BF528 Project 4: Group Frizzled

Yashrajsinh Jadeja - *Data Curator*
Janvee Patel - *Programmer*
Camilla Belamarich - *Analyst*
Zhuorui Sun - *Biologist*

Single Cell RNA-Seq Analysis of Pancreatic Cells

Introduction

The pancreas is a complex organ that can maintain healthy metabolism when properly functioning. It reveals two different types of parenchymal tissue with a diverse set of cell types such as two types of exocrine cells and five types of endocrine cells.[1] The functions of the mammalian pancreas depend on the complex interactions of these distinct cell types. The dysfunction of the pancreas leads to type I/type II diabetes and other pancreatic diseases. To understand the functions and diseases of pancreas, genome-wide information on each cell type is crucial. Although some previous studies focused on the pancreas and pancreatic cells, the gene expression profiles have primarily been described with bulk mixtures.[2-3] The biological signal can be masked because of the variability in cell type proportions. Thus, the single-cell resolution is essential to study cell type subpopulations and their functions.

In the study by Baron et al. (2016), they implemented a single cell RNA sequencing method based on droplets in a set of pancreatic cells to study the human and mouse pancreas, and to better understand the cellular diversity in the pancreas.[4] From this analysis, Baron and researchers worked on over 12,000 pancreatic cells from four human donors and two mice strains. They showed a detailed look at the gene expression program of the mammalian pancreas. Single cell RNA sequencing provided a higher resolution of cellular differences and better understanding of individual cell functions. They mapped the transcriptomes of pancreatic cells to known cell type identities and pointed to the functional roles of them. Through the gene differential expression analysis, they identified several key different genes related to pancreatic biology.

Our goal for this project is to reanalyze the data from the original study by Baron et al (2016) and to replicate some of the results that are presented in the paper. In this project, we processed the barcode reads of a single cell sequencing dataset and performed further analyses to identify clusters and marker genes for distinct cell type populations, and we performed gene enrichment analysis on the marker genes as well.

Data

The data for this study was obtained from NCBI GEO (GEO Accession : GSE84133). The 3 samples selected for further data processing were from a 51 year old female subject. The BMI of this subject was 21.1 and the diabetic condition status was non-diabetic. Human pancreatic islets were obtained from Prodo or NDRI and recovered in CMRLS at 37°C for 24h-48h hours after receipt. Cells were encapsulated using the inDrop platform, into droplets on ice and lysed in the 4nL microfluidic droplets using a final concentration of 0.4% NP-40.

Single cell lysates were subject to reverse transcription at 50°C without purification of RNA (Klein et al., Cell 2015).[6] Cells were barcoded using the inDrop platform (Klein et al., Cell 2015), which makes use of the CEL-Seq protocol for library construction (Hashimshony et al., Cell Reports 2012).[7] The instrument used for sequencing was Illumina HiSeq 2500.

The unique barcode counts were generated individually for all 3 samples. Furthermore, 3 cumulative distribution plots were generated for each sample to observe the distribution of reads across barcodes.

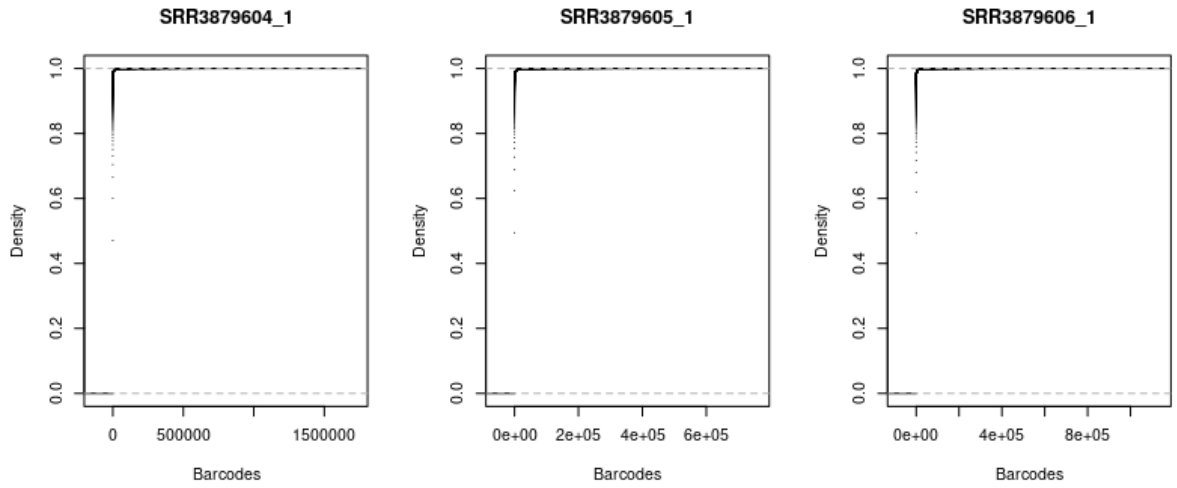

Figure 1. Cumulative Distribution plot for three samples.

This revealed that the barcodes were distributed unevenly with some barcodes having higher counts than many others. The summary statistics for the combined barcodes of 3 samples are detailed in the table below.

Table 1. Summary statistics for barcode distribution.

Min	1st Quartile	Median	Mean	3rd Quartile	Max
1.0	1.0	2.0	396.8	5.0	1604310

This table revealed that the mean count for the barcodes was 396.8. Based on this metric, the non-informative barcodes were filtered out with reads less than the mean count statistic. This ensured that reads with a significant number of counts were retained that would be further used for generating a barcode whitelist.

Salmon Alevin was used to generate the UMI matrix for the data.[8] The reference transcriptome used for indexing was obtained from Gencode (Release 37, GRCh38.p13) and the transcript to map gene file was generated by using data from the 'comprehensive gene annotation' GTF file. This is necessary as Alevin works on transcript level equivalence

classes to resolve potential UMI collision and it also benefits from transcript to gene relation by sharing the information among the equivalence classes. This would enable Salmon to collapse from transcript to the gene level. The mapping statistics generated after generation of UMI matrix are detailed in the table below.

Table 2. Mapping statistics after a successful Alevin run.

Statistic	Value
Transcripts Found	233,613
Transcript to Gene entries	233,652
UMI after deduplicating	24,001,103
Mapping Rate	41.8653%
Reads thrown away due to noisy cellular barcodes	15.4043%

The mapping rate is consistent with what was observed in the original paper.

Methods

Quality Control of the UMI Counts Matrix

The UMI counts matrix was imported using the tximport and the fishpond R packages.[9-10] Within the UMI counts matrix, the Ensembl Gene Identifiers for all the genes were converted to Gene Symbols using an Ensembl Mart resource for the Human Genome (GRCh38.p13) which contained all genes with their associated Ensembl identifier, gene name, and HGNC symbol.[11] The steps for quality control of the UMI counts matrix were performed similarly to Seurat's provided tutorial.[12] Utilizing the Seurat v. 4.0.1 package, a Seurat object was created using the preliminary filters of minimum nonzero count of cells per gene set to 3 and minimum nonzero count of genes per cell set to 50.[13] In determining these filters for complexity of cells (genes per cell) and rarity of genes (cells per genes), an approach was taken similar to an MIT scRNA-Seq Seurat tutorial.[14] The values for genes per cell were determined for each cell by taking a sum of the number of genes with nonzero values in the counts matrix. The values for cells per gene were determined for each gene by taking a sum of the number of cells where the gene was expressed.[14] All cells were ranked by their complexity (genes per cell), and visualized as a scatter plot in Figure 2. The lower inflection point which appeared at approximately 50 was used as the minimum genes per cell filter. The Seurat default filter of 3 for minimum cells per gene and the minimum genes per cell filter mentioned above were used in order to keep a larger portion of cells, and employ more stringent filtering during the filtering of low-quality cells.

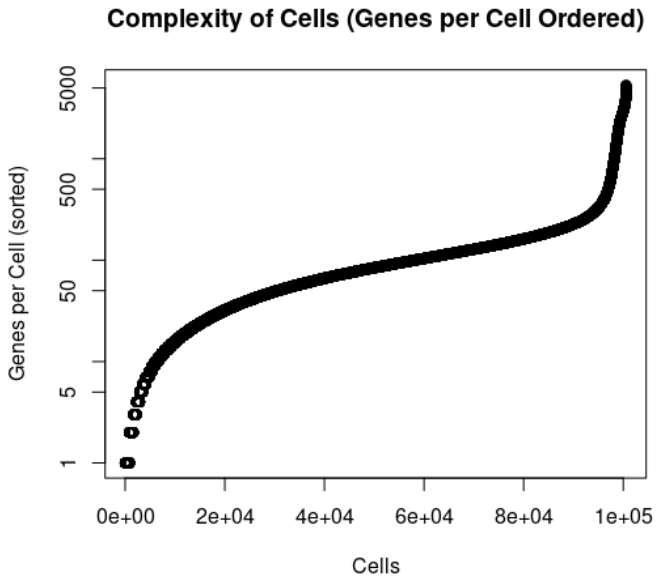

Figure 2. Plot of the Complexity of the Cell (genes/cell) for each Cell in Sorted Format. The x-axis represents all the cells in the unfiltered dataset which is 100,570 cells. All the cells were ordered by their complexity. The y-axis represents the complexity of the cell.

Mitochondrial quality control metrics were determined through Seurat's PercentageFeatureSet function.[12-13] Quality control metrics were visualized using violin plots and feature scatter plots for the number of genes detected in each cell, the number of molecules detected within a cell, and the percentage of reads mapping to mitochondrial genome as shown in Figures 3 and 4. Using these visualizations, low quality cells were filtered out with the metrics of number of genes detected in each cell (nFeature_RNA) between 200 and 4,000, and mitochondrial percentage (percent.mt) less than 48. A higher mitochondrial percentage value was used to account for the observation that mitochondria are present in β-cells and function in glucose-stimulated insulin secretion.[15] Cells with mitochondrial percentages above 48 were filtered out as these could possibly be due to mitochondrial contamination from low-quality or dying cells.[12] Cells with genes per cell greater than 4,000 were filtered out as these could be doublets or multiplets.[12]

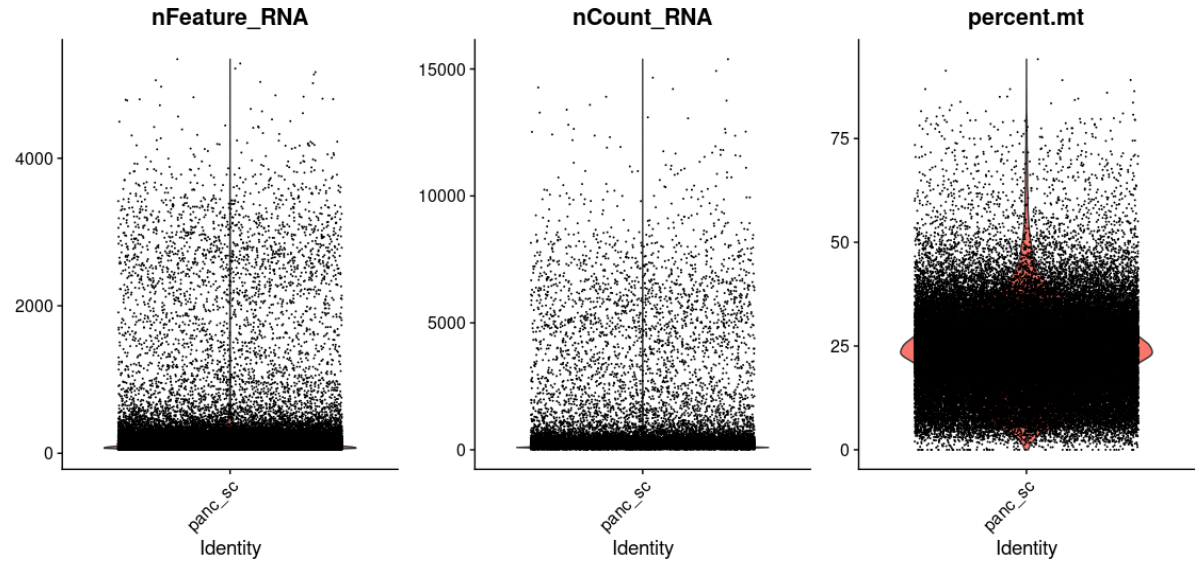

Figure 3. Violin Plots of Number of Genes Detected in Each Cell (nFeature_RNA), Number of Molecules Detected Within a Cell (nCount_RNA), and Mitochondrial Percentage (percent.mt)

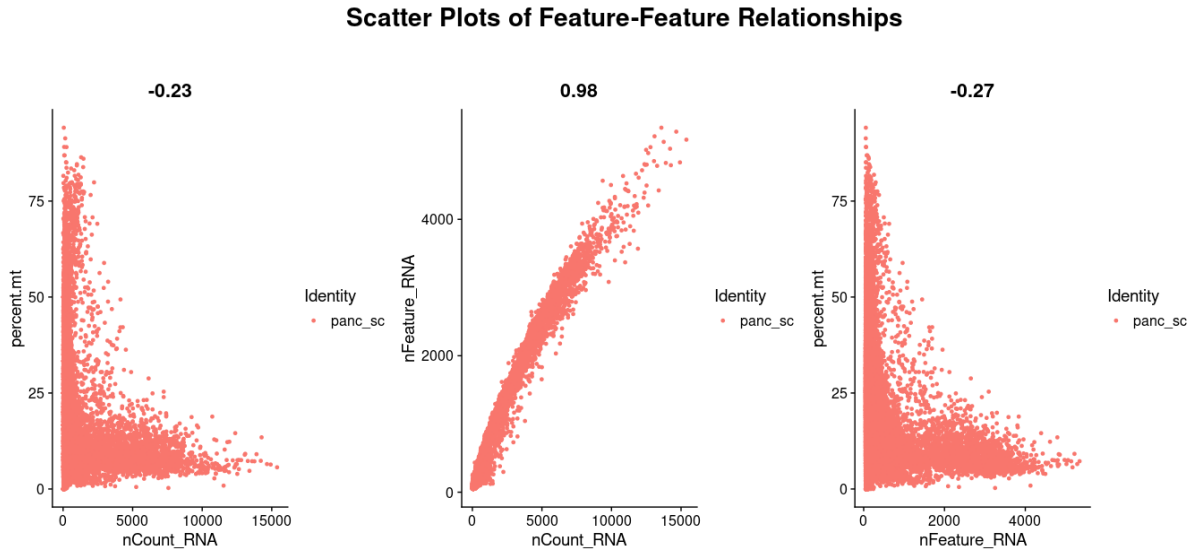

Figure 4. Feature Scatter Plots of Number of Genes Detected in Each Cell (nFeature_RNA), Number of Molecules Detected Within a Cell (nCount_RNA), and Mitochondrial Percentage (percent.mt)

The UMI counts matrix was normalized using log normalization and a scale factor of 10,000 in Seurat's NormalizeData function.[12-13] For feature selection, variance filtering was performed with Seurat's FindVariableFeatures function to filter out low variance genes to get features with high cell-to-cell variation, and the default value of 2,000 high variance features was used.[12-13] Scaling was performed using Seurat's ScaleData function by scaling the expression of each gene so that mean expression across cells was 0 and variance across cells was 1.[12-13] Linear dimensional reduction was performed through principal component analysis (PCA) using Seurat's RunPCA function.[12-13] An

Elbow plot shown in Figure 5 was used to determine the number of principal components to use. According to the approximate bend in the Elbow plot, 10 principal components were used.

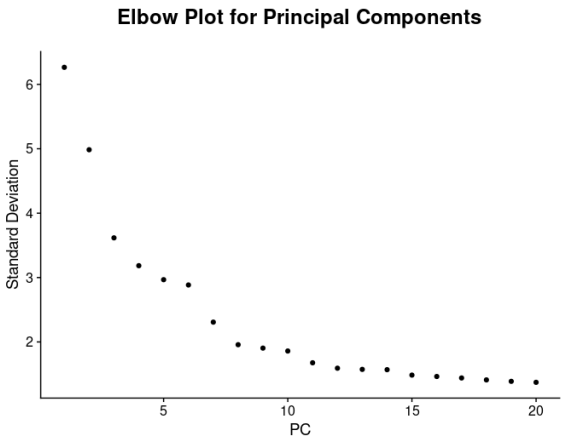

Figure 5. Elbow Plot of the Standard Deviations of the Principal Components. This plot was used to determine the number of principal components to use.

Clustering was performed using Seurat's FindNeighbors function which employed K-nearest neighbors algorithm and shared nearest neighbor (SNN).[12-13] Then, non-linear dimensional reduction was performed using Seurat's RunUMAP function.[12-13] A pie chart displaying the proportions of the cell numbers in each identified cluster shown in Figure 6 was generated using the Graphics R package.[16]

Identifying Marker Genes with Differential Expression

The Rds file containing the clustered data was read into R using the readRDS function from the Seurat package.[13] To identify the marker genes, the FindAllMarkers function from the Seurat package was implemented using a few parameters to obtain clusters with the sufficient amount of marker genes.[13] This function finds markers that define clusters using differential expression.[12] The first parameter was used to extract only the positively enriched marker genes. This was a necessary parameter to use for the subsequent enrichment analysis. Two thresholds were also applied to retrieve a sufficient amount of marker genes for each cluster. A minimum percent of 0.25 and log2 fold change threshold of 0.5 was applied to achieve this. The minimum percent parameter requires that the feature of interest must be detected at a minimum percentage between two groups.[8] Multiple threshold values of these parameters were tested; however, the results obtained did not represent the clusters well enough for the cell type to be identified by the maker genes found for each cluster. In this case, a minimum percent threshold too large, did not produce enough marker genes in that cluster to be identified for cell type. Conversely, a minimum percent of too small produced too many marker genes. This is the same reasoning for increasing the log2 fold change threshold. Additionally, a significance level of 0.05 was applied to the adjusted p-value for the marker genes.

Labelling Clusters with Cell Type

All significant marker genes were assigned to corresponding clusters. These marker genes were used to classify each cluster with a specific cell type. Baron et al (2016) matched the marker genes they found with cell types in their supplemental table S2. This was used as a reference for labelling clusters with the correct subtype. A series of steps was implemented to stay consistent when labelling the clusters with a certain cell type. The first step was to extract the top five marker genes from each cluster by average log2 fold change. Secondly, the reference marker genes were matched to the top two marker genes from each cluster. If the genes matched, the cluster was labelled with the appropriate cell type. However, if the genes did not match, all the differentially expressed marker genes were searched through using the reference marker genes to match from a bigger subset of data. If there was only one match, the cluster was labelled with the corresponding cell type. If there were multiple clusters that had the same reference marker gene, a comparison between the clusters and a literature search was performed. A comparison was made between clusters using Seurat's FindMarkers function, which output a list of genes that differed between the two clusters.[13] These genes were then used in a literature search to determine which cell type matches the differentially expressed genes the best. If multiple clusters were identified with the same cell type, the cluster was labelled accordingly. Lastly, if a cell type could not be identified, the cluster was labeled "Unknown".

Violin plots and a UMAP were plotted to visualize and more accurately label the clusters with the correct cell type. For Unknown clusters, the top two marker genes were plotted across all clusters using Seurat's VinPlot function.[13] This showed the expression of the top marker genes across all clusters. Additionally, Seurat's RunUMAP and DimPlot function displayed the clusters.[9] These plots were especially useful when deciding which cell type to label clusters with multiple or no reference marker genes.

Finding Novel Marker Genes

After labelling clusters with appropriate cell types based on marker genes, there were many instances of marker genes that were just as discriminative of cell type as the marker gene chosen. These novel marker genes were important to report. In order to identify the novel genes, certain thresholds were made stricter compared to the thresholds created to extract the significant marker genes. The minimum percent was increased from 0.25 to 0.8. This increased the minimum percentage of this marker gene to be identified between two groups. The fold change threshold was increased from 0.5 to 1. Lastly, the adjusted p-value was decreased from 0.05 significance level to 0.005 significance level.

Gene Enrichment Analysis

We performed gene enrichment analysis for the marked genes in each clustered cell based on the platform ENRICHR.[17] To analyze the functional roles, the boxes we used here BP indicated for biological processes, MF indicated for molecular function and CC indicated for cellular components. And we can speculate which tissue the cell comes from based on Jensen TISSUES from ENRICHR.

Results

Tables 3 and 4 show the number of cells and genes in the unfiltered and preliminary filtered datasets respectively. After applying the filters of minimum genes per cell < 50 and minimum cells per gene < 3, there were 30,286 cells and 31,565 genes filtered out. Table 5 shows the number of cells remaining after filtering of low quality cells using the filters of genes per cell > 200 and genes per cell < 4,000, and mitochondrial percentage < 48. There were 56,997 cells filtered out to result in a total of 13,287 cells remaining. Table 6 shows the number of genes used after filtering of low-variance genes, and Seurat's default value of 2,000 was implemented for the downstream analysis.[12-13]

Table 3. Summary of the number of cells and genes in the unfiltered dataset. This table shows the values before any filter has been applied.

Number of Cells	Number of Genes
100,570	**60,232**

Table 4. Summary of the number of cells and genes in the preliminary filtered dataset. This table shows the values after applying the filters of minimum features < 50 and minimum cells < 3 when the Seurat object was generated.

Number of Cells	Number of Genes
70,284	**28,667**

Table 5. Summary of the number of cells after filtering of low quality cells. This table shows the value after applying the filters of 200 < genes per cell < 4,000 and mitochondrial percentage < 48.

Number of Cells
13,287

Table 6. Summary of the number of features used after variance filtering. This table shows the number of genes used after filtering out low-variance genes. Seurat's default value was used.

Number of Features
2,000

Table 7. Summary of the number of clusters identified.

Number of Identified Clusters
14

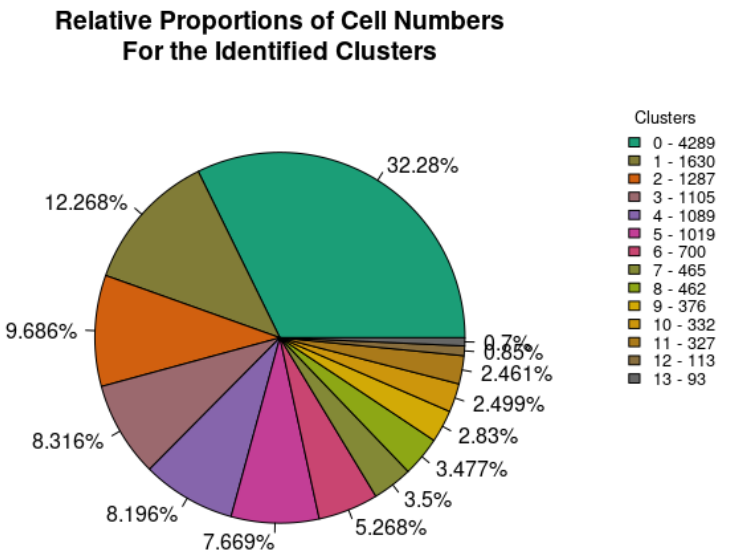

Figure 6. Pie chart displaying the relative proportions of cell numbers for each identified cluster. Percentages are shown for each cluster. There were 14 identified clusters. The legend includes the cluster number followed by the cell numbers for that respective cluster.

Table 7 and Figure 6 show the number of identified clusters and the relative proportions of cell numbers for each of the clusters respectively. There were 14 identified clusters. Based on Figure 6, cluster 0 had the highest number of cells with a value of 4,289 cells which made up approximately a third of the total number of cells used for downstream analysis. Cluster 13 had the lowest number of cells with a value of 93 cells which made up approximately 0.7% of the total number of cells. A trend that can be observed from Figure 6 is that each consecutive cluster has a fewer number of cells compared to the previous cluster. Clusters that have a few number of cells could possibly be attributed to rare cell types within pancreatic tissue.

Labelling Clusters with Cell Type

Table 8 shows how Baron et al (2016) labelled their clusters using marker genes found in their data. Authors used information from various studies to match marker genes with cell types in the pancreas. In this analysis, Table 1 was used as a reference to match the marker genes found to cell type. The top two marker genes from each cluster were first matched to the reference marker genes. *GCG*, *INS*, *SST*, and *KRT19* were matched to the corresponding cell type from Table 8. Next, all the differentially expressed genes were used to match reference marker genes to cell type. *PPY*, *CPA1*, *PDGFRB*, *VWF*, *PECAM1*, and *CD68* were also matched to the corresponding cell type from Table X. *PPY* was matched to both Delta and Gamma cell types; however, *SST* was more significantly differentially expressed in the Delta cell type cluster, so *SST* was a better representation of Delta than *PPY*. Marker genes from Epsilon, Cytotoxic T, and Mast cell types were not found.

Table 8. Baron et al (2016) Reference Marker Genes and Corresponding Cell Type. Supplemental Table S2 from Baron et al (2016) used to label clusters with cell type based on gene markers.

Cell Type	Genes
Alpha	GCG
Beta	INS
Delta	SST
Gamma	PPY
Epsilon	GHRL
Ductal	KRT19
Acinar	CPA1
Stellate	PDGFRB
Vascular	VWF, PECAM1, CD34
Macrophage	CD163, CD68, IgG
Cytotoxic T	CD3, CD8
Mast	TPSAB1, KIT, CPA3

After this matching process, there were still five clusters with unidentified cell types. The first cluster, now labelled Alpha_2, needed further classification. Not only was *GCG* also found significantly differentially expressed in this cluster, *CRYBA2* was one of the top two genes found in this cluster. From a literature search, Muraro et al (2016) identified *CRYBA2* as an alpha cell in the pancreas. The second cluster, now labelled Beta_2, also needed further classification. Similar to Alpha_2, *INS* was also found differentially expressed in this cluster, and one of the top two marker genes, *MAFA*, was consistent with what Muraro et al (2016) found in their paper. *MAFA* was identified to be a beta cell type. Therefore, we labelled this cluster Beta_2. The third cluster, now labelled Delta_2, also contained *SST* differentially expressed. *RBP4*, one of the top two differentially expressed marker genes in this cluster, was also classified as a Delta cell type in the Muraro et al (2016) study. *AQP3*, the other top two marker gene in this cluster, was identified as PP cell type in the Muraro et al (2016) study. PP cell was found to be the new name for the Gamma cell type.[18] Ultimately, this cluster was determined to be Delta cell type for the instance of *SST* and *RBP4*. These three clusters were labelled as a different version of the original cell type to differentiate the cell types when plotted.

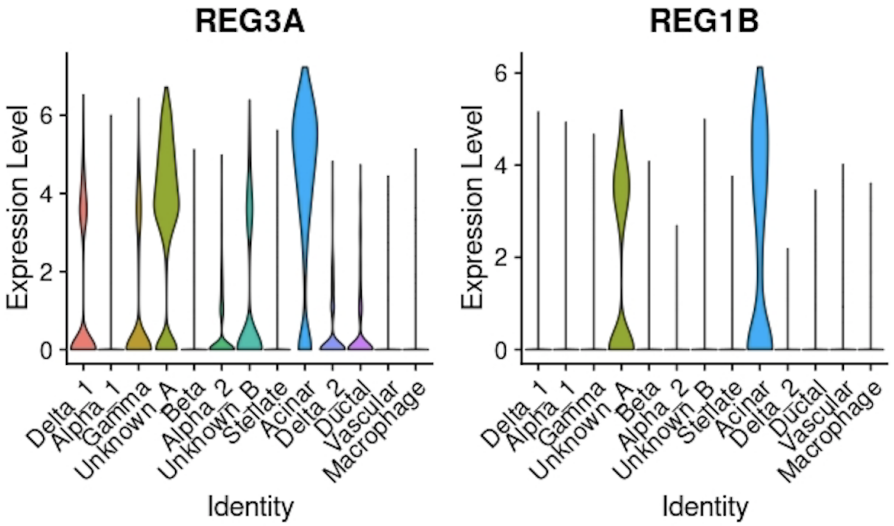

Figure 7. Unknown Cluster A Violin Plot. Expression levels of top maker genes *REG3A* and *REG1B* across all cell types. Both genes show high expression in Acinar cell type. High expression levels are indicated by taller/wider violin plots and low expression is indicated by shorted/narrow violin plots.

The two other clusters were labelled Unknown_A and Unknown_B. Even further classification was needed to make these assumptions. The top two differentially expressed marker genes for Unknown_A were *REG3A* and *REG1B*. A literature search for each gene was performed, but no concrete results were found to classify this cluster with a cell type. Other than being expressed in cluster Unknown_A, both genes were highly expressed in Acinar cell type (Fig. 7). Unknown_A was also clustered relatively close to the Acinar cell type (Fig. 7). The top two differentially expressed marker genes for the Unknown_B cluster were *TACSTD2* and *KRT8*. A literature search for each gene was conducted, but no definitive cell types were identified. Both genes were moderately expressed in Acinar and Ductal cell types. *KRT8* was also highly expressed in Alpha_2 and Delta_2 (Fig. 8). Additionally, Unknown_B was clustered close to Acinar and Ductal in Figure 8.

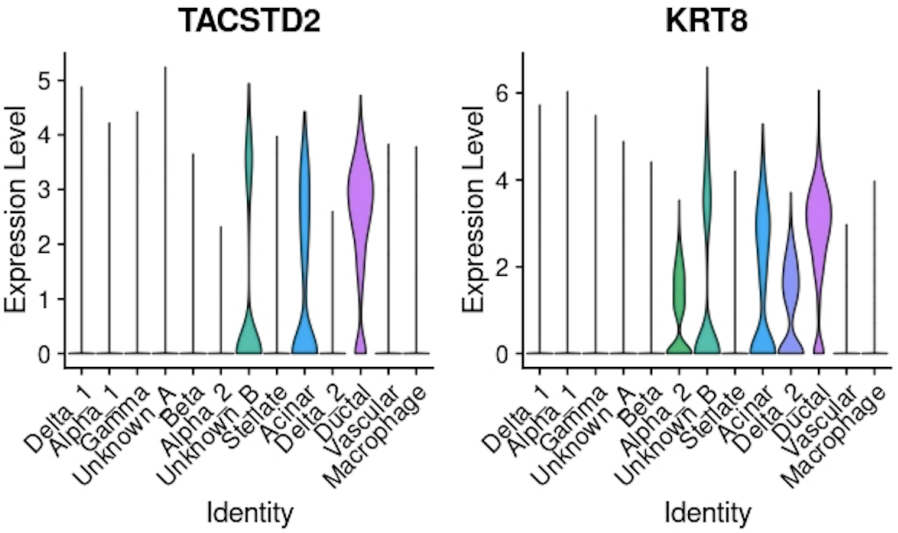

Figure 8. Unknown Cluster B Violin Plot. Expression levels of top maker genes *TACSTD2* and *KRT8* across all cell types. Both genes show high expression in Acinar and Ductal. *KRT8* also shows high expression in Alpha_2 and Delta_2. High expression levels are indicated by taller/wider violin plots and low expression is indicated by shorted/narrow violin plots.

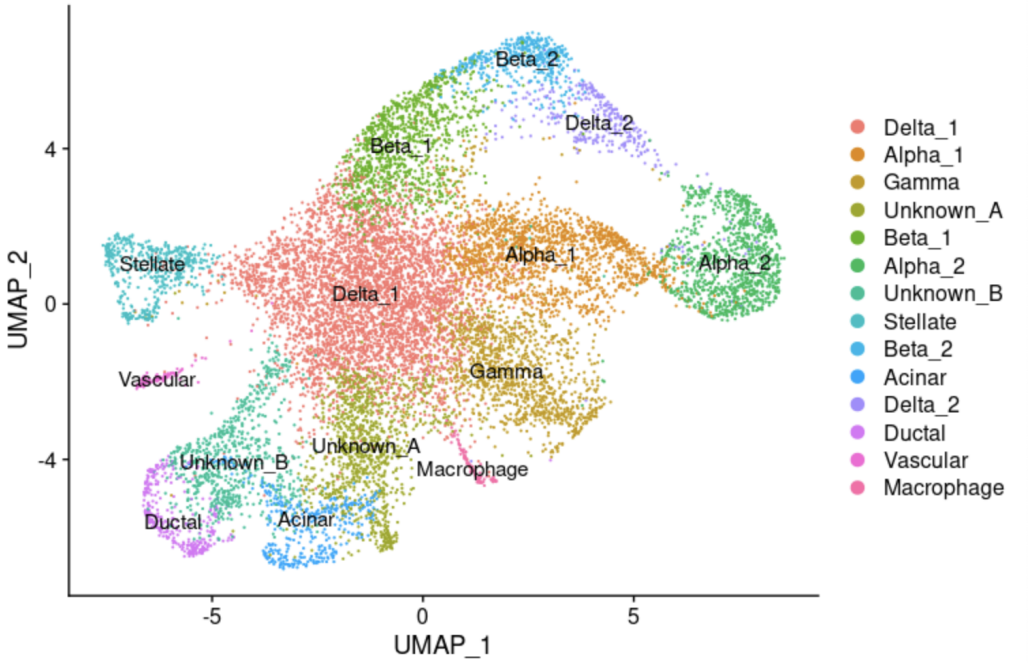

Figure 9. UMAP of Gene Markers Colored by Cell Type. Colors represent each cell type cluster. Less definitive labelled cell types are clustered closer together than more definitive labelled call types.

The cell types were clustered and displayed on the heatmap (Fig. 9). Cell types that had two separate clusters were less definitive than the cell types with just one cluster. However, both Acinar and Ductal were observed to have more than one highly expressed cluster in the heatmap(Fig. 10). Additionally, the Unknown clusters are also less definitive and a potential second or third cluster can be observed for that section of the plot (Fig. 10).

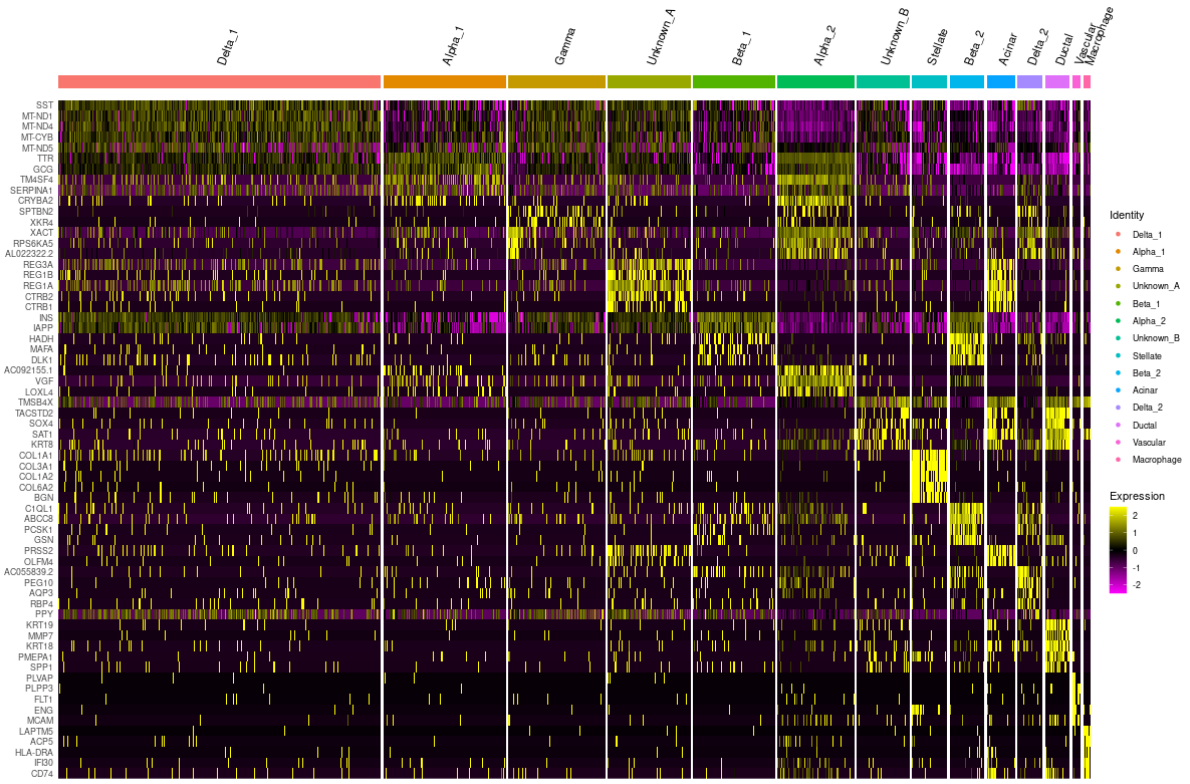

Figure 10. Top Five Marker Genes per Cell Type. Clustered heatmap of expression levels across all cell types showing the top five marker genes per cell type. Cell type is labelled by color. High expression is indicated by yellow clusters and low expression is indicated by purple.

Identifying Novel Marker Genes

After stricter requirements were applied to the novel marker genes, 87 total novel marker genes were found. The top five novel genes for each cell type are shown in Table 9. Gamma, Unknown_A, and Vascular cell types did not have novel genes. Delta_1, Alpha_1, Beta_1, Unknown_B, Stellate, and Dekta_2 cell types had less than five novel genes. Lastly, Alpha_2, Beta_2, Acinar, Ductal, and Macrophage cell types had more than five novel genes.

Table 9. Top Novel Genes Associated with Cell Type. Top novel genes found in each cell type and reported statistics. (*) represent reference marker genes found to be novel genes. (**) represent maker genes found in comparison analysis or literature search.

Marker Gene	P-Value	Average Log2FC	pct.1	pct.2	Adjusted P-Value	Cell Type
SST*	0	1.0023	0.98	0.89	0	**Delta_1**
TTR	0	1.3152	0.988	0.876	0	**Alpha_1**
GCG*	2.7778E-303	1.3420	0.988	0.947	7.9632E-299	**Alpha_1**
INS*	1.5321E-274	1.7142	0.986	0.888	4.3922E-270	**Beta_1**
IAPP	4.9444E-181	1.7216	0.939	0.802	1.4174E-176	**Beta_1**
CRYBA2**	0	1.6340	0.856	0.108	0	**Alpha_2**

VGF	0	1.5512	0.931	0.208	0	**Alpha_2**
TM4SF4	0	1.5134	0.981	0.27	0	**Alpha_2**
SLC7A2	0	1.4096	0.846	0.114	0	**Alpha_2**
CLU	0	1.4040	0.991	0.443	0	**Alpha_2**
TMSB4X	5.3508E-149	1.8638	0.839	0.546	1.5339E-144	**Unknown_B**
COL1A1	0	4.7457	0.867	0.143	0	**Stellate**
TIMP1	7.1433E-157	2.8260	0.826	0.388	2.0478E-152	**Stellate**
MAFA**	0	1.9413	0.825	0.069	0	**Beta_2**
C1QL1	0	1.9087	0.92	0.129	0	**Beta_2**
ABCC8	0	1.8744	0.946	0.198	0	**Beta_2**
HADH	0	1.6262	0.887	0.104	0	**Beta_2**
GAD2	0	1.4076	0.807	0.122	0	**Beta_2**
CTRB2	0	4.1360	0.814	0.119	0	**Acinar**
SPINK1	0	3.5638	0.83	0.138	0	**Acinar**
AL049839.2	0	3.5222	0.875	0.145	0	**Acinar**
SERPINA3	0	3.1904	0.872	0.144	0	**Acinar**
REG1A	8.1464E-205	3.8029	0.87	0.303	2.3353E-200	**Acinar**
IDS	8.7330E-126	1.2231	0.822	0.234	2.5035E-121	**Delta_2**
PCSK1N	7.8332E-85	1.1754	0.961	0.491	2.2455E-80	**Delta_2**
MALAT11	2.1216E-52	1.0763	0.982	0.723	6.0819E-48	**Delta_2**
KRT19*	0	2.9318	0.844	0.069	0	**Ductal**
KRT18	0	2.7623	0.804	0.083	0	**Ductal**
PMEPA1	0	2.7033	0.826	0.071	0	**Ductal**
TACSTD2**	0	2.6233	0.859	0.065	0	**Ductal**
CD24	2.3578E-281	2.1130	0.81	0.12	6.7592E-277	**Ductal**
CTSD1	2.3691E-56	3.6497	0.871	0.289	6.7916E-52	**Macrophage**
TMSB4X3	1.8387E-42	2.1481	0.978	0.559	5.27010E-38	**Macrophage**
FTL	9.6684E-32	3.4188	0.871	0.481	2.7716E-27	**Macrophage**
FTH1	1.4212E-24	2.7721	0.871	0.495	4.0742E-20	**Macrophage**
ACTB	1.1650E-16	1.1891	0.882	0.631	3.3399E-12	**Macrophage**

Gene Enrichment Analysis

The summary of the gene enrichment analysis for clustered cells shown in Table 10. The gene enrichment analysis was performed based on ENRICHR. For each cluster, top 5 marked genes were selected based on p-value. The top GO terms of biological process (BP), molecular function (MF) and cellular component (CC) were selected to show in the table.

Table 10. Summary of the gene enrichment analysis for clustered cells. The summary information of the gene enrichment analysis contained Cluster, top five enriched genes, process, the top GO term for each process and the top tissues for each cluster.

Cluster	Top 5 Enriched Genes	Process	Top GO terms	Top tissues
0	SST MT-ND1 MT-ND4 MT-CYB MT-ND2	BP	mitochondrial ATP synthesis coupled electron transport (GO: 0042775)	Pancreatic islet Cerebrospinal fluid AtT-20 cell
		MF	NADH dehydrogenase (ubiquinone) activity (GO:0008137)	
		CC	mitochondrial respiratory chain complex I (GO:0005747)	
1	TTR GCG TM4SF4 CFC1B CFC1	BP	nodal signaling pathway (GO:0038092)	Pancreatic juice Cestode Kinetoplastid
		MF	activin receptor binding (GO:0070697)	
		CC	secretory granule lumen (GO:0034774)	
2	SPTBN2 XKR4 TYRO3 ZNF320 OR7C1	BP	positive regulation of histone phosphorylation (GO:0033129)	Intestinal mucosa A-172 cell Gastrointestinal endocrine cell
		MF	chondroitin sulfotransferase activity (GO:0034481)	
		CC	integral component of Golgi membrane (GO:0030173)	
3	REG3A REG1B REG1A CTRB2 PRSS2	BP	antimicrobial humoral immune response mediated by antimicrobial peptide (GO:0061844)	Pancreatic juice Bladder Histiocyte
		MF	peptidoglycan binding (GO:0042834)	
		CC	azurophilic granule lumen (GO:0035578)	
4	INS IAPP GNAS HADH MAFA	BP	cellular protein metabolic process (GO:0044267)	Ear Pancreatic beta cell line Pancreatic islet
		MF	insulin-like growth factor receptor binding (GO:0005159) adrenergic receptor binding (GO:0031690)	
		CC	cytosolic ribosome (GO:0022626)	
5	AC092155.1 CRYBA2 VGF TM4SF4 LOXL4	BP	cellular protein metabolic process (GO:0044267)	Eye Pancreatic islet Intestine
		MF	inositol trisphosphate kinase activity (GO:0051766)	
		CC	cytosolic ribosome (GO:0022626)	
6	TMSB4X TACSTD2 SERPINA3 HSPB1 AL049839.2	BP	platelet aggregation (GO:0070527)	Bladder Ascites Vascular system
		MF	cadherin binding involved in cell-cell adhesion (GO:0098641)	
		CC	secretory granule lumen (GO:0034774)	

7	COL1A1 COL3A1 COL1A2 COL6A2 BGN	BP	extracellular matrix organization (GO:0030198)	Bone Intestine Vascular system
		MF	RNA binding (GO:0003723)	
		CC	focal adhesion (GO:0005925)	
8	MAFA C1QL1 ABCC8 PCSK1 GSN	BP	protein targeting to ER (GO:0045047)	Pancreatic islet Vascular system Eye
		MF	RNA binding (GO:0003723)	
		CC	cytosolic ribosome (GO:0022626)	
9	PRSS2 CTRB2 OLFM4 SPINK1 AL049839.2	BP	cotranslational protein targeting to membrane (GO:0006613)	Intestine Pancreatic islet Bladder
		MF	RNA binding (GO:0003723)	
		CC	focal adhesion (GO:0005925)	
10	DPYSL3 PRG4 AC055839.2 PEG10 CASR	BP	SRP-dependent cotranslational protein targeting to membrane (GO:0006614)	AtT-20 cell Medulla oblongata Caudate nucleus
		MF	RNA polymerase II transcription factor binding (GO:0001085)	
		CC	cytosolic ribosome (GO:0022626)	
11	KRT19 MMP7 KRT18 PMEPA1 TACSTD2	BP	neutrophil degranulation (GO:0043312)	Intestine Pancreatic islet Bone
		MF	cadherin binding (GO:0045296)	
		CC	focal adhesion (GO:0005925)	
12	PLVAP PLPP3 FLT1 PECAM1 RGCC	BP	extracellular matrix organization (GO:0030198)	Vascular system Bone Intestine
		MF	collagen binding (GO:0005518)	
		CC	focal adhesion (GO:0005925)	
13	LAPTM5 TYROBP C1QB SDS C1QA	BP	neutrophil degranulation (GO:0043312)	Blood Bone Intestine
		MF	actin binding (GO:0003779)	
		CC	vacuolar lumen (GO:0005775)	

Discussion

By identifying cell types, novel cell type subpopulations, and performing an enrichment analysis, we were able to further classify diverse cell types in the pancreas using a similar workflow as Baron et al (2016). Overall, we were able to identify certain cell types from the marker genes in our data; however, not all cell types were found compared to Baron et al (2016). When we looked at the top two marker genes from each of our clusters, only Alpha, Beta, Delta, and Ductal cell type were identified. When we extended the search to all significant differentially expressed marker genes, six more clusters were labelled with cell type. These cell types were labeled Gamma, Acinar, Stellate, Vascular, and Macrophage. In total, we obtained 14 clusters from our analysis, and our initial matching process identified

nine of the clusters with cell type. Furthermore, we had five clusters with no definitive cell type. We were unable to identify any marker genes for Epsilon, Mast, and Cytotoxic T cell types, which was the main discrepancy between our findings. For the unidentified clusters, we performed literature searches for gene aliases and compared marker genes we found in the unidentified clusters to genes associated with these cell types. Additionally, we loosened the thresholds set to classify our clusters; however, all efforts were determined to be inconclusive. One 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.

Out of the five unlabeled clusters, we associated three of the clusters with already existing cell types. Alpha_2, Beta_2, and Delta_2 were identified based on shared marker genes and past studies that identified certain genes as cell types in the pancreas.[18] Lastly, the two unknown clusters were not identified with definitive cell types. However, the top two genes in cluster Unknown_A showed high expression in the Acinar cell type. Acinar and Unknown_A also had considerable overlap in the UMAP plot, which could mean that these two clusters share similar genes. We would expect this cluster to be identified best with Acinar cell type. The top genes in cluster Unknown_B showed high expression in both Ductal and Acinar with moderately higher expression in Acinar. Unknown_B was seen to be clustered close to both Ductal and Acinar in the UMAP, which could indicate this cluster is associated with one of these cell types.

Since our marker genes did not fully align with the marker genes Baron et al (2016) found, the projection plot and heatmap differed. Additionally, Baron et al (2016) used t-SNE to create their projection plot, while we used UMAP, which could explain the main differences in appearance. The clusters in our heatmap were less clear and definitive than the heatmap in the paper. This could also be due to our differing clusters and marker genes. Baron et al (2016) had three more cell types, Epsilon, Mast, and Cytotoxic T, that we had not discovered in our data set. Furthermore, we had unknown cell type clusters and multiple clusters for the same cell type, which influenced our projection plot and heatmap significantly. We identified many novel marker genes that were just as discriminative of cell type as the marker genes Baron et al (2016) used. These novel genes could adjust our labelled clusters to match what Baron et al (2016) found.

Based on the marker genes, we did gene enrichment analysis for the 14 clusters. By analyzing the top GO terms for each cluster, we got some results to support the previous analysis on cell type. For some clusters the biological process or molecular function from gene enrichment analysis clearly indicated the functions of cells which allowed us to be confident in our findings for some clusters (0,1,2,4,5,7,8,9,10,12,13). And we inferred the cell type for two unknown cell types (cluster 3 and cluster 6) by resulting GO terms and Jensen tissue from ENRICH.

We were able to find some resulting GO terms that support our findings. Cluster 13 contained GO terms neutrophil degradation (GO:0043312), neutrophil activation involved in

immune response (GO:0002283) and neutrophil mediated immunity (GO:0002446) indicated an immune cell, consistent with the previous analysis for macrophage. Cluster 12 contained GO terms like sprouting angiogenesis (GO:0002040) and vascular smooth muscle cell development (GO:0097084). And the tissue contained the Vascular system in a very small p-value. Thus, it indicated a vascular related cell. Cluster 7 contained GO term extracellular matrix organization (GO:0030198), and the Jensen tissue contained stellate cells in a very small p-value. For stellate cells, it is the major cell type involved in liver fibrosis, which is the formation of scar tissue in response to liver damage[19]. For clusters 0, 1, 2, 4, 5, 8, 9 and 10, the top tissues for these cells contained Pancreatic juice or Pancreatic islet, and they contained a lot of lumen like cellular components which strongly implies that these cells were secretory cells with secretory function. For cluster 11, we were unable to find clearly GO terms supporting our previous analysis for the ductal cell. But the Jensen tissue result showed that the Intestine and Pancreatic islet with a small p-value for this cell come from.

For cluster 3 and cluster 6, we got the unknown cell type from the marker genes. By previous analysis on marker genes and resulting GO terms from ENRICH, we inferred the cell type for these two clusters. For cluster 3, the Jensen tissue results contained a significant small p-value for pancreatic juice, so this cell should be present in the pancreatic juice. As we mentioned before, the top genes for this cluster showed high expression in Acinar cell type, this cell type may have similar functions with Acinar cells and also present in pancreatic juice. For cluster 6, it contained the GO terms in BP like negative regulation of actin filament polymerization (GO:0030837) and in MF like protein binding involved in cell-cell adhesion and other bindings. And the Jensen tissue result showed that the Bladder, Ascites and Vascular system with a small p-value. The top genes in cluster 6 showed high expression in both Ductal and Acinar cell type. These findings could indicate this cluster should be near Ductal and had similar functions with Ductal.

Above all, we identified 14 clusters and determined most cell types with their marker genes. The gene enrichment analysis provided support for the identified cell types by cell functions and biological processes. Single cell RNA sequencing (sc-RNA seq) provided a higher resolution of cellular differences and better understanding of individual cell functions. By sc-RNA seq, we can explore the complex systems beyond the different cell types and clarify the cell functions. It allows us to analyze the connection between certain types of cells and diseases, better understand disease mechanisms on the gene and cellular level. Furthermore, targeted and suitable treatments can be developed which has direct implications for the healthcare industry.

Conclusion

In conclusion, we reproduced some similar results from Baron et al.(2016). By analyzing the single-cell sequencing data from a 51-year-old female donor, we successfully found 14 clusters in the pancreas with marker genes and identified nine cell types. The original paper from Baron et al. (2016) identified 15 cell types. This difference can be explained by the subset of data we analyzed. Overall, single cell RNA-seq provides a higher resolution of cellular differences that allows researchers to better explain pancreatic biology

and better understand cellular diversity and individual cell functions.in the pancreas. Applications for this research in the pancreas can hopefully, in the future, give more profound insight into pancreatic biology and address pancreatic diseases such as diabetes that affect a great portion of the world's population.

References

1. Types of Cells in the Pancreas. (2020, August 13). Retrieved April 13, 2021, from https://med.libretexts.org/@go/page/7777
2. Grün D., Muraro M.J., Boisset J.-C., Wiebrands K., Lyubimova A., Dharmadhikari G., van den Born M., van Es J., Jansen E., Clevers H. De novo prediction of stem cell identity using single-cell transcriptome data. Cell Stem Cell. 2016;19:266–277.
3. Bugliani M., Liechti R., Cheon H., Suleiman M., Marselli L., Kirkpatrick C., Filipponi F., Boggi U., Xenarios I., Syed F. Microarray analysis of isolated human islet transcriptome in type 2 diabetes and the role of the ubiquitin-proteasome system in pancreatic beta cell dysfunction. Mol. Cell. Endocrinol. 2013;367:1–10.
4. 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. PMID: 27667365
5. inDrops pipeline : https://github.com/indrops/indrops
6. Klein, Allon M., et al. "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells." Cell 161.5 (2015): 1187-1201.
7. Hashimshony, Tamar, et al. "CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification." Cell reports 2.3 (2012): 666-673.
8. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525-527.
9. Charlotte Soneson, Michael I. Love, Mark D. Robinson (2015): Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research
10. Zhu, A., Srivastava, A., Ibrahim, J.G., Patro, R., Love, M.I. Nonparametric expression analysis using inferential replicate counts Nucleic Acids Research (2019)
11. Sarah E Hunt, William McLaren, Laurent Gil, Anja Thormann, Helen Schuilenburg, Dan Sheppard, Andrew Parton, Irina M Armean, Stephen J Trevanion, Paul Flicek, Fiona Cunningham Ensembl variation resources Database Volume 2018 doi:10.1093/database/bay119
12. Seurat—Guided clustering tutorial. (n.d.). https://satijalab.org/seurat/articles/pbmc3k_tutorial.html
13. Hao and Hao et al. Integrated analysis of multimodal single-cell data. bioRxiv (2020) [Seurat V4]
14. Single-cell RNA-seq Demo (10x non-small cell lung cancer). (n.d.). http://barc.wi.mit.edu/education/hot_topics/scRNAseq_2020/SingleCell_Seurat_2020.html
15. Wollheim, C. B., & Maechler, P. (2002). B-cell mitochondria and insulin secretion: Messenger role of nucleotides and metabolites. Diabetes, 51(suppl 1), S37–S42. https://doi.org/10.2337/diabetes.51.2007.S37
16. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
17. 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).

18. Muraro, Mauro J et al. “A Single-Cell Transcriptome Atlas of the Human Pancreas.” Cell systems vol. 3,4 (2016): 385-394.e3. doi:10.1016/j.cels.2016.09.002
19. Stanciu A, Cotutiu C, Amalinei C (2002). "New data about ITO cells". *Rev Med Chir Soc Med Nat Iasi*. **107** (2): 235–239. PMID 12638266