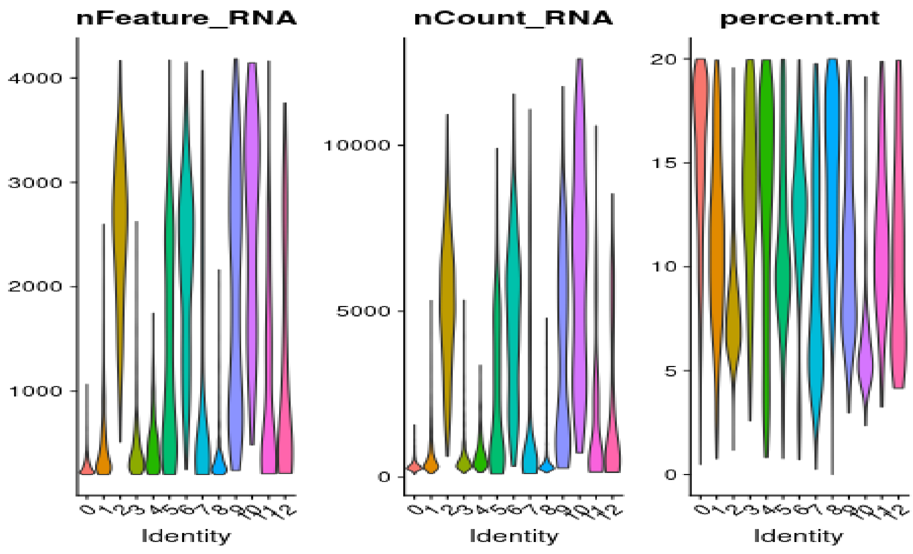

Figure 4: *Quality Control metrics of the UMI count matrix. The violin plot was generated using the filtered dataset on the cells with feature counts of 2,500 or less than 200 and have > 5% of mitochondrial counts. nFeature_RNA represents the number of genes per cell. nCounts_RNA represents the number of*

molecular counts with genes per cell and percent.mt represents the percent of reads that map to the mitochondrial genome.

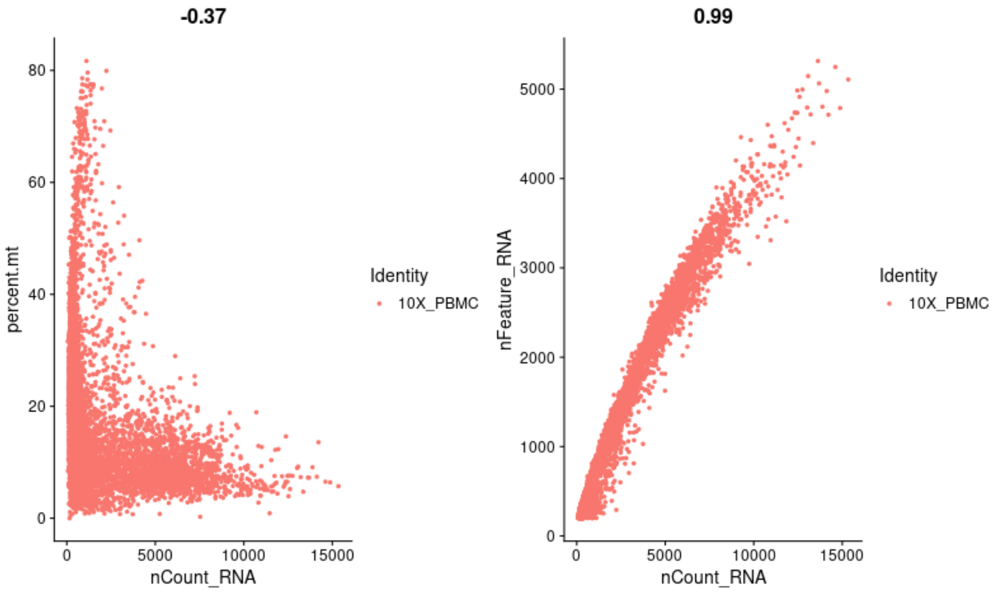

Figure 5: *Scatter plot correlation of nCount_RNA,*
between percent.mt(left) and nFeature_RNA (right).

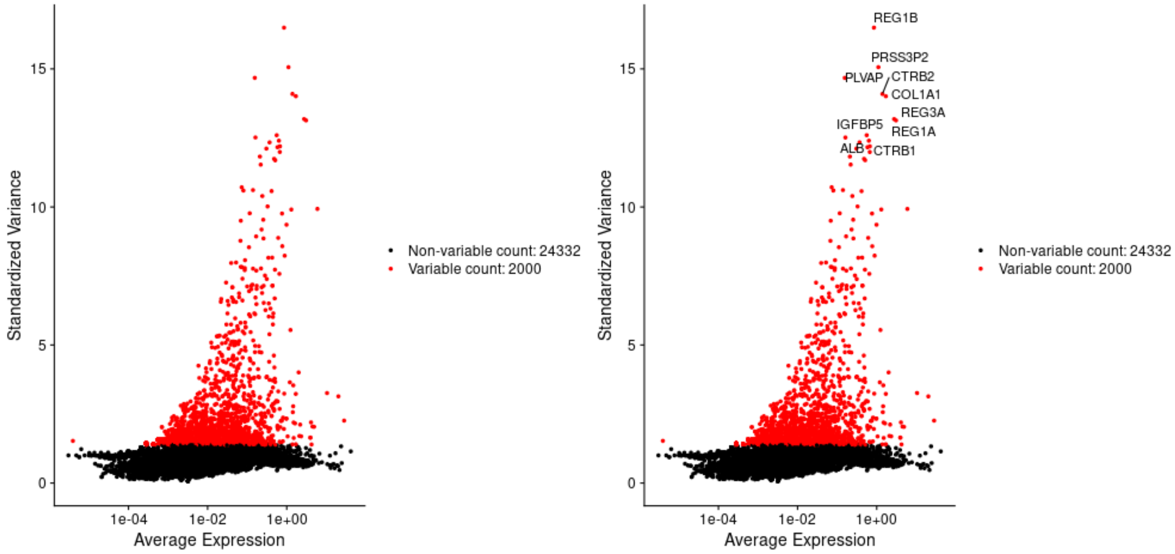

Figure 6*: Average expression & Standardized Variance scatter plot.*
The scatter plot on the ***left*** *represents those without the top 10 most variable genes, and the scatter plot on the* ***right*** *represents the top 10 most variable genes.*

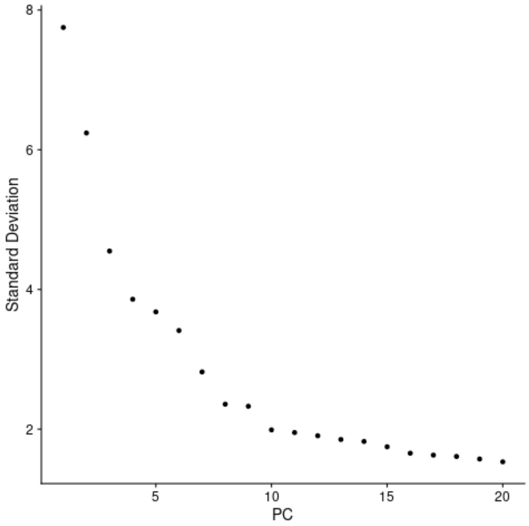

Figure 7: *Elbow plot displays the percentage variance of PCs.*
After a PC of approximately 7, the PCs counts stabilize. However, a fall is noticed in the relationship between standard deviation and PCs. This indicates that around PC 7 or 8, it could be used as a PC cut-off mark.

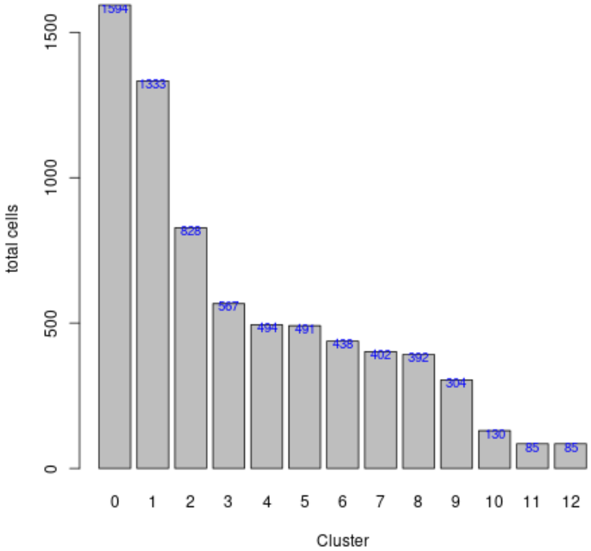

Figure 8: *Histogram of relative proportions of the number of cells in each cluster.*
The x-axis represents the number of clusters from 0 to 11, and the y-axis represents the number of cells.

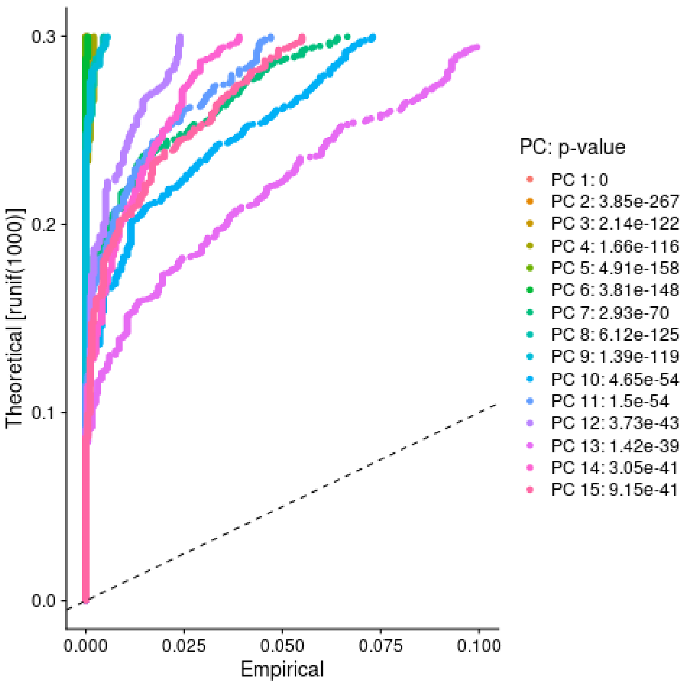

Figure 9: *The jackstraw plot for each PC.*
This plot indicates each principal component with its relative p-value. All the principal components show a deviation from the expected statistic, the dotted line.

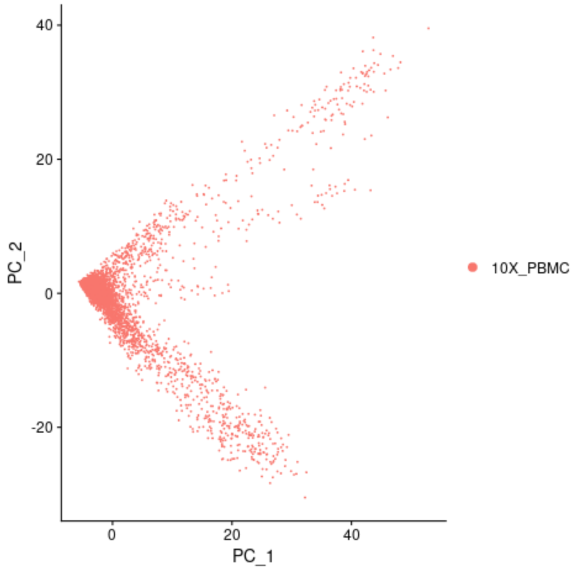

Figure 10: *a PC scatter plot indicating the gene-gene relationship.*
This dimensional PC scatter plot represents the top two components of PCA.

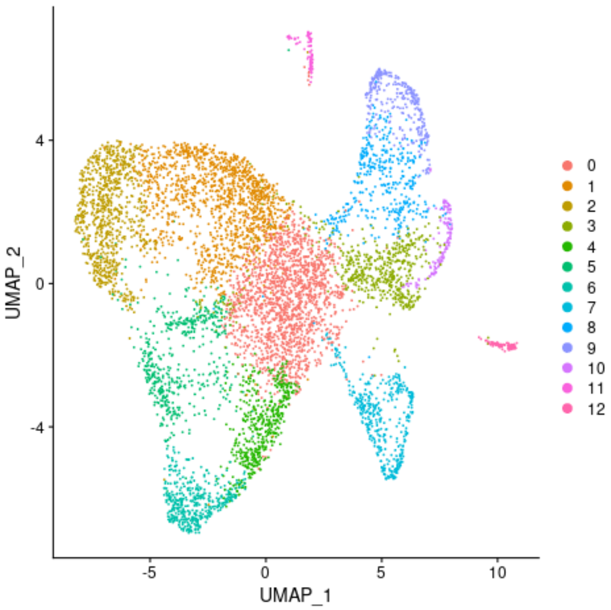

Figure 11: *UMAP for cluster cells.*
The colors represent specific cell types within the cluster map.

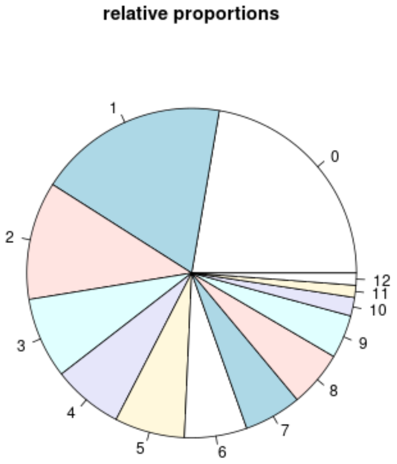

Figure 12: *Pie chart indicating the relative proportions of cell numbers for each identified cluster. Each color in the diagram represents a different cluster.*

Identifying Marker Genes for each cluster

After filtering for low-quality cells and low variance genes, 13 clusters were identified. To have a preliminary understanding of our marker genes, we plot a Violin plot for the different marker genes to visualize their expression in other clusters as shown below.

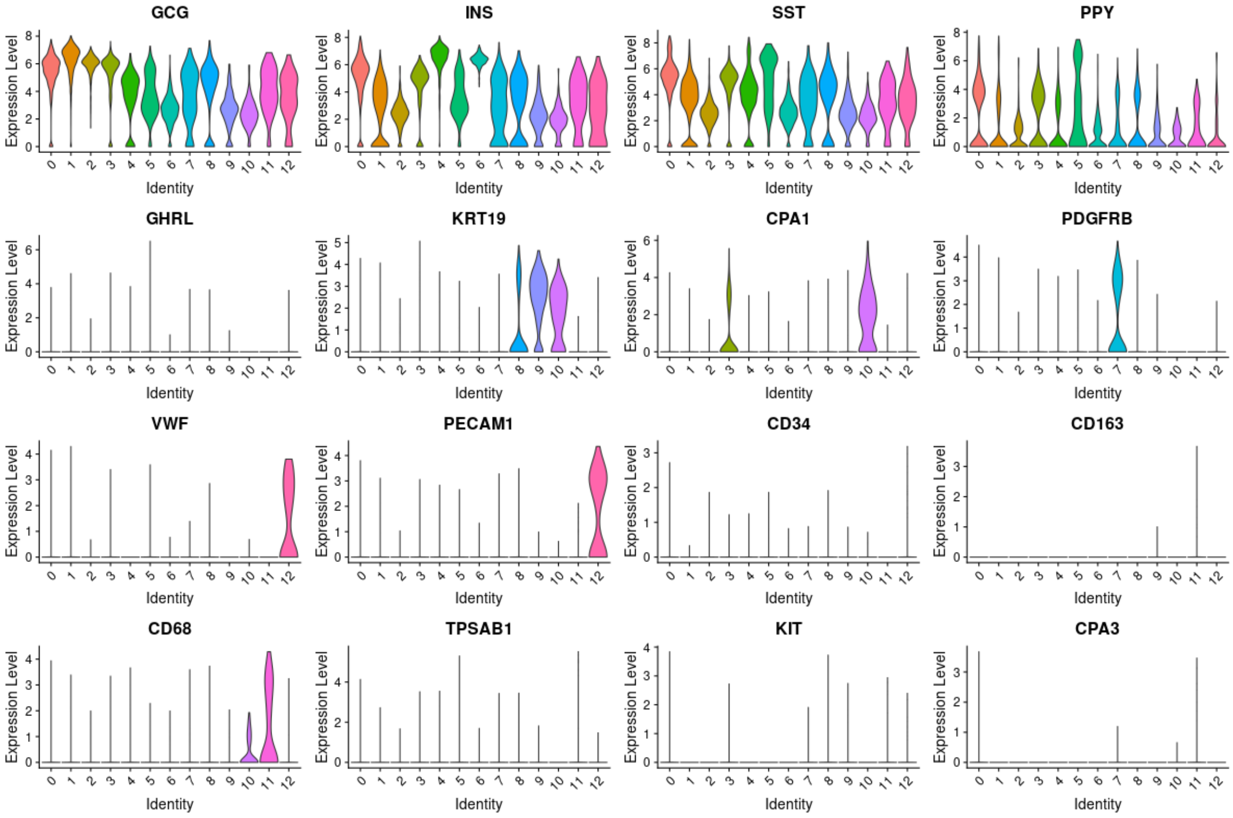

Figure 13: *Violin plot representing the expression of selected marker genes across all 13 clusters.*

We can observe that GCG, INS, SST, and PPY are expressed in almost every cluster. Clusters that represent these markers with the highest percentage and Log2 fold change were classified as the respective cell types based on Table 1 shown below. Based on the supplementary material in Baron et al. [1], IgG, CD3, and CD8 were also recognized as markers but were not found in our analysis. Hence, we don't make use of these markers. VWF and PECAM1 are exclusively expressed in Cluster 12 and can be classified as the Vascular cell type without any ambiguity.

A Feature Plot would be another method of visualization that would help us envision feature expression in low-dimensional space (in the form of a UMAP) across various clusters. As we can observe from the plot, these results are in accordance with the Violin Plot.

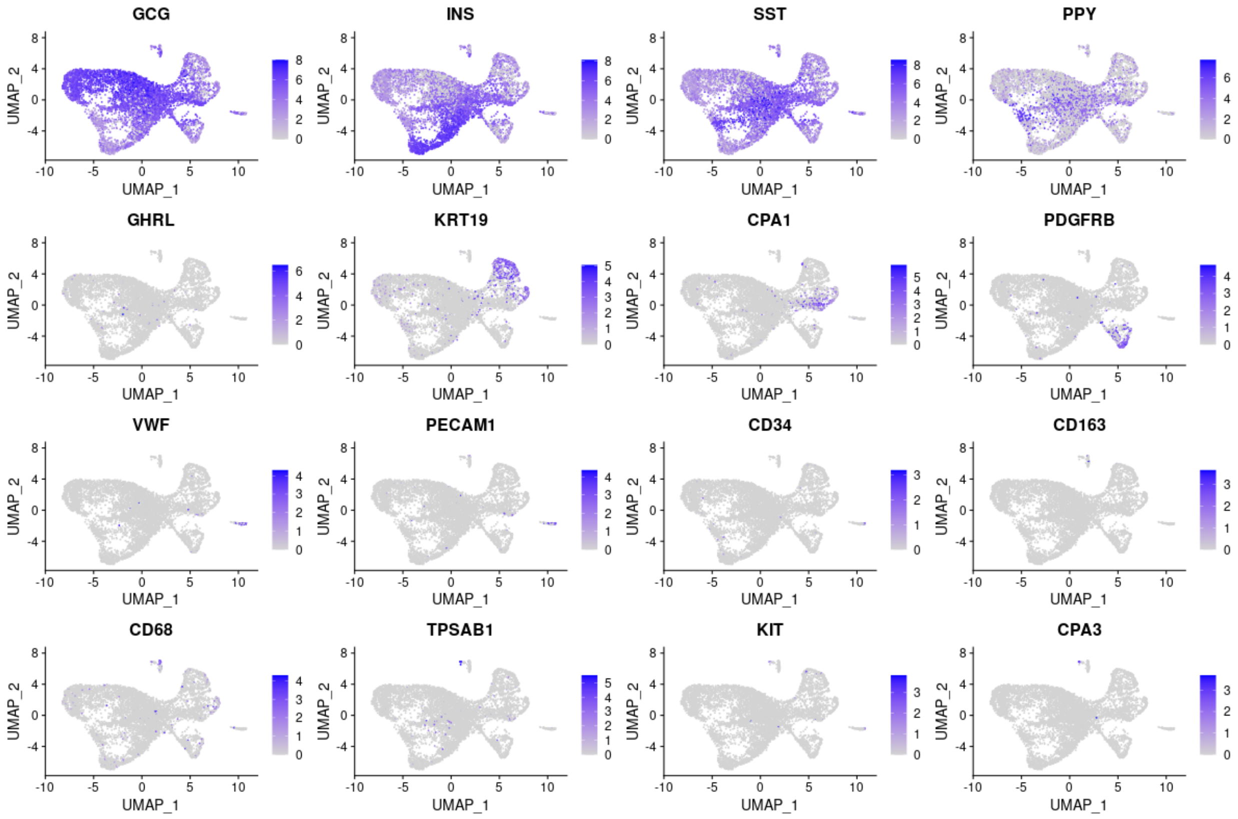

Figure 14: *Feature Plot representing the expression of markers in a UMAP plot.*

Cluster number	Cell type	Marker gene
0	Delta	SST
1,2	Alpha	GCG
3,10	Acinar	CPA1
4,6	Beta	INS
5	Gamma	PPY
7	Stellate	PDGFRB
8,9	Ductal	KRT19
11	Macrophage	CD68
12	Vascular	VWF, PECAM1

Table 3: *Cell types and marker genes for each cluster. Cell types were assigned based on marker gene expression. Marker genes were chosen based on Baron et al. [1] Table S2.*

We can visualize the clusters in low dimensional space using the RunUMAP() function on the first 30 dimensions of the PCA clustered dataset. Labels of the cluster numbers have been replaced with the cell type based on Table 1. Looking at the UMAP components 1 and 2 plotted against each other, we can see that the vascular and macrophage cell types are very distinct and located far from each other and other cell types. Clusters with the same cell types are clustered together, like Alpha, Ductal, Acinar, and Beta Cell types. On observing closely, some of them also merge into each other at the boundary, confirming that these clusters are similar to each other and are probably specific sub-populations of the primary cell type.

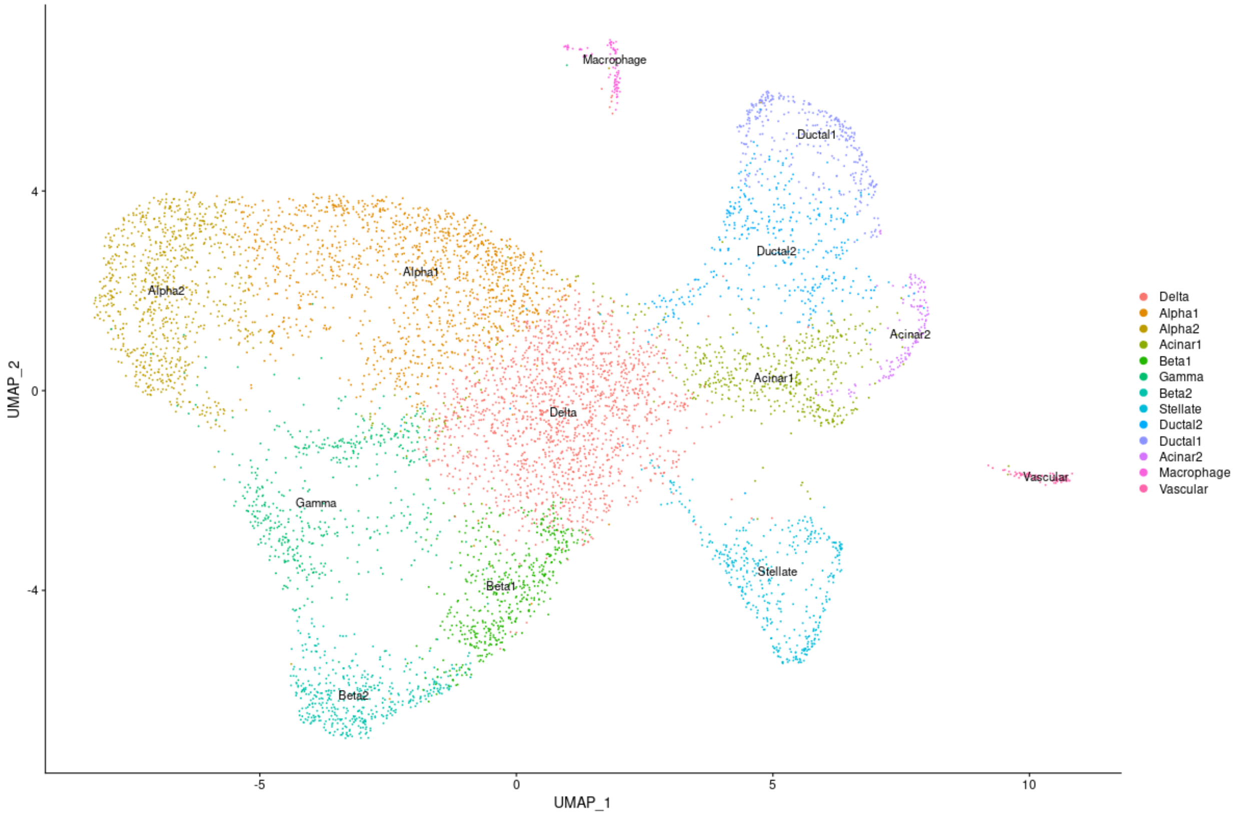

Figure 15: *Dimensionality of cell clusters through a UMAP plot. Clusters are labeled by the particular cell type assigned, as shown in Table 3.*

To visualize the top novel marker genes per cluster, a heatmap (with at most five markers for every cluster) was created to represent the marker genes' expression across the various clusters. For viewing which clusters are closely related to each other or are even the same cell type or a sub-population belonging to a cell type, one can observe if a similar set of markers are expressed in both or many clusters. For instance, clusters 3 and 10 have a similar expression of the marker genes REG1A, REG3A, etc., which tells us that they might be acinar cells.

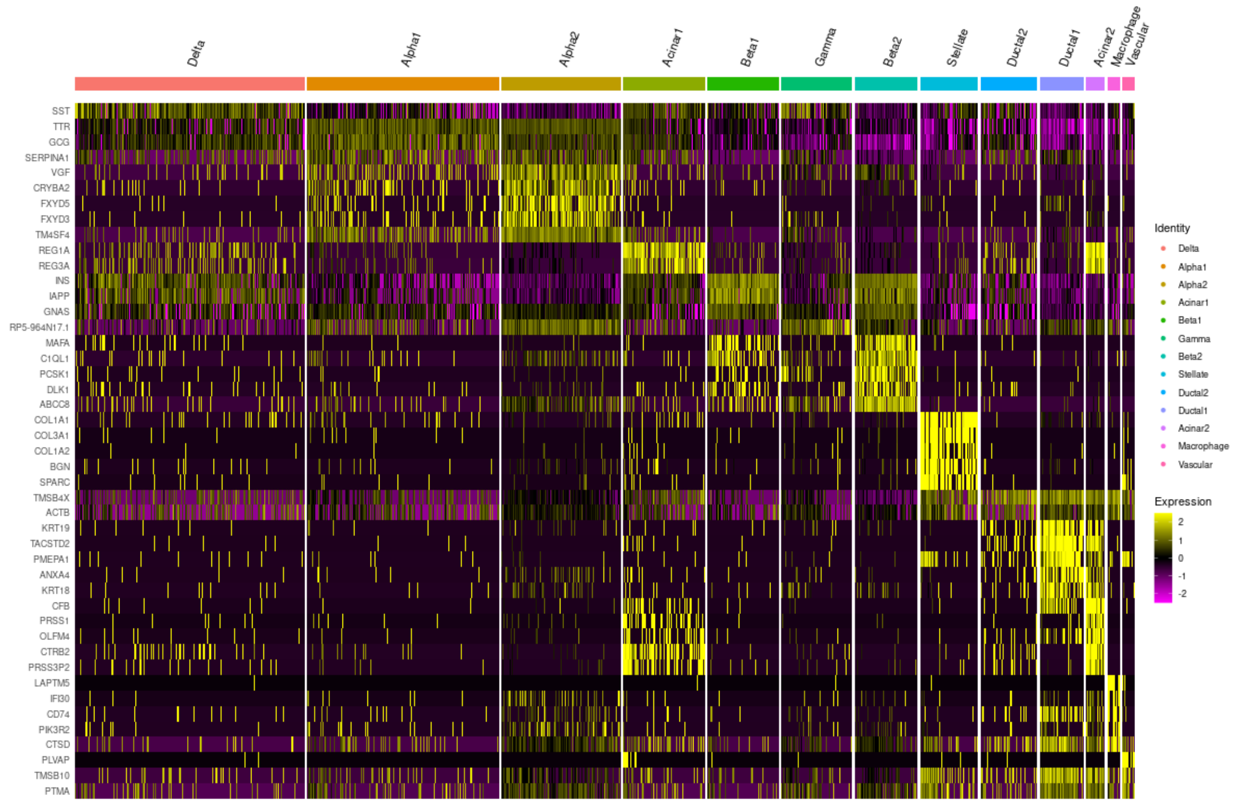

Figure 16: *A heatmap representing the expression of the top novel marker genes in every cluster. Each row shows the expression level of one marker gene, and the columns are grouped by their respective clusters. The yellow color indicates a higher expression level, while purple shows a decrease in expression level.*

Discussion

Single cell RNA-seq provides a higher resolution of cellular differences, allowing researchers to understand cellular diversity better. It can help us to better understand disease mechanisms at the gene and cellular level. We attempted to recreate the single-cell RNA-seq analysis conducted by Baron et al. [1] to verify the cell types. Overall, we could reproduce several but not all components of their results. This could be attributed to the fact that we only chose the three 51-year-old female samples and not all of them. The pre-processing of the data into UMI raw counts matrix and parameters for subsetting and clustering the data might differ from the Baron et al. study [1].

The Baron et al. study originally identified 14 different clusters and attributed them to different cell types. Our study identified 13 clusters, out of which Acinar, Alpha, Beta, and Ductal cell types belong to 2 clusters each, and Gamma, Delta, Stellate, Macrophage, and Vascular cell types belong to 1 cluster each. The same cell type might be labeled into two clusters since more markers are required to subdivide these specific subpopulations of cell types.

Overall, we have identified fewer cell types compared to the original article. However, the cell species generally correspond to the results reported in the original article. The Baron et al. study also identifies Epsilon, Endothelial, Mast, Cytotoxic T cell, and Schwann cell types which weren't identified by our study due to a lack of appropriate marker genes. Additionally, the stellate cell type was subdivided into quiescent and activated cell types in the Baron et al. study. Our study didn't make this distinction. One of the reasons for the misclassification and incomplete result could be our custom identification of barcodes. Since we do not have access to the original barcode, valid barcodes were customized and padded manually. Another possible explanation for the missing cell types is that upstream processing, setting thresholds, and clustering could have filtered out the marker genes associated with these cell types.

Conclusion

Overall, the reproducibility of the findings from Baron et al.[1] the study was not perfectly replicated due to minor inconsistencies. Despite that, the identified cell types are generally consistent with the original article. We believe that a better result reproduction could be gained when using more samples to perform the analysis. Other hardships encountered in this study were due to the overlap of shared genes between different clusters that might have resulted in similar shared genes between different cell types. A more accurate reproduction of the original may be facilitated by adjusting all the parameters in the functions, from filtering to adjusting thresholds, finding different clustering methods, and using more examples from the full dataset. The study provided a novel method to characterize transcriptome profiling using single cell RNA-Seq and the results to study disease pathology.

References

1. Baron, Maayan, Adrian Veres, Samuel L. Wolock, Aubrey L. Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, et al. 2016. "A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-Cell Population Structure." Cell Systems 3 (4): 346–60.e4.
2. Franzén O, Gan LM, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data exploration. Database (Oxford). 2019 Jan 1;2019:baz046. doi: 10.1093/database/baz046. PMID: 30951143; PMCID: PMC6450036.
3. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell.* 2015;161:1187–1201.
4. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods.
5. Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*. 2019;47(D1): D766-D773.