# **Project 4 Write-up**

Abhishek Thakar (Programmer), Allison Nau (Analyst), Mae Rose Gott (Data Curator), and Sheila Yee (Biologist)

# **Introduction**

The pancreas, a complex organ involved in the digestive and endocrine system, is essential for regulation of blood sugar levels via secretion of hormones like insulin and glucagon as well as the secretion of digestive enzymes to aid in proper digestion and maintenance of stomach pH. Pancreatic function is characterized by a heterogeneity of cell types. Dysfunction of the pancreas can lead to serious illnesses such as type 1 and 2 diabetes, pancreatitis, and cancer. As such, it remains clinically important to characterize the diversity of pancreatic cell types to understand their role in pancreatic dysfunction.

To address the need of classifying previously known cell types and novel subpopulations, Baron et al. (2016) carried out single-cell transcriptomics of individual pancreatic cells from human donors and mouse models. To characterize the cellular differences between individual cells, the authors implemented the inDrop method which assigns unique RNA barcodes to each cell and analyzes the gene expression profile through droplet microfluidics. Previously characterized cell types were identified along with two additional distinct subpopulations in beta and ductal cells.

In this study, we will attempt to reproduce the findings of Baron et al. To explore the heterogeneity of cell types within a single type of tissue, we will use single-cell RNA sequencing data from a 51-year-old female donor. Analyses will be performed to identify rare and novel subpopulations of cell types, discover marker genes, and determine cell type functions. In particular, our analysis will be mainly directed in using Seurat, an R package used for QC and exploration of single cell RNA-seq data (scRNA-seq). Through this package, we will be able to filter, normalize and detect highly variable features to define clusters by differential expression.

**Data**

Data sourced for this project was found on GEO accession number GSM2230758 from a study published by Baron et al. (2016). Sample was sourced from a female donor, aged 51 with a BMI of 21.1. All samples from the Baron et al (2016) study were obtained from either Prodo or NDRI. Once received, the sample was then recovered in CMRLS at 37°C for 24-48 hours. Cells were encapsulated and barcoded using the inDrop platform and the total RNA was sequenced on Illumina HiSeq 2500.

Quality Control measures were run on the cell barcodes by counting the number of reads by 19 bp barcodes. From the whitelist and the sample data, a UMI matrix was generated.

# **Methods**

To generate the UMI matrix, we used salmon alevin (v1.4.0) on the acquired sample data. To run salmon alevin, a whitelist, transcript ID to gene map file, and an index were created. The whitelist was created by isolating the barcodes from the data and then joining the data from the 3 sample files to create a list of barcodes. The list was then put through a program that sorted by number of instances and then filtered using R (v4.0.2) and saved as a csv. Infrequent barcodes were then filtered out using the mean count.

The map and index used to generate the UMI matrix was created by using the Gencode’s v37 comprehensive gene annotations (CHR regions) and transcript sequences (CHR region) respectively. The map was created using the zless command while the index was created using salmon index. The codes to create map, index, and UMI map were run as qsub scripts using a super computing cluster.

There is precomputed alevin data in the project directory for BF 528. Once imported into R, the package tximport was used. The alevin output file contained information on gene expression levels, gene identification codes, and cell sample barcodes for approximately 30,000 individual cells. The next step required filtering and further downstream analysis for which the package Seurat was used after tximport converted the alevin outputs files into a matrix. The data is also scaled and principal component analysis (PCA) is used for dimensionality reduction. The elbow method is used to identify the optimal number of dimensions and then clustering is performed.

The Seurat object clustered in a previous step was loaded into R and analyzed using the R packages Seurat (Hao 2020, Stuart 2019, Butler 2018, Satija 2015), dplyr (Hadley 2020), and janitor (Firke 2021). Differential expression analysis was run using the Seurat function FindAllMarkers, with the minimum percent of detection set to 25% and the log2 fold change threshold set to 0.25. Only positive (increased) expression was considered for each cluster at this step. Thresholds for minimum percent of detection and fold change were tried at multiple values (0.1, 0.25, 0.3, 0.5). 0.25 provided a balance between minimizing the number of genes identified in multiple clusters and being able to still identify the smaller clusters. For the purpose of downstream analysis, genes with an adjusted p-value less than 0.05 (or 0.01 for GSEA) were considered differentially expressed.

The top expressed genes for each cluster, marker genes used in Baron et al, and marker genes used in Muraro et al 2016 were visualized using a series of violin plots (Figure 6, 7). The five most significant genes from each cluster were used to make a heatmap of log2 fold change (Figure 13). Clusters were visualized using a UMAP, then feature plots were made of gene expression overlaid over the UMAP (Figure 8, 9). These plots and the significant differentially expressed genes from each cluster were used to identify the cell type in each cluster. In addition to the genes associated with each cell type in Baron et al, Muraro et al, and a literature search, databases were used to help identify clusters: GeneCards (Stelzer 2016), The Human Protein Atlas (Uhlen 2015), and PangloaDB (Franzen 2019).

Once cluster identities were determined, cluster names were added using the Seurat function RenameIdents and all figures were remade with the cluster labels added in.

For each cluster, differentially expressed genes were compared with the gene used to define the cluster. If the gene was at least as significant by adjusted p-value, had at least as large of a fold change, and had at least as large of a difference of percent membership, the gene was considered a novel marker for that cluster. Difference of percent membership was calculated by subtracting the positive gene frequency outside of the cluster from the positive gene frequency inside the cluster (Table 1).

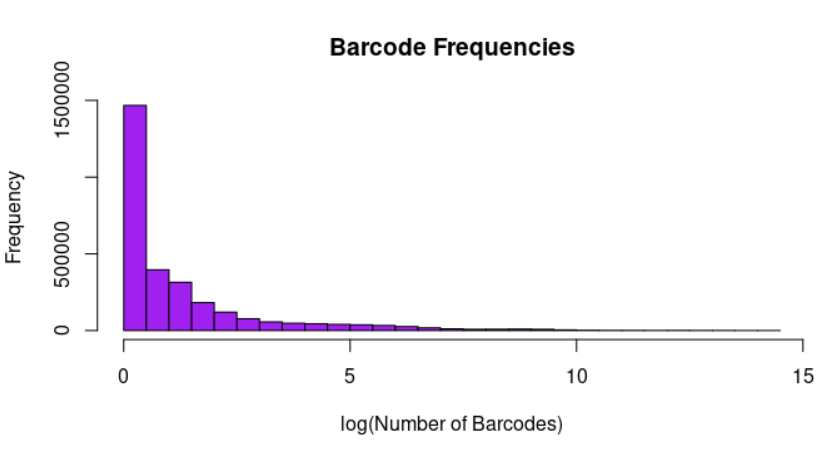
To investigate the difference between subsets present in more than one cluster, a direct comparison was made by re-running differential expression between the clusters of the same cell type, using stricter threshold and considering both increased and decreased expression. For the two alpha cells clusters (Figure 10) and for the two beta clusters (Figure 11), minimum percent membership was set to 50%, log2 fold change threshold was set to 0.5, and the minimum difference between membership percent was set to 0.5 (50%). Both positive and negative changes were considered for this comparison. Violin plots were generated to investigate the statistical significant differential expressed genes.

To investigate the cluster of an unknown cell type(s) differential expression was run again between this cluster and all other clusters with stricter thresholds, with minimum percent membership set to 50%, log2 fold change threshold set to 0.5, and minimum difference in percent membership set to 0.2. Both increased and decreased expression was considered (Figure 12).

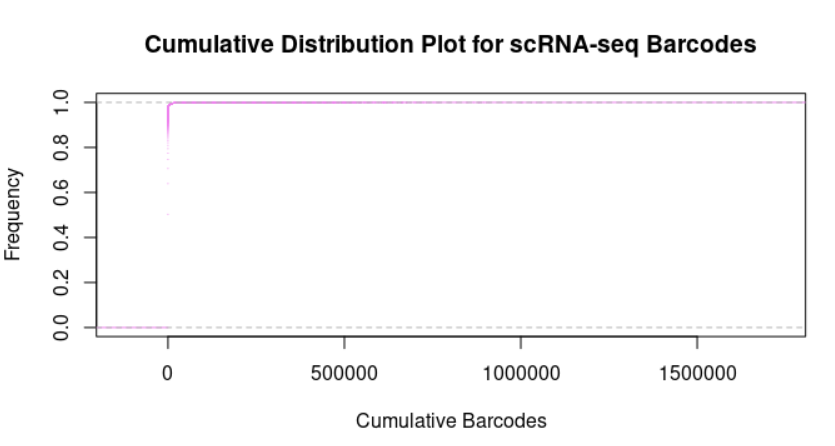
Differential expression analysis and plot generation runs in less than 10 minutes when run on a shared computing cluster.

Enrichment analysis was performed on each cluster to validate their associated biological function using Enrichr (Chen et al., 2013; Kuleshov et al., 2016; Xie et al., 2021). Marker genes were filtered to have positive change in expression with an adjusted p-value less than 0.01 (Table 2).

# **Results**



**Figure 1.** Sequencing Barcode Frequencies Before Filtering. Nearly 3 million barcodes are represented.



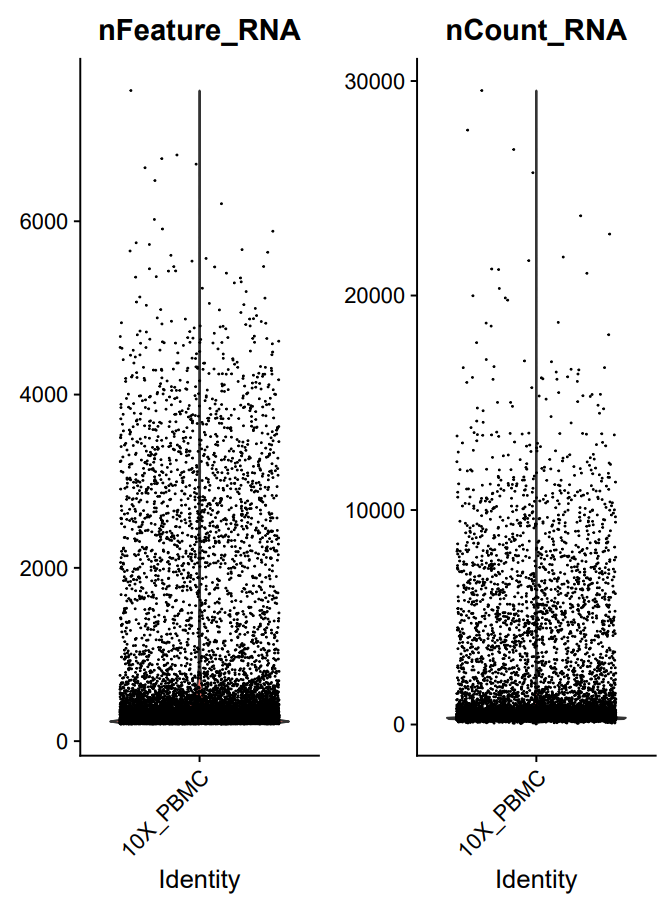
**Figure 2.**Cumulative Distribution Plot.

The creation of the whitelist reduced the number of included barcodes from 2914895 to 100759, which ended up being less than the number of barcodes that were originally included (384^2).

The script that created the map took approximately 1 minute and generated 234484 transcripts to gene ids. The index took approximately 8 minutes to generate and found 143310191 kmers, 1473642 contigs, and reported a sequence length of 187519451

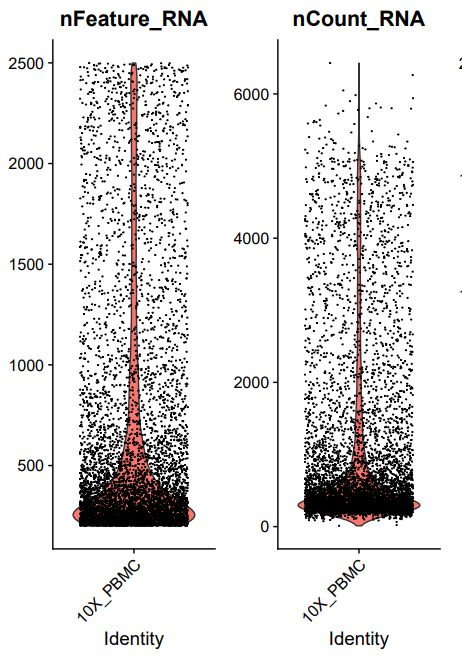
Salmon Alevin took approximately 2 hours and 11 minutes to finish running. The run returned 187519451 UMI after deduplicating. The output had 28456950 bidirected edges and 13937437 unidirected edges while skipping 798 barcodes due to a no mapped read.

From the raw alevin file, generating a feature and cell count from the gene expression levels we see that the total cell count reaches 30,000 while the gene counts per cell reach 7,000. From the article and supplementary information, a feature count that high tends to indicate an increased amount of doublets and therefore filtration was necessary, prior to normalization.



**Figure 3.** Prefiltered plot of alevin output cell and feature count.

Upon filtering using feature count and percent.mt guidelines in R (feature count greater than 200 and less than 2500 and percent.mt less than 20), the new subset was normalized. Using the log normalize function with a scaling factor of 10,000 on the subset, a second filtering step was performed to help remove low variance genes. This post normalization filtering step helps remove cells that may have abnormally high/low expressions of specific genes and do not affect downstream analysis.



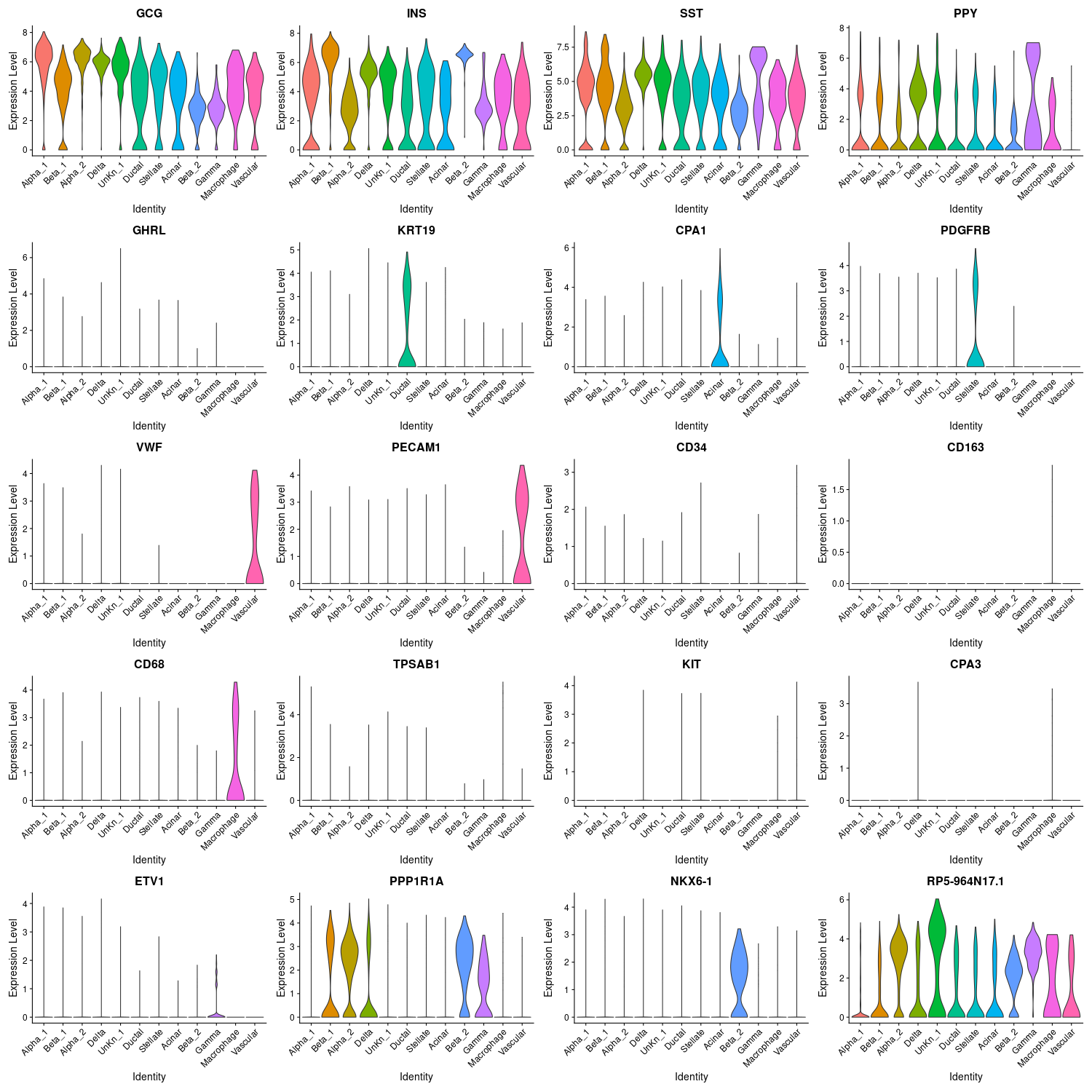
**Figure 4.** Plot of alevin output cell with gene count after filtering.

After the first filtering step, normalization, and variance filtering the total cell count with higher quality resulted in approximately 14,000 cell counts that reduced 12 clusters. Varying the feature count threshold between 200-2500 allowed us to generate 12 clusters

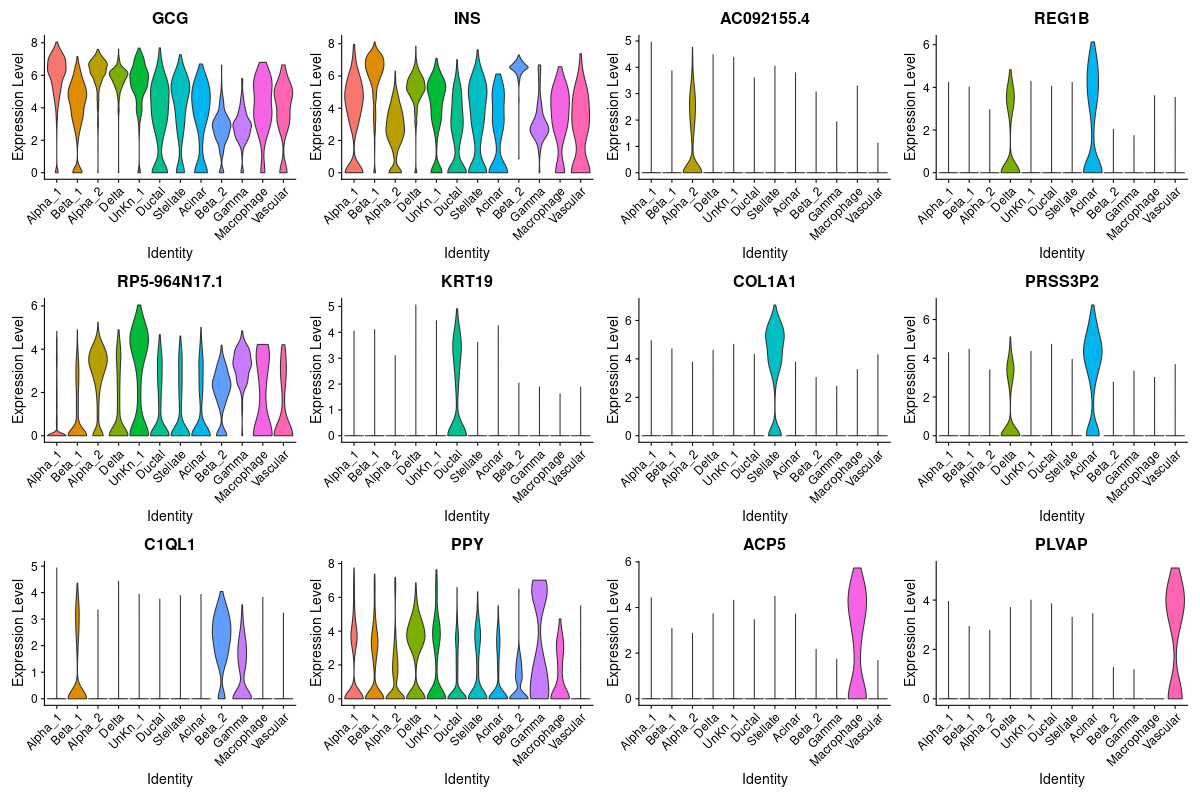
****

**Figure 5.** Clustering performed with UMAP with a B) barplot to compare cell count sizes of each cluster.

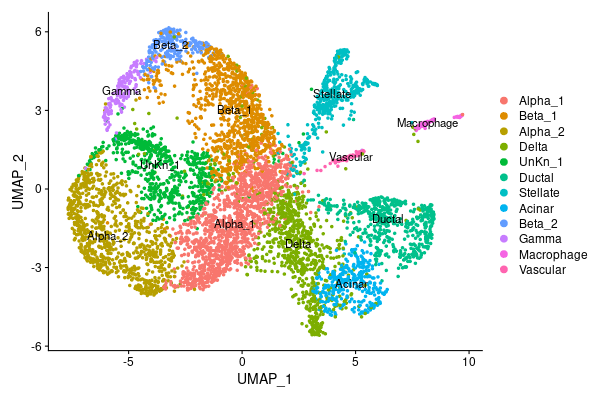
Differential expression between clusters was used to define what cell type each cluster was (Figure 6, Baron 2016, Muraro 2016). Cell subsets were compared with the markers used to define subsets in Baron et al. The two clusters that expressed significantly higher GCG were labeled as alpha subsets, compared with the one cluster in Baron et al. The two clusters that expressed significantly higher INS were considered beta subsets, compared with the one cluster in Baron et al. The clusters that expressed significantly higher SST, PPY, KRT19, CPA1, and PDGFRB were considered delta cells, gamma cells (pancreatic polypeptide cells), ductal cells, acinar, and stellate respectively. The cluster significantly higher for VWR and PECAM1 was considered vascular cells. Unlike Baron et al, neither vascular cells nor any of the other clusters expressed significantly higher CD34. The cluster significantly higher for CD68, we considered macrophage. Unlike Baron et al, neither this “macrophage” cluster, nor any other, expressed significantly more CD163. Unlike Baron et al, we did not have CD3, CD8, nor IgG in this dataset. We were unable to identify clusters associated with epsilon cells (GHRL), cytotoxic T cells (CD3, CD8, TRAC, other TCR genes), and mast cells (TPSAB1, KIT, CPA3).



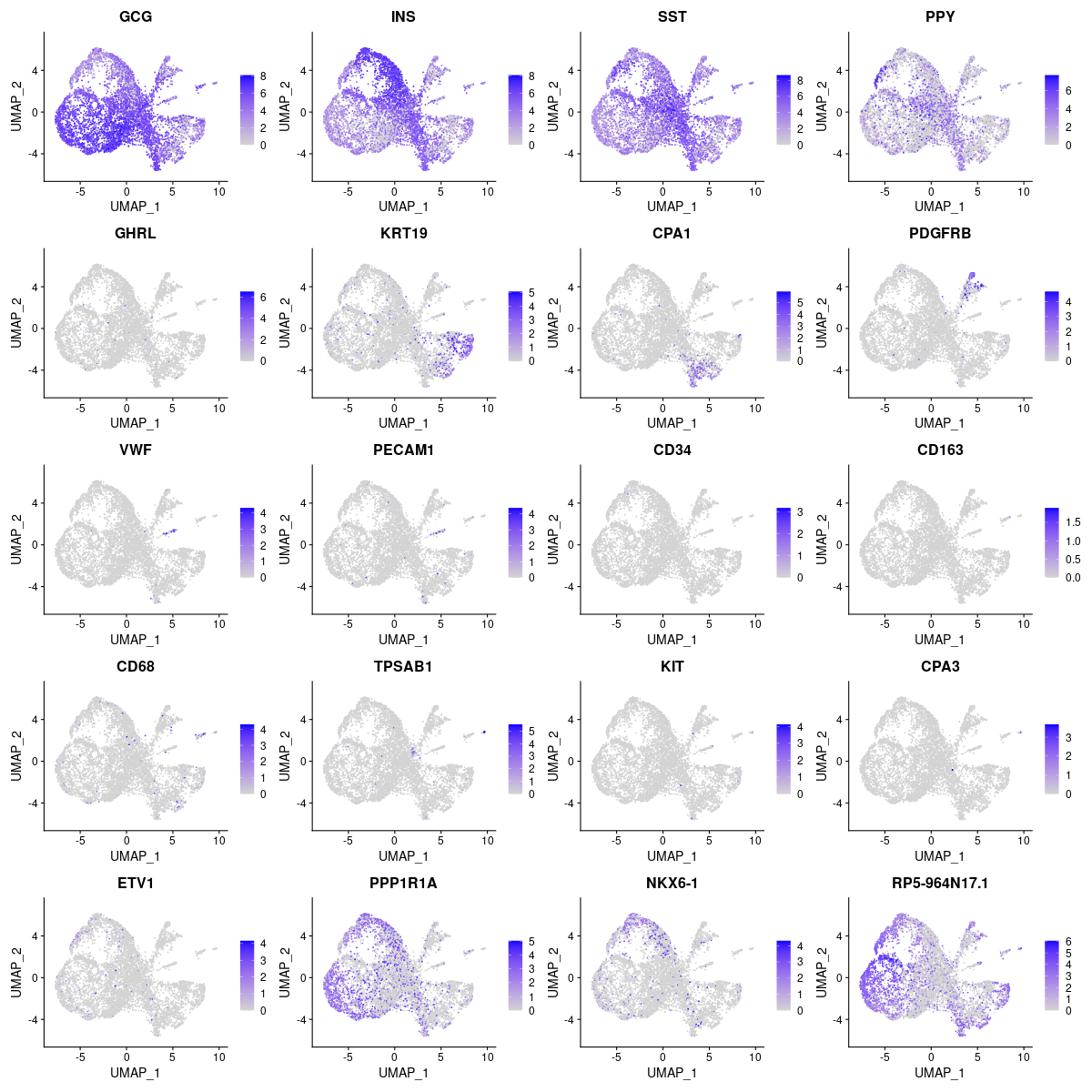
**Figure 6**. Expression of genes of interest across clusters. GCG (alpha), INS (beta), SST (delta), PPY (gamma, pancreatic polypeptide), GHRL (epsilon), KRT19 (ductal), CPA1 (acinar), PDGFRB (stellate), VWF (vascular), PECAM1 (vascular), CD34 (vascular), CD163 (macrophage), CD68 (macrophage), TPSAB1 (mast), KIT (mast), and CPA3 (mast) are all genes used in Baron et al to define cell subsets. ETV1 is another gene associated with gamma (pancreatic polypeptide) cells (Muraro 2016). PPP1R1A is significantly differentially expressed between the two alpha clusters. NKX6-1 is significantly differentially expressed between the two beta clusters. RP5-964N17.1 is the most differentially expressed gene for the unknown cluster UnKn\_1.



**Figure 7**. Expression level of most differentially expressed genes for each cluster, in order.

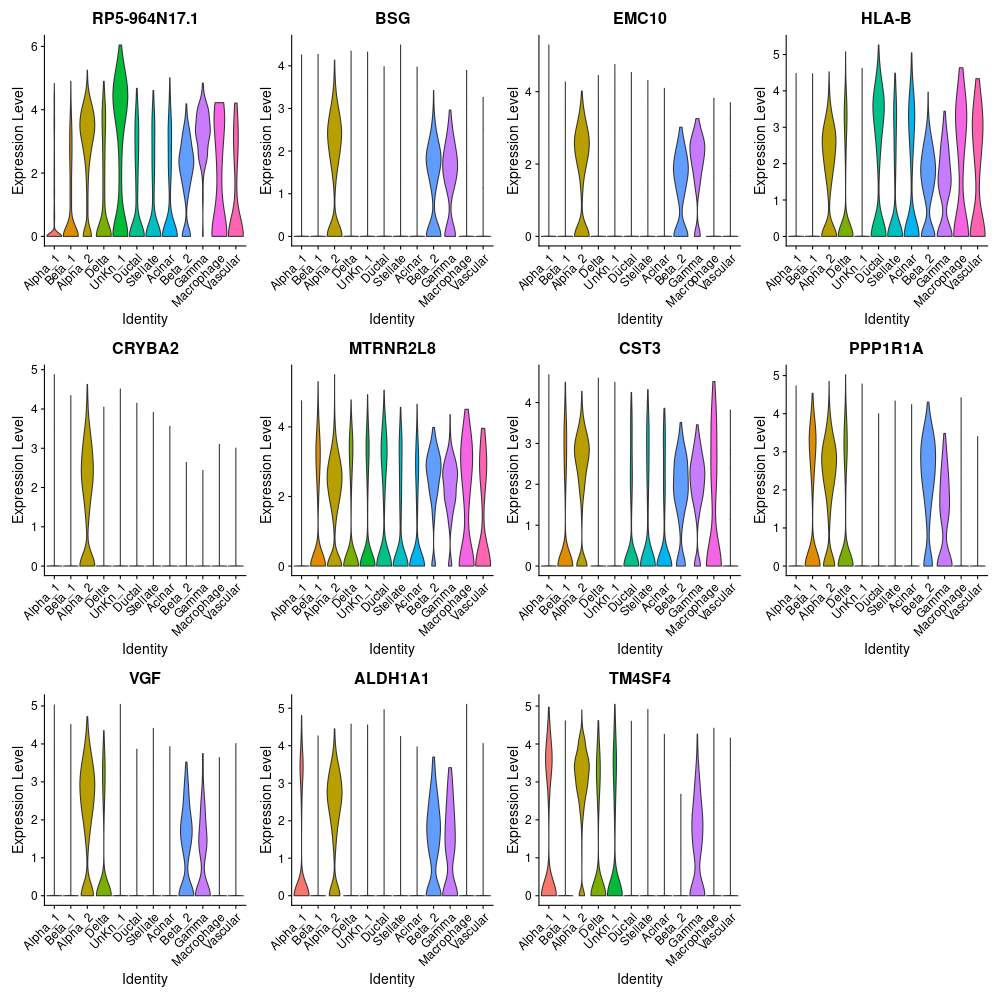


**Figure 8**. UMAP dimensionality reduction representing gene clusters. Cluster names are in the same order as cluster number in Figure 5.



**Figure 9**. Feature plots. Expression of genes of interest overlaid on UMAP.

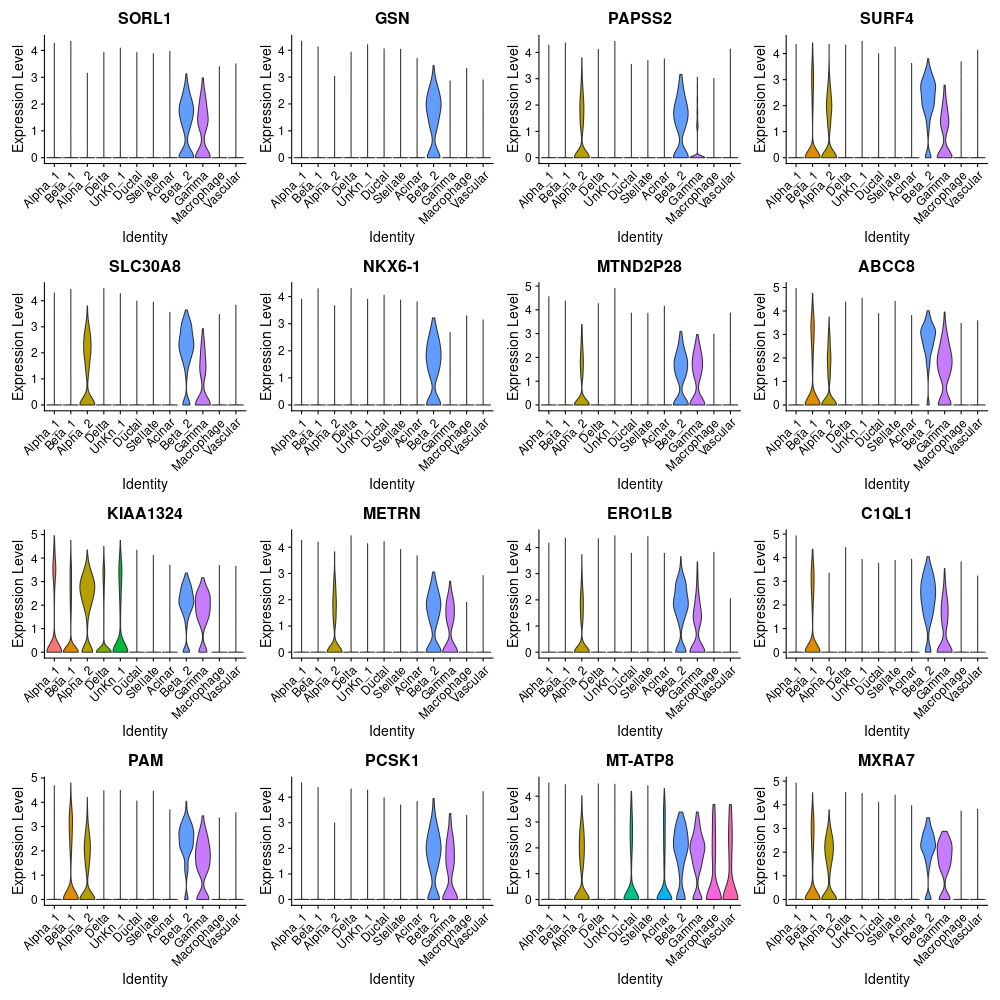
There were two clusters identified as alpha subsets. Alpha\_2 expressed significantly more of each of the 8 genes in Figure 10. Alpha\_2 expressed more CRYBA2, which is a known alpha marker (Muraro 2016). Alpha\_2 also expressed more PP1R1A, which may be important in hormonal control of glycogen metabolism (GeneCards, Stelzer 2016). Alpha\_2 also expresses more ALDH1A1, a protein involved in alcohol metabolism (GeneCards, Stelzer 2016). Perhaps Alpha\_2 is more metabolically active than Alpha\_1. For the alpha cluster Alpha\_1, there were no novel marker genes just as definitive as GCG. Cluster alpha\_2 had 25 genes at least as definitive as GCG (Table 1). Alpha\_2 had novel marker genes more significant than GCG. CRBA2, TTR, TM4SF4, GC, IRX2, and LOXL4 are known alpha markers (Muraro 2016).



**Figure 10**. Expression level, by cluster, of all genes that are significantly differentially expressed between alpha\_1 and alpha\_2 clusters.

There were two clusters identified as beta cells. Cluster Beta\_2 expressed significantly more of each of the genes in Figure 11. Beta\_2 expressed significantly more SLC30A8, a protein only associated with islets of Langerhans in the pancreas, where cells secrete insulin and glucagon (GeneCards Stelzer 2016, Oxford dictionary). Beta\_2 also expressed significantly more NKX6-1, which is required for the development of beta cells in the pancreas (GeneCards, Stelzer 2016). Beta\_2 expressed significantly more Pancreatic Peptidylglycine Alpha-Amidating Monooxygenase (PAM) (GeneCards, Stelzer 2016). Beta\_2 expressed significantly more ABCC8, a gene involved in protein transport across membranes; mutations in this gene are associated with dysregulated insulin secretion (GeneCards, Stelzer 2016). For beta cluster beta\_1, there were no novel marker genes just as definitive as INS. Beta\_2, following a similar pattern as alpha\_2, has 4 genes at least as definitive as INS (Table 1). MAFA, is a known beta marker gene (Muraro 2016).

Delta, stellate, acinar, macrophage, and vascular cells had respectively 4, 7, 13, 8, and 2 novel marker genes just as definitive as the genes used to define each respective cluster (Table 1). Like CPA1, CTRB2 is a known acinar marker gene (Muraro 2016).

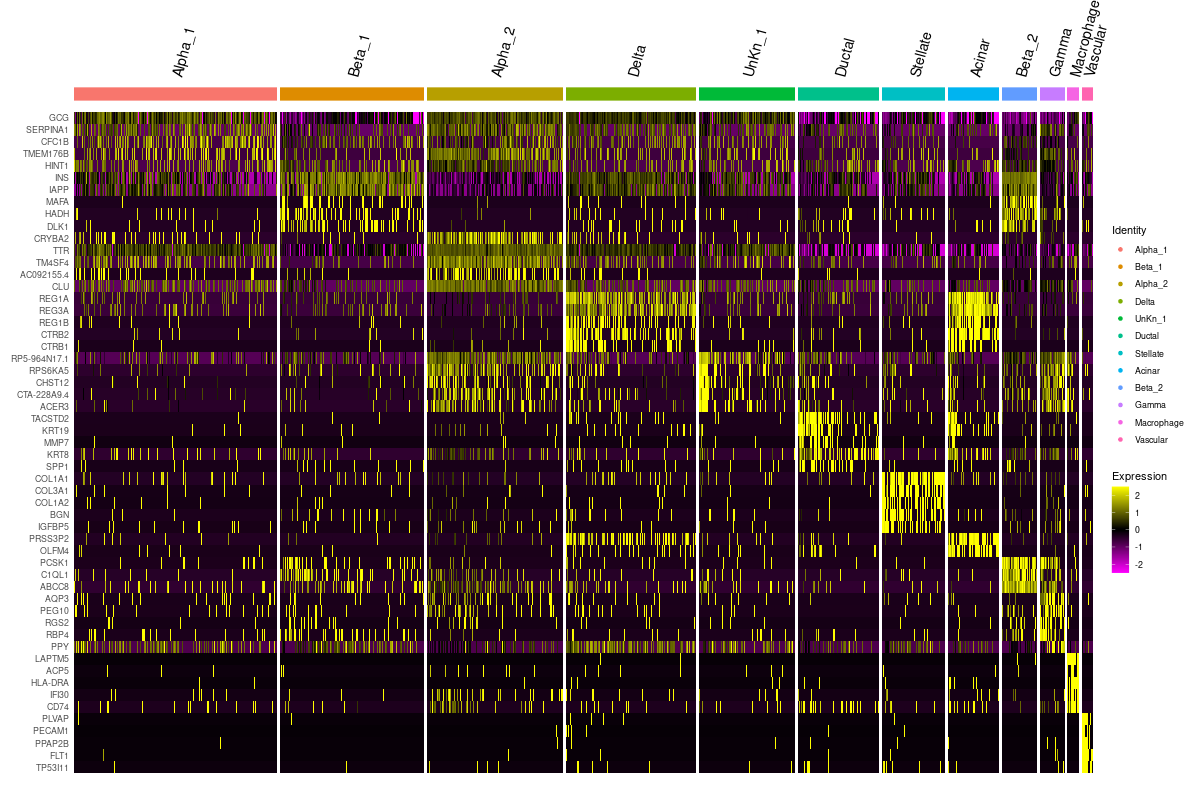


**Figure 11**. Expression level, by cluster, of all genes that are significantly differentially expressed between beta\_1 and beta\_2 clusters.

Unknown cluster UnKn\_1 expressed significantly less of each of the genes in Figure 12, except for RP5-964N17.1, which it expressed significantly more of. RP5-964N17.1 is a LncRNA (Ensembl, Hunt 2018). It expresses significantly less of TMSB4X and TPT1, which are associated with cell proliferation (GeneCards, Stelzer 2016). UnKn\_1 expresses significantly less of the ribosomal encoding proteins RPL10, RPS28, RPL41, RPL19, RPS19, RPS23, RPS12, RPLP2, RPS27A, RPL15, RPS24, RPL13A, RPS2, RPL30, RPLP0, RPS27, and RPS8 (GeneCards, Stelzer 2016). Respectively, FTH1 and FTL encode the heavy and light chain of ferritin, the protein that is the main intracellular storage of iron, involved in iron uptake and release (GeneCards, Stelzer 2016). UnKn1 expresses many of known pancreatic cell type markers, like GCG, INS, SST, and PPY, just at a lower expression level than the clusters associated with those cell types (Figure 6). Perhaps this is a senescent or inactivated cluster of alpha, beta, delta, or gamma cells, or a mix of any of the four.



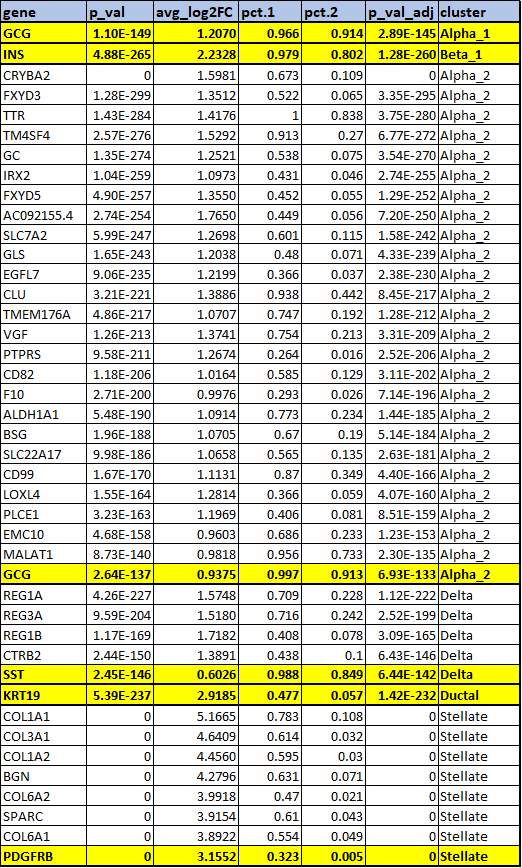
**Figure 12**. Expression level, by cluster, of all genes that are significantly differentially expressed in the unknown cluster UnKn\_1.

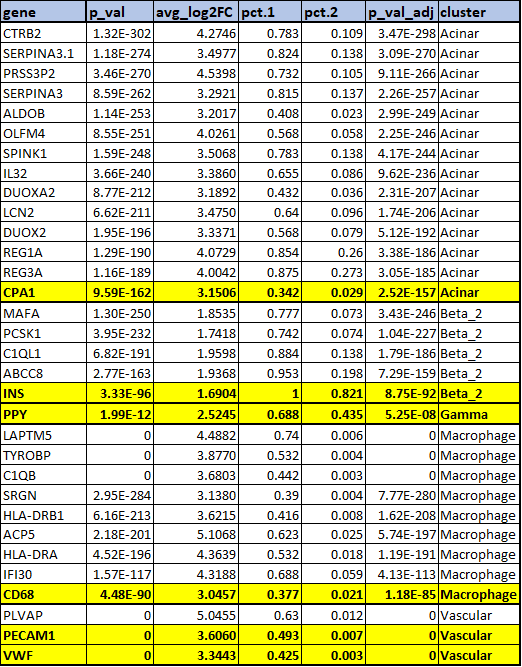


**Figure 13**. Heatmap of normalized log2 fold change of the 5 most significant genes from each cluster.

As seen in Figure 6 and Figure 12, the signal intensity for marker genes for each cluster is not as definite and clean as presented in Baron et al. This could be due to any number of factors, including scaling and normalization.

**Table 1**. Novel marker genes. Marker genes used to define each cluster are highlighted. Novel marker genes are genes that were just as significant as the genes used to define the cluster. Novel marker genes were determined by comparing adjusted p-value, log2 fold change, and difference of fraction of cells detected within the cluster and fraction of cells detected outside of the cluster with the values from the original marker gene (pct.1-pct.2).





**Table 2.** Gene set enrichment analysis on marker genes from each cluster using Enrichr. The following cell clusters were analyzed: endocrine cell types (alpha\_1, alpha\_2, beta\_1, beta\_2, delta, gamma; in green), exocrine cell types (acinar and duct; in blue), immune cell types (macrophage; in red), stellate cells (in yellow), vascular cells (in purple), and an unknown cell type (in gray). Marker genes with an adjusted p-value of less than 0.01 were used to identify cell type functions.



# **Discussion**

Baron et al presented a heatmap and tSNE plot in their Figure 1 that showed that all clusters were clearly distinct, and well defined according to the genes they used to define them. Our heatmap, violin plot showing the most significantly differentially expressed genes from each cluster, and violin plot of marker genes shows a much messier picture of the relationships between cell types, even when focused on a short list of genes as Baron et al presented (Figure 12, 6, 7). The scaling in Baron et al’s heatmap showing either clearly positive or clearly negative GCG expression across clusters does not compare with our GCG results. How we cleaned, processed and normalized the data, how we calculated differentially expressed genes, and the scaling we used may be responsible for some of the discrepancy. Our UMAP does not show the clearly distinct clusters that are shown in Baron et al (Figure 7). We have clearly distinct stellate, vascular, and macrophage clusters, but our other cell types are not so distinctly defined. Our UnKn\_1 cluster appears like it may be an inactivate or senescent cell type. Baron et al did label a quiescent stellate cluster; our UnKn1 inactivated cluster does not express PDGFRB, but does share some marker expression with alpha, beta, delta, and gamma cells.

When comparing our dimensionality reduction plots (Figure 5, 8) with the original paper Figure 1D, you can see that there are two different visualizations used (UMAP vs T-SNE). Therefore, some of the differences in results can be attributed to the differences in upstream analysis such as use of Salmon alignment method vs Bowtie, a change in mapping rates, and the use of different filtering thresholds, which could drastically affect the clustering. It is unclear whether it is possible to change the clustering method in Seurat from nearest neighbor to the one used in the paper (hierarchical clustering).

We had fewer major cell types than Baron et al looked at, and other subsets (such as alpha and beta) separated into subsets. This could have been due to multiple factors. Alpha and beta cells are major cell types in the pancreas, and subsets with different activity levels (or functions) may be expected. Our missing subsets, epsilon, cytotoxic T cells, and mast cells are expected to be smaller frequencies in the pancreas compared with some of the subsets we did find (Baron 2016, Muraro 2016). In addition to the already listed factors, this could be due to the set-up and processing of the data. Unlike the authors, we did not have access to the original barcodes so we were unable to select barcodes and allow for up to 2 mismatches when grouping. Instead, we filtered out the least frequent barcodes, which may have affected some of the smaller subsets as the final barcode count was somewhat smaller than the assumed number of barcodes that the study started with.

We were able to identify 9 cell types from 12 different clusters. This included 4 endocrine cells (alpha, beta, delta, gamma), 2 endocrine cells (acinar and ductal), 1 immune cell (macrophage), stellate cells, and vascular cells (Table 2). Overall, the results of the gene enrichment analysis correspond accurately and provide some support to each clustered cell type. Both clusters of beta cells are enriched by the insulin secretion pathway. Different states of beta cells have been characterized by previous studies; in its regular functional state, beta cells secrete insulin in response to metabolic demand and to regulate levels of glucose while in its dysfunctional state, impaired insulin secretion leads to the onset of diabetes (Cerf, 2013). Alpha, beta, and gamma cells are enriched with ribosome-related terms while beta and gamma are enriched with insulin secretion. Delta cells were enriched with gastric acid and pancreatic secretion, supporting their known somatostatin secretion function. Somatostatin inhibits gastric acid, insulin, and glucagon secretion (Mani et al., 2015). The alpha, beta, gamma, and delta clusters are all pancreatic endocrine cells that share some of the same genes which would perhaps explain the overlap in biological function and enrichment pathways, particularly with their involvement with hormonal regulation and secretion.

Furthermore, pathway enrichment terms also corresponded well with the remaining clusters. For example, the cluster of stellate cells were enriched with focal adhesion and extracellular matrix (ECM) pathways, results that were corroborated by Baron et al. and Buchholz et al., 2005. Activated stellate cells assume a myofibroblast-like phenotype in response to pancreatic inflammation or carcinogenesis by depositing excessive amounts of ECM to replace cellular regions damaged by inflammation or creating a fibrotic matrix optimal for tumor growth and progression, respectively (Ferdek et al., 2017; Phillips, 2012). This strongly suggests that this cluster of stellate cells is of the activated phenotype.

The unknown cluster is enriched with the transcription factor NF-kappa B signaling pathway which has been well characterized for its role in the immune response. This pathway is involved in the development of acute pancreatitis, tumor progression, and pancreatic ductal carcinoma (Rakonczay Z et al., 2008; Pramanik et al., 2018). It suggests that the cells in this unknown cluster are immune cell types. There are considerably fewer marker genes in this cluster relative to other clusters that we characterized which may be a result of upstream filtering, clustering, and labeling steps.

**Conclusion**

Overall, our results were found to be synonymous with those of Baron et al. Most of the cell types from Figure 1 of the original paper were identified using the Seurat R package. We were able to successfully characterize the biological significance of most of the cell clusters. Difficulties encountered in this process was due to the overlap of shared genes amongst different clusters that resulted in similar enrichment terms between different cell types. For each step of this project, from data curator to biologist, it felt that any subjective tweaks (filtering, thresholds, etc.) could lead to some difference in the downstream analysis and biological significance. Testing different nFeature\_RNA, nCount\_RNA, and percent.mt subsets to see how the cluster counts and results differ may help standardize the analysis. More accurate reproducibility may be facilitated with analysis of the entire dataset from the original study. In general, the study provides a clear illustration of the potential to harness the power of single-cell RNA-seq to systematically characterize the transcriptomics of thousands of cells in order to advance understanding dysfunction of disease at a molecular level.

**References**

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

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A.

NCBI GEO: archive for functional genomics data sets--update.

Nucleic Acids Res. 2013 Jan;41(Database issue):D991-5.

Butler et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol (2018) [Seurat V2]

Charlotte Soneson, Michael I. Love, Mark D. Robinson (2015): Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research

[Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A.](https://www.ncbi.nlm.nih.gov/pubmed/23586463)

[Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics. 2013; 128(14)*.](https://www.ncbi.nlm.nih.gov/pubmed/23586463)

Cerf M. E. (2013). Beta cell dysfunction and insulin resistance. *Frontiers in endocrinology*, *4*, 37. <https://doi.org/10.3389/fendo.2013.00037>

Edgar R, Domrachev M, Lash AE.

Gene Expression Omnibus: NCBI gene expression and hybridization array data repository

Nucleic Acids Res. 2002 Jan 1;30(1):207-10

Ferdek, P. E., & Jakubowska, M. A. (2017). Biology of pancreatic stellate cells-more than just pancreatic cancer. *Pflugers Archiv : European journal of physiology*, *469*(9), 1039–1050. <https://doi.org/10.1007/s00424-017-1968-0>

Franzén, O., Gan, L., &amp; Björkegren, J. L. (2019). Panglaodb: A web server for exploration of mouse and human single-cell rna sequencing data. Database, 2019. doi:10.1093/database/baz046

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.1. [https://CRAN.R-project.org/package=dplyr](https://cran.r-project.org/package=dplyr)

Hao and Hao et al. Integrated analysis of multimodal single-cell data. bioRxiv (2020) [Seurat V4]

Hunt, S. E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., Parton, A., Armean, I. M., Trevanion, S. J., Flicek, P., & Cunningham, F. (2018). Ensembl variation resources. 2018, 12.

ISLETS of LANGERHANS: Definition of islets OF LANGERHANS by Oxford dictionary on LEXICO.COM also meaning of islets OF LANGERHANS. (n.d.). Retrieved April 27, 2021, from https://www.lexico.com/definition/islets\_of\_langerhans

[Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A.](https://www.ncbi.nlm.nih.gov/pubmed/27141961)

[Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research. 2016;*](https://www.ncbi.nlm.nih.gov/pubmed/27141961) [*gkw377*](https://doi.org/10.1093/nar/gkw377).

Muraro, M., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelese, M., Carlotti, F., de Koning, E., van Oudenaarden, A. (2016). A single-cell transcriptome atlas of the human pancreas. Cell Systems, 3(4). doi:10.1016/j.cels.2016.09.002

Phillips P. Pancreatic stellate cells and fibrosis. In: Grippo PJ, Munshi HG, editors. Pancreatic Cancer and Tumor Microenvironment. Trivandrum (India): Transworld Research Network; 2012. Chapter 3. Available from: https://www.ncbi.nlm.nih.gov/books/NBK98937/

Pramanik, K. C., Makena, M. R., Bhowmick, K., & Pandey, M. K. (2018). Advancement of NF-κB Signaling Pathway: A Novel Target in Pancreatic Cancer. *International journal of molecular sciences*, *19*(12), 3890. https://doi.org/10.3390/ijms19123890

Rakonczay, Z., Jr, Hegyi, P., Takács, T., McCarroll, J., & Saluja, A. K. (2008). The role of NF-kappaB activation in the pathogenesis of acute pancreatitis. *Gut*, *57*(2), 259–267. https://doi.org/10.1136/gut.2007.124115

Safran, M., Lancet, D. (2016). The GeneCards suite: From gene data mining to disease genome sequence analyses. Current Protocols in Bioinformatics, 54(1). doi:10.1002/cpbi.5

Sam Firke (2021). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.1.0. [https://CRAN.R-project.org/package=janitor](https://cran.r-project.org/package=janitor)

Satija and Farrell et al. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol (2015) [Seurat V1]

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N.,

Stuart and Butler et al. Comprehensive Integration of Single-Cell Data. Cell (2019) [Seurat V3]

Thomas Lin Pedersen (2020). patchwork: The Composer of Plots. R package version 1.1.1. [https://CRAN.R-project.org/package=patchwork](https://cran.r-project.org/package=patchwork)

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., … Pontén, F. (2015). Tissue-based map of the human proteome. Science, 347(6220), 1260419. https://doi.org/10.1126/science.1260419

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

[Xie Z, Bailey A, Kuleshov MV, Clarke DJB., Evangelista JE, Jenkins SL, Lachmann A, Wojciechowicz ML, Kropiwnicki E, Jagodnik KM, Jeon M, & Ma’ayan A.](https://currentprotocols.onlinelibrary.wiley.com/doi/10.1002/cpz1.90)

[Gene set knowledge discovery with Enrichr. *Current Protocols, 1, e90. 2021. doi: 10.1002/cpz1.90*](https://currentprotocols.onlinelibrary.wiley.com/doi/10.1002/cpz1.90)

Zhu, A., Srivastava, A., Ibrahim, J.G., Patro, R., Love, M.I Nonparametric expression analysis using inferential replicate counts Nucleic Acids Research (2019)