Single Cell RNA-Seq Analysis of Pancreatic Cells

Project 4: Team wheeler
Data Curator: Vishala Mishra (continues to be absent?)
Programmer: Reina Chau
Analyst: Ariel Xue
Biologist: Jessica Fetterman

Introduction

The pancreas consists of a number of diverse cell types related to its endocrine and exocrine functions. The endocrine cell types serve to regulate blood glucose levels in fasted and fed states, while the exocrine cell types produce the gastric enzymes that break down food items in the small intestine. The endocrine cells make-up a small fraction of the entire cell population and become dysfunctional in diabetes. Most studies of the mechanisms of the pancreatic endocrine cells have been restricted to animal models, extrapolating the findings to diabetes in humans. However, key differences in the organization of the endocrine cells in the pancreas have been noted between humans and mice.[1, 2] Whether the differences in pancreatic histology are reflected by differences in the cell populations or transcriptomes of the pancreatic endocrine cells is not understood. Baron *et al* sought to compare the cell types present in human and mouse pancreatic samples using single cell transcriptomics, which we tried to recapitulate in our analyses using different methodologies.[3] We used salmon alevin to map and quantify the reads after filtering out barcodes with low frequency counts. We further performed quality control on the quantified reads using Seurat package from Bioconductor and excluded low-quality cells and genes from downstream analysis. The gene expression measurements were normalized and principal components analysis (PCA) was performed to identity the dimensionality of the single-cell dataset. We then used UMAP to evaluate how the cells clustered based upon their transcriptomes and identified the differentially expressed genes that defined the clusters using K-nearest neighborhood (**KNN**). Lastly, we assessed the biological significance of the differentially expressed genes in each cell cluster with Metascape.[4]

Methods

Processing Single Cell Sequencing Reads

In this study by Baron *et al*, single cell sequencing was performed on approximately 10,000 human pancreatic cells collected at autopsy (N=4 donors) and approximately 10,000 murine pancreatic cells from 2 different mouse strains using inDrop.[3, 5] inDrop encapsulates each of the individual cells within a single droplet containing all components needed for reverse transcription, converting the RNA to cDNA. Paired-end sequencing was performed with an Illumina Hiseq 2500 with an average of 100,000 reads per cell.

In our analysis of the samples, 13 human single cell RNA-seq samples obtained from four donors were provided with preprocessed sequencing libraries. The barcodes in these samples were padded to be exactly 19 bases + 6 UMI bases = 25 bases. Out of the 13 given samples, we further investigated the samples from a 51-year-old female donor (SRR3879604, SRR3879605, and SRR3879606) and performed cell-by-cell quantification of the UMI counts.

Before the UMI counts were quantified, we evaluated the distribution of the reads among the barcodes. **Figure 1** shows a cumulative distribution of the barcode counts for all three samples. The majority of the barcodes had a very low frequency count (probably exactly 1 read per barcode). Therefore, to filter out these infrequent barcodes from each sample, we eliminated reads that did not meet the respective mean value of all the barcode counts. The mean of the three samples are 279, 193, and 193 respectively.

The barcodes that passed mean count were quantified with salmon alevin, an alignment tool for fast transcript quantification from single cell RNA-seq data.[6] The salmon output contains statistics on

how well the reads mapped to the reference genome. From the output log of salmon, we obtained a mapping rate of approximately 42.8%, which is quite low. After examining the quantification log, we found that ~3% reads are excluded due to noisy cellular barcodes (likely due to PCR-related artifact), and 13,071 barcodes were skipped due to no mapped reads. Therefore, we suspect the high noise resulted in our low mapping rate.

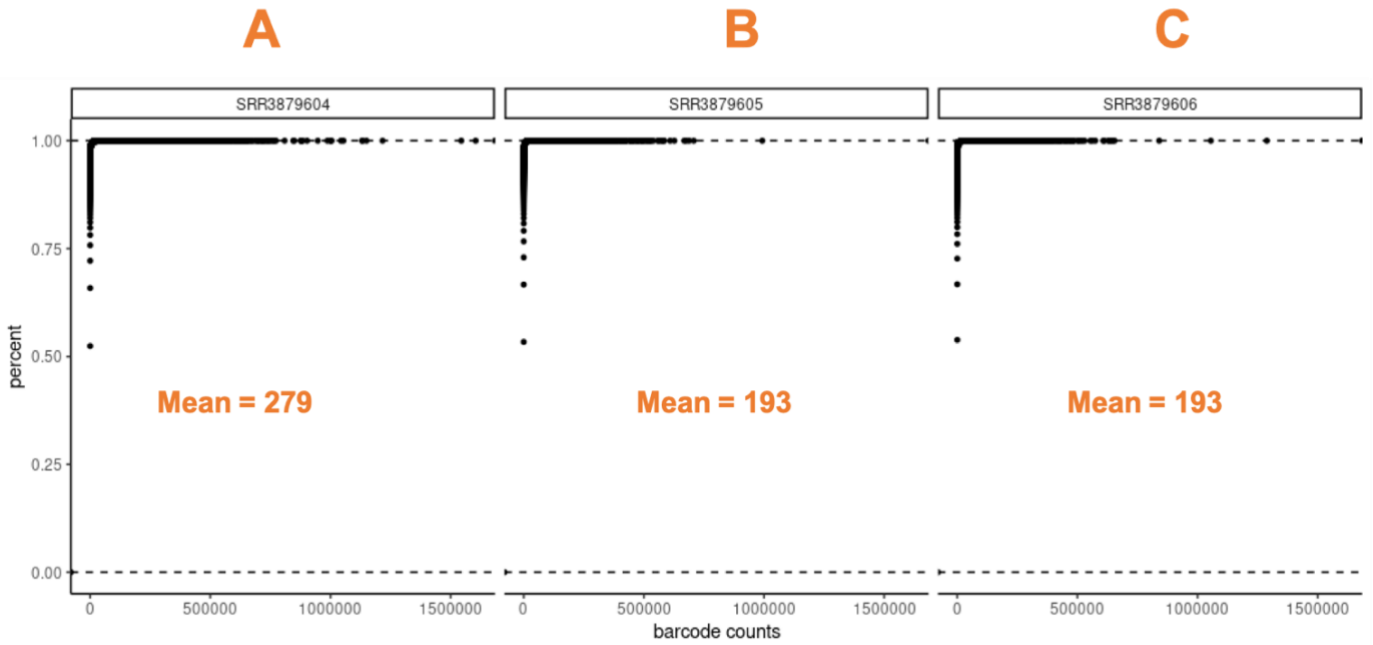

Figure 1: **The Distribution of Reads Among the Barcodes**. Cumulative distribution of the barcode counts for the 51-year-old female donor samples SRR3879604 (**A**), SRR3879605 (**B**), and SRR3879606 (**C**). The majority of the barcode counts for all three samples had a very low frequency (probably exactly 1 read per barcode). The infrequent barcodes were filtered out by excluding barcodes that have total reads less than the respective mean of all the barcode counts. The mean of the three samples are 279, 193 and 193 respectively.

Quality Control and UMI counts Matrix Filtering

After the samples were quantified with salmon, we obtained a UMI count matrix with 60,232 genes and 130,986 cells. To filter out the low-quality cells and genes for the downstream analysis, we utilized the Seurat package from Bioconductor to perform quality control on the UMI count matrix.[7] First, we created a Seurat object of the UMI counts matrix and only the cells that had a minimum count of 10 and genes with a minimum count of 200 were retained. Our initial filter resulted in 21,495 genes and 10,059 cells. Next, we visualized the feature counts using a violin plot (**Figure 2**) to identify and exclude additional low-quality cells. Typically, a low-quality cell contains very few genes, and likewise, cell doublets or multiplets can exhibit an aberrantly high gene count. Therefore, by removing cells that have unique feature counts less than 200 and over 4000, we retained a total of 21,495 genes and 9,885 cells. Furthermore, if a cell is low-quality or dying, the cell will exhibit extensive mitochondrial contamination. Thus, we excluded cells that had >10% mitochondrial counts, leaving us with a final sample set of 21,495 genes and 2,685 cells.

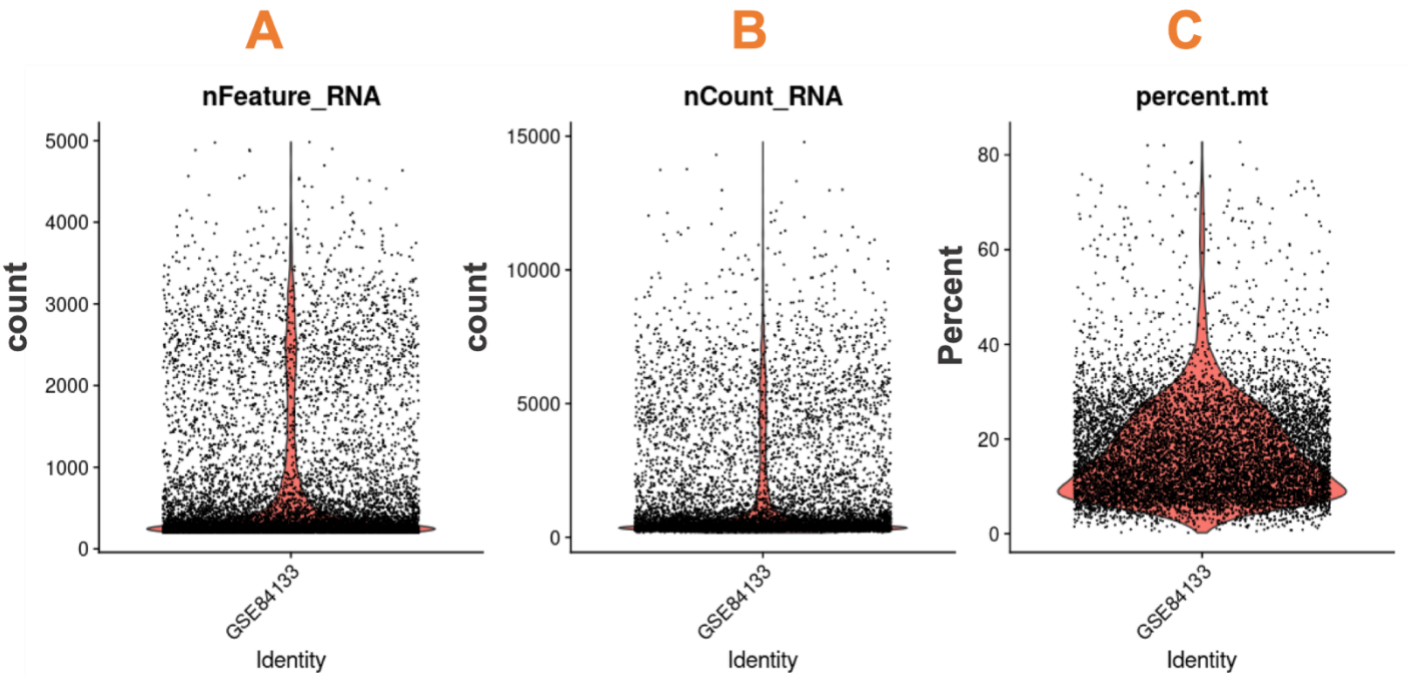

Figure 2: Violin plots were used to visualize the quality control metrics including the total number of features in a cell (**A**), total molecules detected within a cell (**B**), and the percent of reads that mapped to the mitochondrial genome (**C**) with the UMI counts matrix. The low-quality cells are filtered out by nFeature_RNA < 200 or nFeature_RNA > 4000 or percent.mt > 10%.

After we filtered the low-quality cells from the UMI counts matrix, we normalized the feature expression measurements for each cell by the total expression, then multiplied the normalized values by a scale factor of 10,000 and log-transformed the expression values. The normalization and log transformation reveal cell-to-cell variation in transcript levels (i.e. gene transcripts that are highly expressed in some cells, and lowly expressed in others), aiding in the identification of differentially expressed genes. **Figure 3** shows the top 10 highly variable genes (*REG1B*, *PRSS2*, *PLVAP*, *REG3A*, *CTRB2*, *PPY*, *REG1A*, *PRSS1*, *CTRB1* and *ALB*) that are outliers based upon the mean variability plot from Seurat.

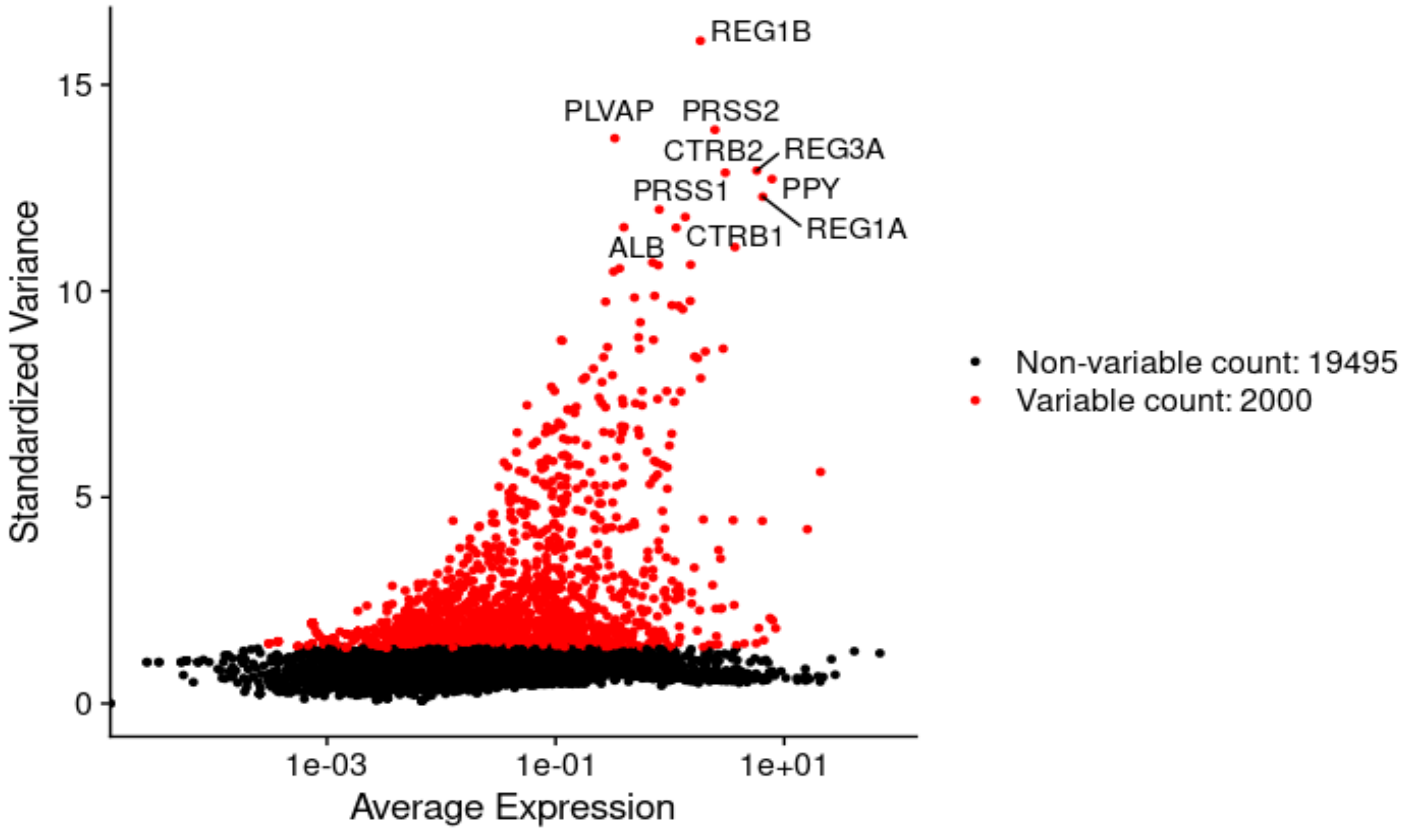

Figure 3: Top ten highly variable genes that are outliers based upon the mean variability plot from Seurat

Principal Component Analysis (PCA) and Cell Clustering

We scaled the normalized expression of each gene so that the mean expression across cells was 0 and the variance was 1. Next, we performed a PCA on the scaled data to identify the principal components (PCs) that explained the most variation in the dataset. **Figure 4** shows the first two components of the PCA. The first PC explains about 7.9% of the variance while the second PC explained ~6.6% of the variance. **Figure 5** shows an elbow plot, which ranks the PCs based on the percentage of variance explained by each PC. We observed an “elbow” around PC15, suggesting that the majority of the variance is explained by the first 15 PCs.

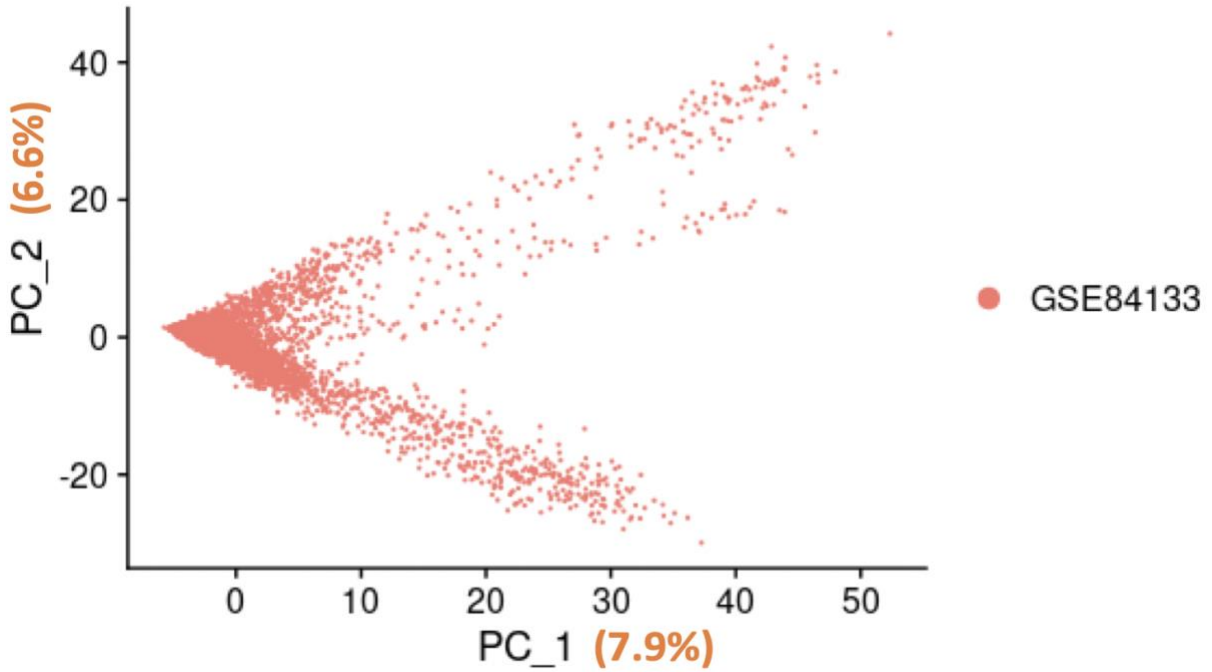

Figure 4: Dimension plot of the top two components of PCA. The 1PC explained ~7.9% of the overall variance while PC2 explained about 6.6% of the variance.

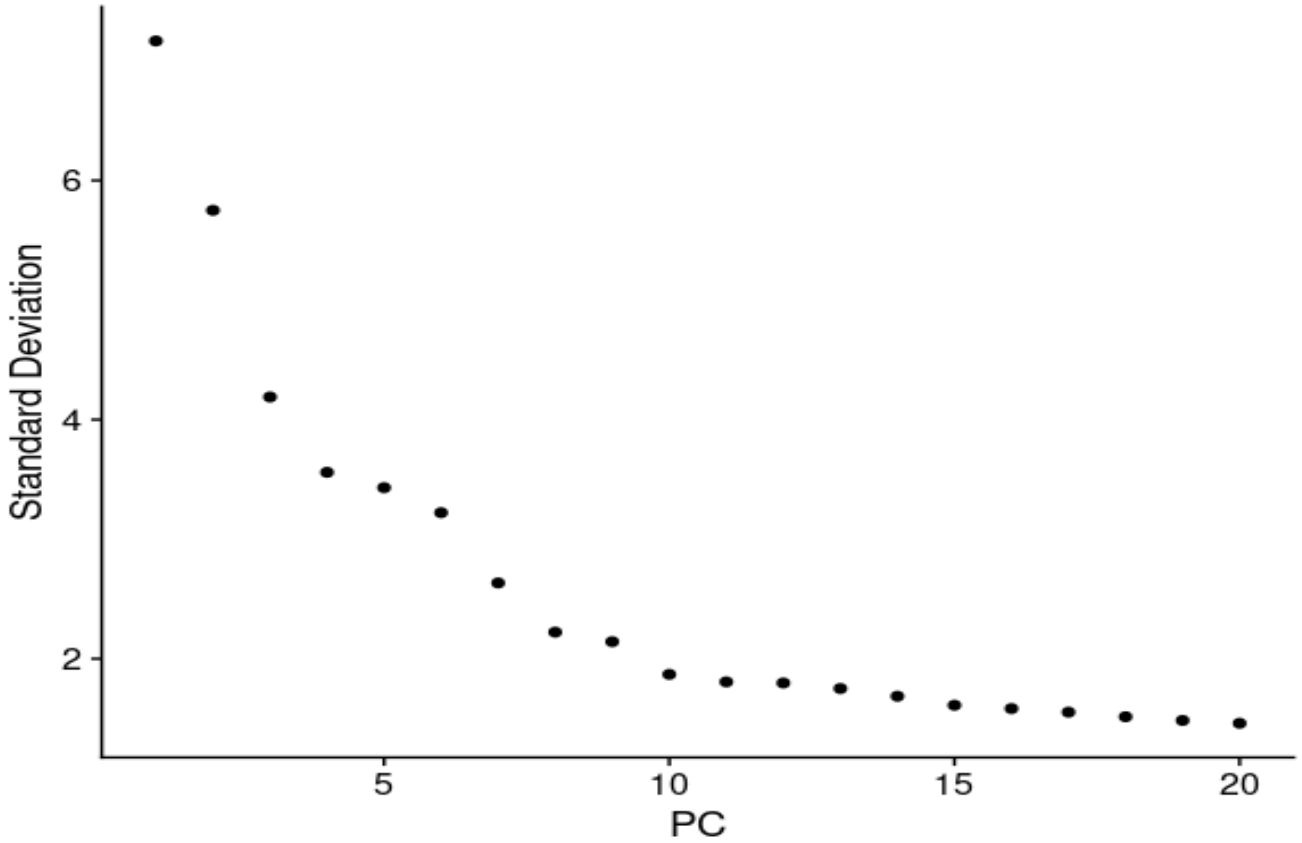

Figure 5: Elbow plot of the first 20 components of the PCA. The variance explained by the PCs are shown in descending order from the PCs explaining the greatest amount of the variation in the data to

the PCs explaining the smallest amount of the variation. An “elbow” was observed around PC15, suggesting that the majority of the variance is explained by the first 15 PCs.

Once we knew the dimensionality of the data (15 PCs), we calculated the Euclidean distance in the PCA space and implemented a K-nearest neighbor (**KNN**) algorithm to cluster cells with similar feature expression patterns, partitioning the cells into highly interconnected communities. The KNN is a supervised machine learning algorithm that is used to solve both classification and regression problems.[7] The KNN algorithm is based upon the premise that similar objects exist in close proximity to each other. After the cells were classified into their relative nearest neighborhood, we applied a modularity optimization technique known as the Louvain algorithm or SLM, to cluster the cells together. The SLM method was implemented using the *FindClusters* function in the Seurat package, which contains a resolution parameter that sets the ‘granularity’ of the downstream clustering, with increased values leading to a greater number of clusters. The Louvain method is an algorithm that is used for identifying communities in large networks. We obtained a total of 10 communities. The relative proportions of cell numbers identified within these 10 communities are reported in **Figure 6**.

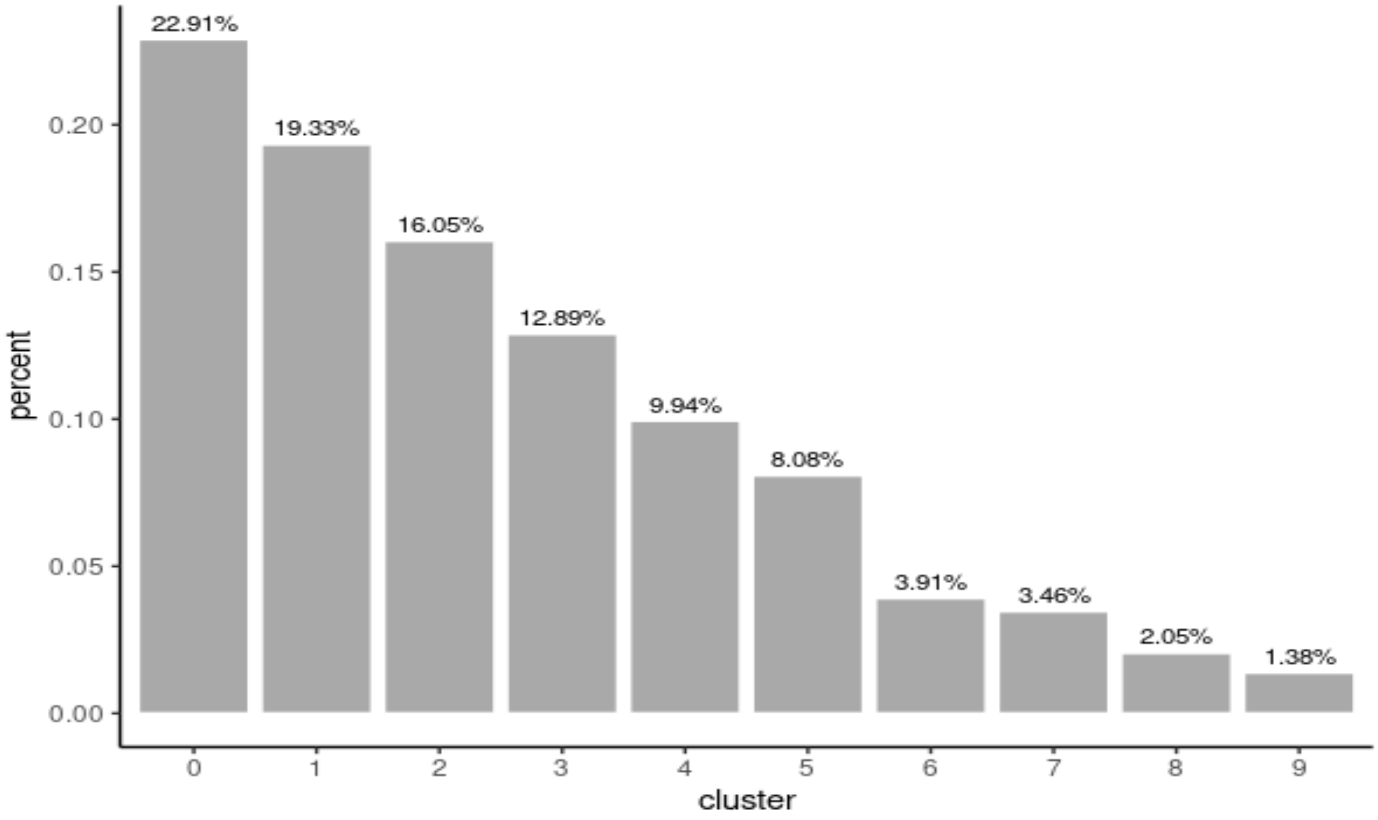

Figure 6: Communities of Clustered Cells. Relative proportions of cell numbers identified for each of the 10 communities using K-nearest neighborhood algorithm.

Gene Enrichment Analysis

The differentially expressed genes identified for each cluster were filtered based upon a fold change >0 and an adjusted p-value of <0.05. The filtered resulting gene lists for each cell cluster were put into Metascape in order to identify the top enriched pathways.[8]

Results

Identification of Gene Markers

The Seurat package was further utilized to identify the differentially expressed genes defining each of the 13 clusters provided in the RDS file.[7] The command *FindAllMarkers* was used for the loaded Seurat object with the parameters for minimum gene expression percentage detected of 0.25 and a minimum average log2 fold change of 0.25. As the log2 fold change refers to the log-ratio of a gene's average of 1.19 fold change in the cluster compared to the rest of the cells, 19% is considered an appropriate threshold for defining significance in a large data set. After testing several potential thresholds for log2 fold change, 0.25 could effectively assign cell types to each cluster without filtering out significant markers. This command compared genes in each cluster with the rest of the clusters, identifying 6,487 DEGs as potential gene markers for all clusters. The top 2 DEGs with highest log2 fold change for each cluster were reported as cluster-defining genes in **Table 1**. The top 10 DEGs with highest log2 fold change were shown via command *DoHeatmap* in **Figure 7**.

Table 1. Top 2 gene markers for each cluster with highest log2 fold change.

Cluster ID	Markers
0	SP100, PRRG3
1	DLK1, INS
2	FN1, COL1A1
3	REG1B, REG1A
4	TTR, GCG
5	CXCL1, KRT19
6	INS, EEF1A2
7	ACER3, AL022322.2
8	TTR, GC
9	COL1A1, COL3A1
10	CRP, KRT18
11	ALDOB, PRSS2
12	ACP5, AC007192.1

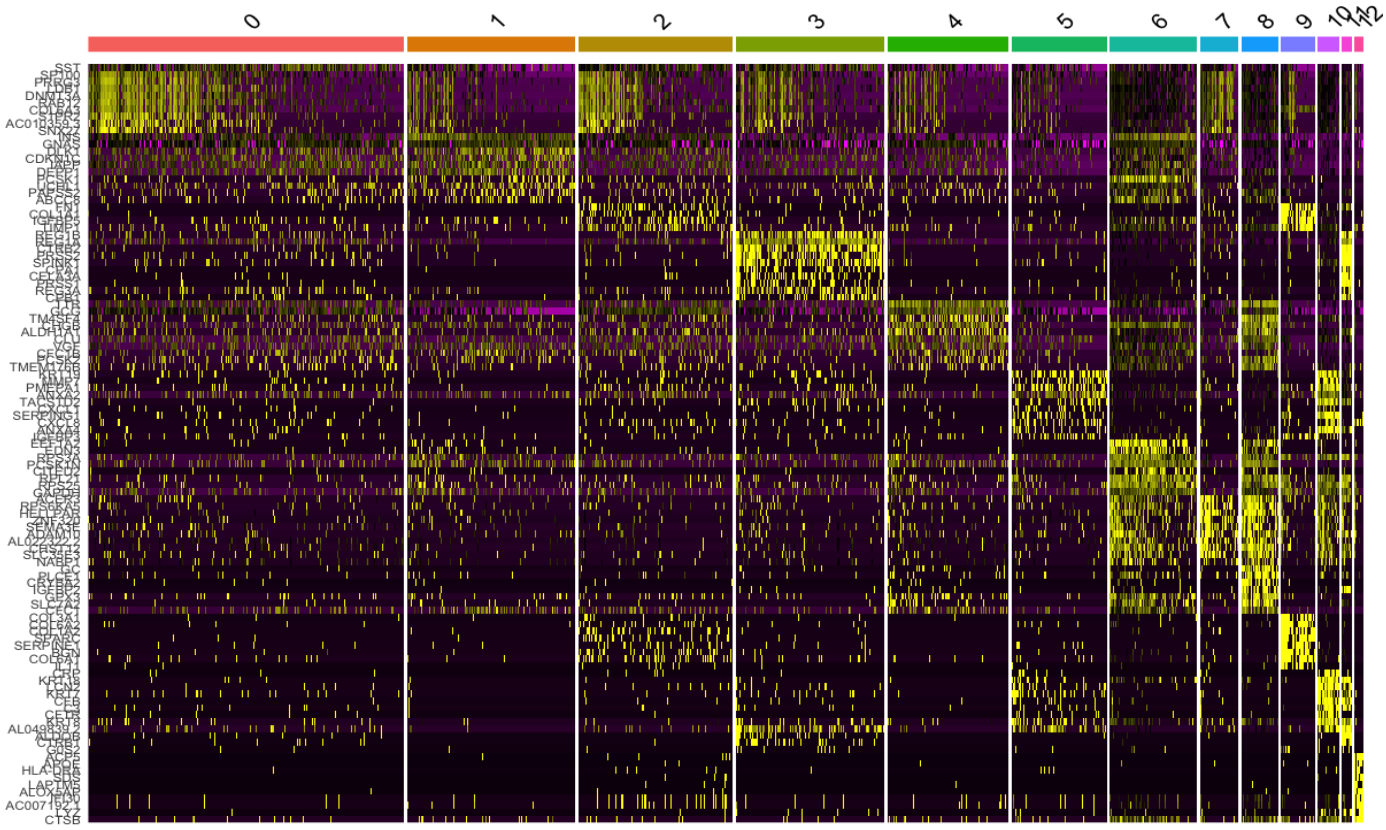

Figure 7. Top 10 gene markers with the highest log2 fold change for each of the 12 cell clusters. The X axis is cluster ID (also represented by the blocks of color) and the Y axis are the gene names.

Labeling the Cell Clusters Based Upon Expression of Marker Genes

Baron *et al* provided a list of identified cell types with their respective markers in their report (**Table 2**), and this list is referred here for the cell type assignment as well.[3] Based upon the assumption that a gene with a log2FC larger than 0.25 is considered as a marker gene, we further assumed that if we find one cell type's known gene marker in the DEG list with minimum log2FC of 0.25 detected above, the corresponding cluster(s) could be labeled as such cell type. To implement this assumption, each known marker was searched in the previously obtained DEG list and then all of the clusters expressing the gene markers were labeled with the corresponding cell type. In order to the distribution of the gene markers, UMAP figures are produced for each marker to help determine cell types (**Figure 8**). The UMAPs show that some known markers are highly expressed in multiple clusters while some other markers show low expression in all clusters. This is consistent with the direct search in the marker gene list, and the final assigning results are recorded in **Table 2** - If one cluster contains multiple known markers, it will be labeled with multiple cell types; Similarly, if one known marker is found in multiple clusters, all these clusters will be labeled with the same cell type; If a known marker is not found in any of the clusters, it will be reported as none.

The results show that cluster 0 is a combination of delta and gamma cells and cluster 3 is a combination of delta and acinar cells. No clusters for the cell types of epsilon, quiescent stellate, activated stellate, endothelial, cytotoxic, mast were found (**Table 2**). UMAP was used to visualize the clustered cell with labeled cell types (**Figure 9**) except for cluster 7 and 9 because there are no known markers found (**Table 2**) and their cell types remained to be discussed.

The Seurat object we used was first normalized with method *LogNormalize* in order to obtain log normalized UMI counts, and then **Figure 10** was made to show the heatmap of log normalized UMI

counts for the top 2 marker genes of each cluster, the corresponding marker for each cluster is also shown in **Table 1**. This heatmap of UMI counts does not clearly suggest distinct cell types, which is consistent with **Figure 7** using log2FC as expressions, because some markers are highly expressed in multiple clusters while some markers could not represent the whole cluster well. For example, *GCG* as one of the top 2 marker genes in cluster 4, showed high values for UMI counts in many other clusters such as cluster 0, 2, 3, 8, 10, 11. **Figure 10** also shows that the top 2 marker genes in cluster 0 (*SP100* and *PRRG3*) with highest log2FC could not fully represent the whole clusters because only part of the UMI counts show high expression in this cluster. Another problem it reflects is the top 2 marker genes in cluster 2 (*FN1*, *COL1A1*) actually have much lower expression levels compared to other clusters' markers in cluster 2, such as *SP100* and *GCG*. At the same time, *FN1* and *COL1A1* have much higher expression in cluster 9 (log2FC of 3.77 and 4.49 respectively) compared to that in cluster 2 (log2FC of 1.95 and 1.63 respectively). Due to the definition of marker gene was solely on differential expression conveyed by log2FC, many problems as mentioned above appeared and these explanations and interpretations will be discussed later.

Table 2. Known markers for each cell type and the corresponding clusters identified.

Known Markers	Cell type	Labeled Clusters
GCG	Alpha	2, 4, 8
INS	Beta	1, 6
SST	Delta	0, 3
PPY	Gamma	0
GHRL	Epsilon	None
KRT19	Ductal	5, 10
CPA1	Acinar	3, 11
RGS5	Quiescent stellate	None
PDGFRA	Activated stellate	None
VWF	Endothelial	None
SDS	Macrophage	12
TRAC	Cytotoxic	None
TPSAB1	Mast	None

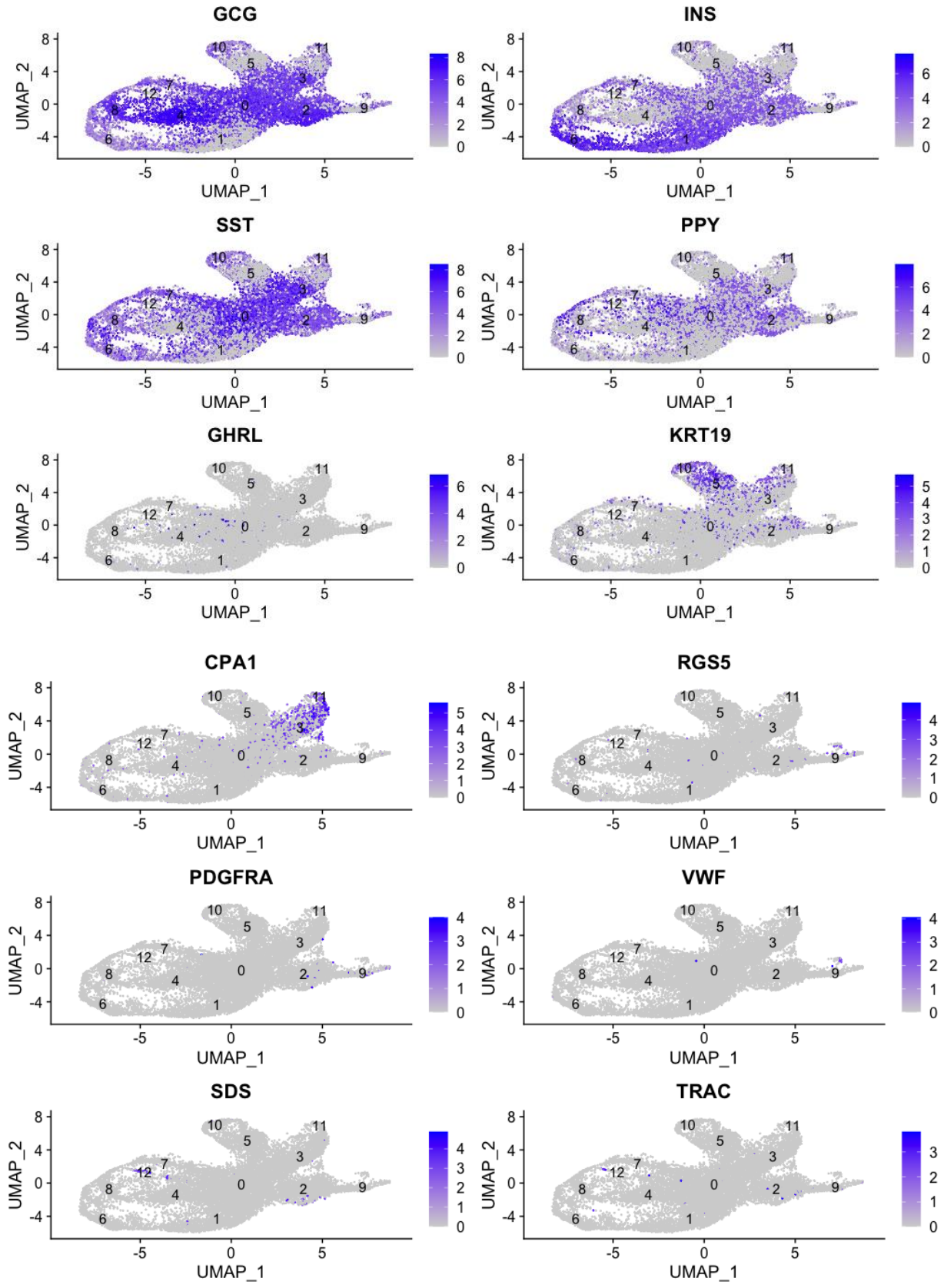

Figure 8. UMAPs for each known marker gene showing the distribution of its expression in different clusters. The darker the blue, the greater the expression of the marker gene.

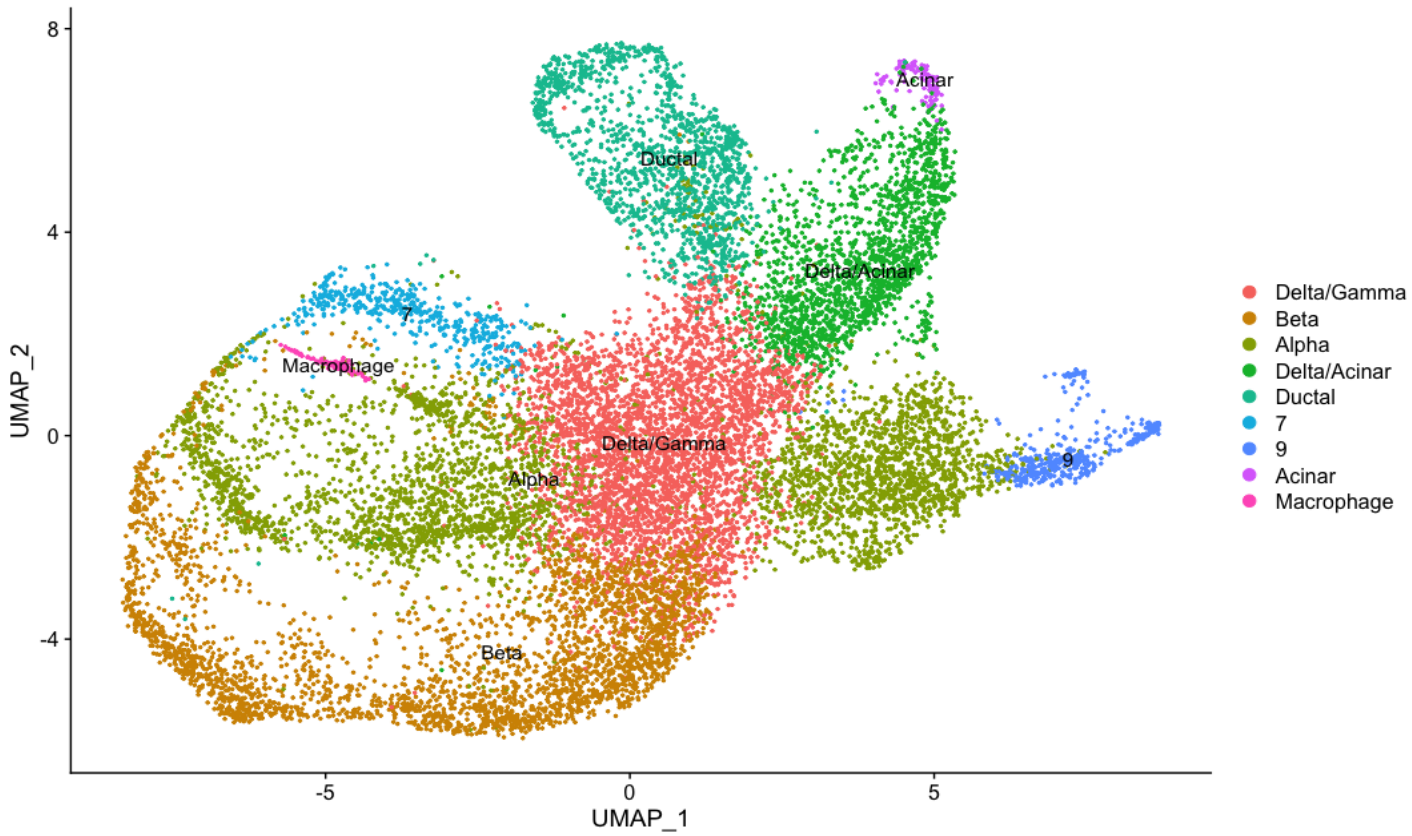

Figure 9. UMAP for clustered cells with labeled cell types. 7 and 9 are still cluster IDs which indicate that they are not yet able to be labeled by known markers.

Figure 10. Heatmap of log normalized UMI counts for the top 2 marker genes in each cluster.

Identification of Enriched Pathways in Each Cluster

In order to identify the pathways that were enriched by cell cluster, the differentially expressed genes for each cell cluster were filtered based upon a fold change >0 and an adjusted p-value of <0.05. The filtered resulting gene lists put into Metascape.[6] The top pathways that were enriched in each cluster are presented in **Table 3**. Gene enrichment using all of the differentially expressed genes defining each cluster resulted in a very similar list of top enriched pathways as the filtered differentially expressed genes. The similar pathways identified using the filtered and unfiltered differentially expressed datasets likely reflects the greater contribution of genes with higher fold changes that likely drive the enrichment scores.

Table 3. Gene Enrichment by Cluster

	GO	Description	Log10(q)
Cluster 0	GO:0006119	Oxidative phosphorylation	-4.06
n=27 genes	GO:0001666	Response to hypoxia	-0.80
	GO:0009152	Purine ribonucleotide biosynthetic process	0
	R-HSA-373076	Class A/1 (Rhodopsin-like receptors)	0
Cluster 1	GO:0010817	Regulation of Hormone Levels	-6.6
n=33 genes	ko04940	Type 1 diabetes mellitus	-3.3
	GO:0016486	Peptide hormone processing	-2.1
	M106	PID HNF3B pathway	-1.7
	hsa04913	Ovarian steroidogenesis	-1.5
	GO:0043408	Regulation of MAPK cascade	-0.9
	GO:0007507	Heart development	-0.6
	GO:0097529	Myeloid leukocyte migration	0.0
Cluster 2	R-HSA-1474228	Degradation of the extracellular matrix	-1.6
n=16 genes	GO:0043434	Response to peptide hormone	-1.6
	R-HSA-8957275	Post-translational protein phosphorylation	-1.4
	hsa04714	Thermogenesis	-0.4
Cluster 3	R-HSA-156902	Peptide chain elongation	-48.4
n=64 genes	CORUM:308	60S ribosomal subunit, cytoplasmic	-19.4
	GO:002181	Cytoplasmic translation	-7.9
	GO:0042255	Ribosome assembly	-5.9
	GO:0044278	Cell wall disruption in other organism	-5.0
	ko04974	Protein digestion and absorption	-3.4
	GO:0045055	Regulated exocytosis	-2.9
	GO:1902176	Negative regulation of oxidative stress-induced intrinsic apoptotic signaling pathway	-2.6
	GO:0034248	Regulation of cellular amide metabolic process	-1.7
	GO:0051438	Regulation of ubiquitin-protein transferase activity	-1.4
Cluster 4	GO:0010817	Regulation of hormone levels	-4.3
n=29 genes	R-HSA-8957275	Post-translational protein phosphorylation	-3.0
	R-HSA-2980736	Peptide hormone metabolism	-2.0
	GO:0042445	Hormone metabolic process	-1.9
	GO:0002274	Myeloid leukocyte migration	-1.1
	GO:0010951	Negative regulation of endopeptidase activity	-0.9

	GO:0019730	Antimicrobial humoral response	-0.5
	GO:0097529	Myeloid leukocyte migration	-0.1
Cluster 5	GO:0045055	Regulated exocytosis	-8.9
n=64 genes	GO:0009611	Response to wounding	-6.9
	R-HSA-109582	Hemostasis	-6.7
	WP3888	VEGFA-VEGFR2 signaling pathway	-6.4
	R-HSA-1280215	Cytokine signaling in immune system	-5.3
	WP4659	Gastrin signaling pathway	-5.2
	M5885	NABA matrisome associated	-5.1
	WP4754	IL-18 signaling pathway	-5.1
	GO:0019083	Viral transcription	-4.0
	GO:002480	Antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent	-3.5
	GO:0030260	Entry into host cell	-3.3
Cluster 6	R-HSA-156842	Eukaryotic translation elongation	-63.7
n=673 genes	ko04141	Protein processing in endoplasmic reticulum	-17.8
	WP111	Electron transport chain (OXPHOS system in mitochondria)	-16.4
	CORUM:5380	TRBP containing complex (DICER, RPL7A, EIF6, MOV10, and subunits of the 60S ribosomal particle)	-13.8
	GO:0010817	Regulation of hormone levels	-10.7
	GO:0010035	Response to inorganic substance	-9.9
	GO:0042273	Ribosomal large subunit biogenesis	-9.7
	GO:0045055	Regulated exocytosis	-9.2
	WP3888	VEGFA-VEGFR2 signaling pathway	-8.3
	hsa04911	Insulin secretion	-8.0
Cluster 7	WP111	Electron transport chain (OXPHOS system)	-8.9
n=146 genes	GO:0015988	Energy coupled proton transmembrane transport, against electrochemical gradient	-5.5
	ko04260	Cardiac muscle contraction	-3.1
	hsa03008	Ribosome biogenesis in eukaryotes	-2.1
	GO:1900118	Negative regulation of execution phase of apoptosis	-1.2
	M163	PID LIS1 pathway	-1.0
	GO:0030072	Peptide hormone secretion	-0.5
	R-HSA-422475	Axon guidance	-0.5
	GO:0031532	Actin cytoskeleton reorganization	-0.4
	GO:0009152	Purine ribonucleotide biosynthetic process	-0.2
	GO:0008333	Endosome to lysosome transport	-0.1
Cluster 8	R-HSA-156842	Eukaryotic translation elongation	-32.0
n=949	R-HSA-1428517	Citric acid cycle and respiratory electron transport	-18.7
	GO:0044257	Cellular protein catabolic process	-6.8
	GO:006457	Protein folding	-6.5
	GO:0045055	Regulated exocytosis	-6.4
	GO:0022613	Ribonucleoprotein complex biogenesis	-6.1

	ko04141	Protein processing in endoplasmic reticulum	-5.2
	GO:0009914	Hormone transport of small molecules	-5.1
	R-HSA-199991	Membrane Trafficking	-5.0
	CORUM:5380	TRBP containing complex (DICER, RPL7A, EIF6, MOV10, and subunits of the 60S ribosomal particle)	-5.0
Cluster 9	R-HSA-156842	Eukaryotic translation elongation	-80.9
n=247 genes	WP3888	VEGFA-VEGFR2 signaling pathway	-32.6
	R-HSA-1474244	Extracellular matrix organization	-26.8
	GO:0001568	Blood vessel development	-16.8
	R-HSA-114608	Platelet degranulation	-16.3
	R-HSA-381426	Regulation of insulin-like growth factor (IGF) transport and uptake by insulin-like growth factor binding protein (IGFBPs)	-12.6
	R-HSA-1280215	Cytokine signaling in immune system	-11.2
	GO:0030199	Collagen fibril organization	-11.1
	GO:0045055	Regulated exocytosis	-11.1
	GO:0070848	Response to growth factor	-10.7
Cluster 10	R-HSA-156842	Eukaryotic translation elongation	-39.8
n= 987	WP3888	VEGFA-VEGFR2 signaling pathway	-33.5
	GO:0045055	Regulated exocytosis	-30.6
	GO:0043043	Peptide biosynthetic process	-30.0
	GO:0030155	Regulation of cell adhesion	-27.7
	GO:0030029	Actin filament-based process	-25.3
	GO:0009611	Response to wounding	-23.5
	GO:0097190	Apoptotic signaling pathway	-23.1
	GO:0040017	Positive regulation of locomotion	-21.4
	GO:0030855	Epithelial cell differentiation	-21.4
Cluster 11	CORUM:306	Ribosome, cytoplasmic	-96.9
n= 781	CORUM:5380	TRBP containing complex (DICER, RPL7A, EIF6, MOV10, and subunits of the 60S ribosomal particle)	-26.4
	WP3888	VEGFA-VEGFR2 signaling pathway	-26.0
	GO:0002274	Myeloid leukocyte migration	-16.6
	GO:0042255	Ribosome assembly	-15.6
	GO:0097190	Apoptotic signaling pathway	-14.2
	WP107	Translation factors	-12.9
	GO:0010942	Positive regulation of cell death	-12.3
	ko04141	Protein processing in endoplasmic reticulum	-10.7
	GO:0006417	Regulation of translation	-10.1
Cluster 12	GO:0002274	Myeloid leukocyte activation	-32.7
n= 98	ko04142	Lysosome	-15.4
	WP3888	VEGFA-VEGFR2 signaling pathway	-10.8
	GO:0002683	Negative regulation of immune system process	-10.0
	GO:0002253	Activation of immune response	-9.8
	WP3937	Microglia pathogen phagocytosis pathway	-7.1
	GO:0002697	Regulation of immune effector process	-7.0

	ko04145	Phagosome	-6.9
	R-HSA-109582	Hemostasis	-6.5
	GO:0019882	Antigen processing and presentation	-6.4

Discussion

We identified a total of 12 cell clusters in the human pancreatic sample, which was lower than the 14 clusters identified in the manuscript by Baron et al.[3] We further sought to identify the potential cell type composition of each of the clusters we identified based upon their transcriptomes. However, the assignment of a cell type based upon the top DEGs for each cluster was not straightforward and for some clusters, we were unable to assign a cell type due to the absence of the genes considered to be cell type-defining used by Baron and colleagues.[3] Using gene enrichment analyses allowed us to further assign cell types to cell clusters, but not always- some of the clusters had gene enrichments in pathways that were not cell type specific (e.g., mitochondrial oxidative phosphorylation, ribosomes).

Our UMAP (**Figure 9**) showed a different projection for the cell types with Figure 1D reported by Baron et al. This inconsistency could potentially be due to: 1) differences in input data - there is variance in upstream filtering and clustering of raw data, and 2) differences in analytical methodology and software package - here we used Seurat while the original paper used bseqsc. However, based upon the projects in Figure 1D by Baron et al, we estimate that cluster 9 in our dataset is neighbored to alpha cells, and close to the delta/acinar group.[3] From Baron *et al*'s Figure 1D, there is one such cell type that has a similar neighborhood - activated stellate cells.[3] Thus, with all the information available, cluster 9 may be activated stellate cells, but the cell type for cluster 7 is still unclear.

Similarly, the heatmap of the log normalized UMI counts in Figure 10 is inconsistent with Figure 1B reported by Baron *et al* with our heatmap lacking obvious marker genes for each cluster. Multiple clusters had high expression of the same cell type-defining genes. The marker genes in one cluster may not fully reflect the expression of the entire cluster. For example, *SP100* and *PRRG3* in cluster 0 were only expressed by about half of the cells in the cluster and likely drove the higher average fold change for these genes compared to the other clusters, which may be a result of using differential expression to determine the marker genes. As the top marker genes are here defined by highest average log2FC within a cluster compared to the average log2FC for the rest of cells, firstly they are only a representation of the average expression in a cluster, and secondly, they are based on a relative level compared to the whole sample data. Therefore, the marker genes used might not be meaningful on a biological level.

Using the top DEGs in each cluster as markers, only *INS* in cluster 2 and *GCG* in cluster 4 were among the top 2 marker genes detected. Cluster 7 and 9 could not be assigned to a specific cell type using known markers because none of the cell type defining markers had a log2FC larger than 0.25. Clusters 0, 1, and 2 are the clusters with largest numbers of cells but their highest average log2FC for the top DEGs are 1.39 (*SP100*), 1.81 (*DLK1*), and 1.95 (*FN1*), respectively, which is relatively low compared to other top genes such as *ACP5* (average log2FC=5.64) in cluster 12 and *COL1A1* (average log2FC=4.49) in cluster 9. Coincidentally, both cluster 2 and cluster 9 have expressed a gene marker of *COL1A1*, but their actual log2FC have almost a 3-fold difference in expression (1.63 vs. 4.49). If we determined the DEGs that separate cluster 2 and cluster 9 from each other, it is interesting to see that the top gene is *GCG,* which shows an average log2FC of 3.00 (with pct.1 = 0.922, pct.2 = 0.548, and adjusted p-value = 7.88e-127). As *GCG* is one of the known markers for alpha cells, this is why cluster 2 was assigned to the alpha cell type while cluster 9 was not. This is another example of how top DEGs cannot be used directly as marker genes - if we consider the top 2 DEGs in cluster 2, including *COL1A1,* as the marker genes for its cell type, cluster 9 with its higher

expression of the same marker gene (*COL1A1*), we would not be able to delineate cluster 9 cells from cluster 2.

Although the projection of the cell clusters in our UMAP is not entirely consistent with the original paper, some similarities were observed - delta cells and acinar cells are close to each other, which possibly explains why cluster 3 could not fully separate the 2 cell types based upon the marker genes. We also observed three clusters that were considered to be alpha cells based upon their differential expression of marker genes are not neighbored to each other (cluster 2, 4 and 8) and are separated by beta/gamma cells. This is probably due to the subjective threshold of log2FC of 0.25- cluster 2 has an average log2FC of 0.37 for the marker gene *GCG* and if a more stringent threshold for DEG cell type-defining genes was considered (e.g., ≥0.50), cluster 2 would likely no longer be labeled as alpha cells.

Cluster 9 cells had high *COL1A1* and *COL3A1* expression with none of the other cell clusters having detectable expression of these 2 genes. We could not readily classify the cell type of cluster 9, however, as the two collagen genes with high expression are not necessarily specific to a cell type. *ALDOB* expresses only strongly in cluster 11 but not cluster 3. As we previously detected acinar cells in both cluster 3 and 11, it may be important to use *ALDOB* to separate these two clusters further apart. Also, *ACP5* is only highly expressed in cluster 12, which was identified as macrophage cells, so ACP5 might be another marker gene for such cell type. Our data suggests that the top DEGs are insufficient for detecting the cell type defining genes as the threshold may be subjective and would benefit from a uniform standard.

We further attempted to classify the cell types using gene enrichment analyses. The cells in cluster 0 expressed both *SST* and *PPY*, which define the delta and gamma cells, respectively. The pathway enrichment analysis for cluster 0 was non-specific with mitochondrial respiration and a response to hypoxia being the top pathways, which provides little additional information on the specific cell type of the cells belonging to this cluster.

Cluster 1 likely contains beta cells as the cells in this cluster express insulin (*INS*) and have gene enrichment in the type 1 diabetes mellitus pathway, which would be expected to be enriched in beta cells, which become dysfunctional in diabetes. The cells belonging to cluster 6 are likely endocrine cells due to the enrichment of genes in insulin secretion and regulation of hormone levels. Of note, insulin was one of the differentially expressed genes defining cluster 6, suggesting that this cluster may also contain beta cells. Clusters 1 and 6 were both defined by the expression of insulin, indicative of beta cells, which is consistent with the findings of Baron *et al* who found heterogeneity in the gene expression patterns in the beta cells. Similar to the findings of the authors, cluster 6 cells had an enrichment in differentially expressed genes involved in pathways related to the endoplasmic reticulum and unfolded protein response that was not observed in the cluster 1 cells.

Cluster 2 is a cell type that is responsive to hormonal regulation based upon the gene enrichment pathways, which indicate a response to peptide hormone and thermogenesis. Cluster 3 cells are likely exocrine cells due to the gene enrichment in protein digestion and absorption, which would be expected of cells that produce and secrete digestive enzymes. Consistent with this, *CPA1* was a differentially expressed gene identified in cluster 3, which defines acinar cells. Based upon the gene enrichment in hormone pathways (regulation of hormone levels, peptide hormone metabolism, hormone metabolic process), cluster 4 cells are likely an endocrine cell type, although the enriched pathways provide little additional insight into which specific endocrine cell type. The cells in cluster 5 are also likely endocrine cells due to the enrichment of differentially expressed genes in gastrin signaling- gastrin is a hormone that stimulates cells in the stomach to produce and secrete acids and

stimulates motility of the gastro-intestinal system. Interestingly, cluster 5 cells express *KRT19*, which is a ductal cell defining gene.

Cluster 7 cells are likely a neuronal cell type, perhaps Schwann cells, due to their enrichment in genes in the neuronal development pathway PID LIS1 and axon guidance. Due to their enrichment for genes involved in hormone transport, it is possible that the cells in cluster 8 are involved in the endocrine functions of the pancreas. Cluster 9 cells may be endothelial cells due to the enrichment of genes in VEGFA-VEGFR2 signaling and blood vessel development. Cells in cluster 10 are likely ductal cells based upon their expression of *KRT19* and the gene enrichment for epithelial cell differentiation. Cluster 11 cells expressed *CPA1*, which is a gene that's expression defines acinar cells. Lastly, cluster 12 cells are likely immune cells due to their expression of genes enriched in leukocyte activation and immune system responses. Further, cluster 12 cells expressed *SDS*, which is a gene expressed specifically in macrophages.

Conclusion

Despite using multiple methods to classify the cell types for the clusters in the pancreatic samples, we were unable to make cell type assignments for a number of the clusters due to the absence of clear marker gene expression and the top enriched pathways being non-specific cellular processes. The differences in our findings from those by Baron and colleagues could be the result of different methods such as the use of different alignment tools and annotation.

References

1. Grube D and Bohn R. The microanatomy of human islets of Langerhans, with special reference to somatostatin (D-) cells. *Arch Histol Jpn*. 1983;46:327-53.
2. Orci L and Unger RH. Functional subdivision of islets of Langerhans and possible role of D cells. *Lancet*. 1975;2:1243-4.
3. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM, Melton DA and Yanai I. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst*. 2016;3:346-360 e4.
4. Huang da W, Sherman BT and Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4:44-57.
5. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA and Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*. 2015;161:1187-1201.
6. Patro R, Duggal G, Love MI, Irizarry RA and Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017;14:417-419.
7. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, Hao Y, Stoeckius M, Smibert P and Satija R. Comprehensive Integration of Single-Cell Data. *Cell*. 2019;177:1888-1902 e21.
8. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C and Chanda SK. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun*. 2019;10:1523.