Go Ogata as Data Curator
Liz Murphy as Programmer
Haroon Qureshi as Analyst
Meghana Peshwa as Biologist

Single Cell RNA-Seq Analysis of Pancreatic Cells

Introduction:

Gene expression is a powerful tool that allows for the characterization of samples and capturing a sample response to different stimuli. While powerful gene expression is capturing the overall expression of the sample and fails to capture relationships within the sample. A solution to this problem is single cell sequencing, which allows for identification of cell types and proportions of cells within a sample. Baron et al. used this technology to characterize the difference between pancreas of mouse models and in humans. Baron et al. were able to confirm the 15 different cell types that were previously characterized. In addition to confirming that 15 cell types, they were able to find heterogeneity in the regulation of genes related to functional maturation and in the combination of bulk gene expression data, found disease-associated differential expression. This project was performed in the intent to recreate the data and to confirm their findings using different bioinformatic tools

Data:

In Barin et al. , the human tissue was obtained from the Harvard University Faculty of Art and Sciences and was inDrop single cell sequencing. The sequencing was performed using paired end sequencing on Illumina Hiseq 2500 machines. In total, 13 human samples were sequenced using this method. The replication of the data and results were run using samples from a 51 year old female donor. The fasta files associated with the 51 year old female were SRR3879604, SRR3879605, SRR3879606. Samples of the original paper had a complex barcoding scheme and the files used in replicating the experiment were processed and file *_1_bc.fastq.gz was used instead to decide which barcodes would be included in the analysis.

Total Reads	1324837961
Mapping Rate	37.39%
Whitelist barcodes	143399
Mean umis per cell	152

Mean genes per cell	101

Table Salmon Output The salmon alevin showed a rather low mapping rate and shows that the Mean umis per cell across the 3 samples are 152 with the mean genes per cell is 101.

Methods:

Data was QC controlled using trimmomatic and the barcodes of individual samples were read to calculate the number of reads per distinct barcode. The reads per distinct barcode were used to generate a whitelist of barcodes that represent valid cells that contain data that is not relevant. The criteria for whitelisting of the barcode was determined by the filtering out barcodes that were less than 10 as the median of the read counts were 10 for all 3 fasta files. The whitelist generated from the previous step was used for generating the UMI matrix with salmon alevin. In addition to the previously generated whitelist and the 3 pairs of fasta files, a SA salmon index and gtf file of hg38_primary build obtained from refgenie. The gtf file was further processed to obtain a transcript ID to gene map.

The UMI counts matrix was imported using the R packages tximport and fishpond. Quality control for cell and gene-based filtering were performed based on Seurat's Guided Clustering Tutorial[1] and a tutorial from the Harvard Chan Bioinformatics Core[2]. A Seurat object was created with the initial filters of minimum nonzero count of cells per gene set to 3, and minimum nonzero count of genes per cell set to 20. More strict filtering of cells was performed based on the number of genes detected per cell (nFeature_RNA), UMI counts (transcripts) per cell (nCount_RNA), number of genes per UMI (log10GenesPerUMI), and the mitochondrial percentage (percent.mt). Low quality cells, with fewer than 200 or greater than 4000 genes, less than 500 UMIs, mitochondrial percentage greater than 20%, and log10(genes per UMI) less than 0.85 were removed. These filtering parameters account for cells that may not have been sequenced deeply enough, possible doublets, low complexity cells, and dead or dying cells with mitochondrial contamination. Feature scatter plots and violin plots of the Seurat object metadata were used to determine these parameters, as showN in Figures 1 and 2. The counts matrix was further filtered by gene, removing genes with zero expression for all cells, and keeping only those genes expressed in at least 10 cells. These low expression genes can reduce the average expression of cells if not removed.

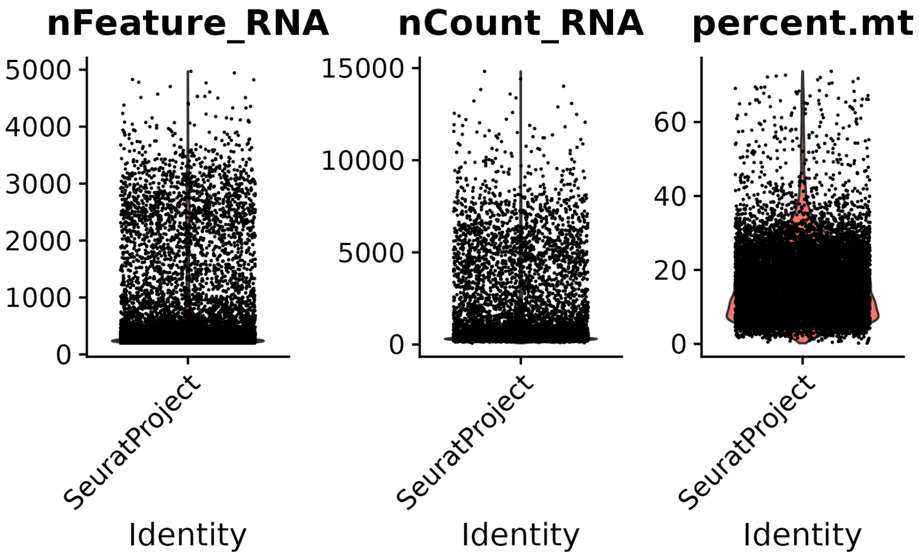


Figure 1. Violin plots displaying the distribution of UMI matrix filtering parameters nFeature_RNA (genes per cell), nCount_RNA (transcripts per cell) and percent.mt (mitochondrial percentage).

The filtered UMI matrix was then normalized using Seurat's NormalizeData function. Feature selection was performed, using Seurat default parameters which selected the 2,000 most variable features. The data were normalized in preparation for linear dimensionality reduction via PCA, using the RunPCA function. An elbow plot, as shown in figure 3, was used to determine the principal components capturing the most variance in the dataset, and 8 dimensions were used for clustering in the next step. The cells were clustered using the Seurat function FindNeighbors, which uses a k-nearest neighbors algorithm. The clusters were then saved to an RDS file.

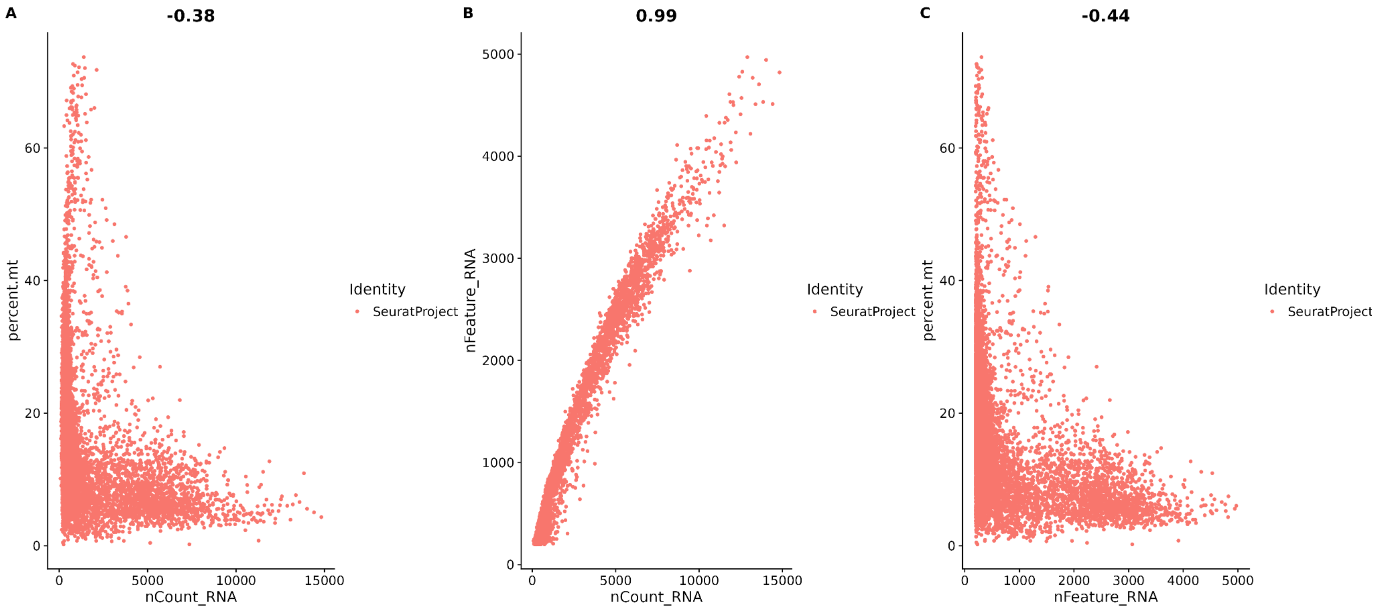


Figure 2. Feature Scatter plots displaying filtering parameters. **A)** percent.mt vs. nCount_RNA, **B)** nFeature_RNA vs.nCount_RNA, **C)** percent.mt vs. nFeature_RNA

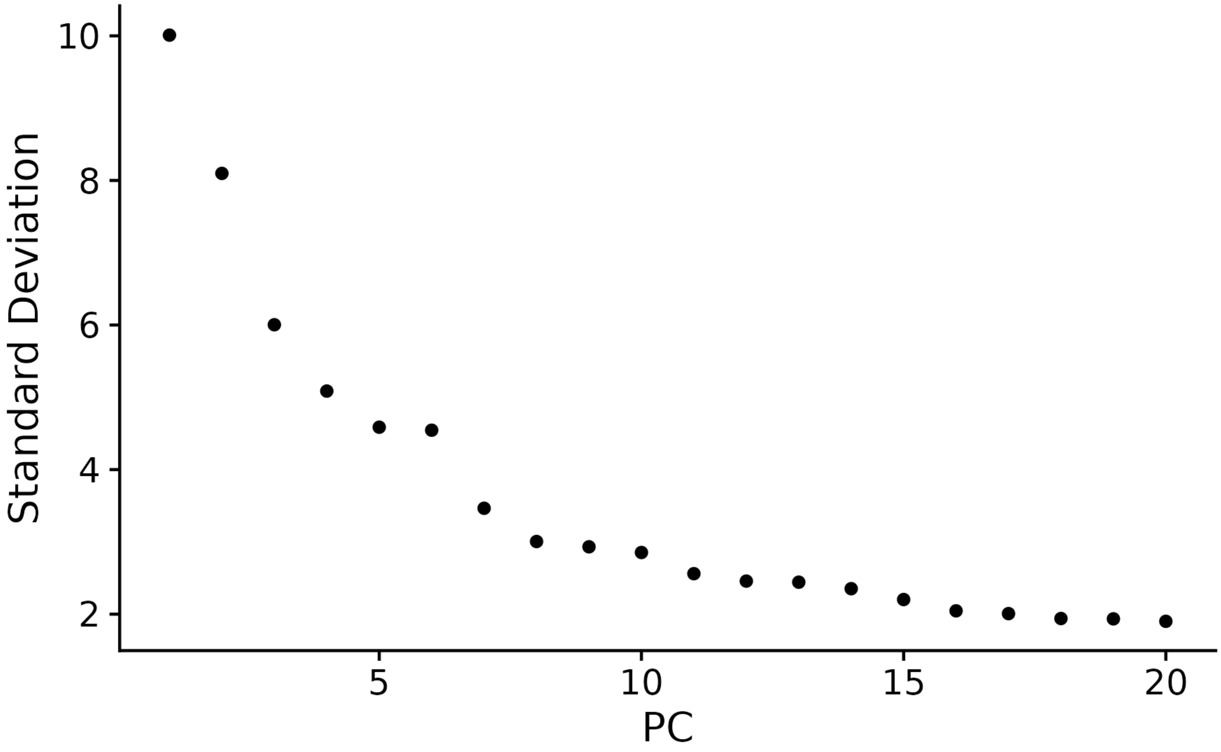
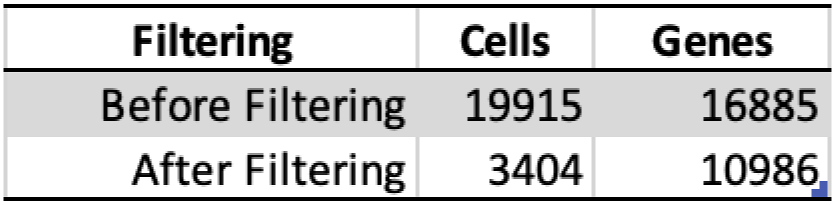


Figure 3. Elbow plot displaying the variance captured by the first 20 principal components after PCA. An “elbow” is visible around PC8.

Using the generated clusters, the subtypes of the cells were determined using the FindAllMarkers() function in the Seurat package. This function utilized the Wilcoxon Rank Sum Test to compare medians between populations to determine differential expression, from which the marker genes for every cluster were identified. A minimum threshold percentage of 0.25 was implemented while the biomarkers were generated, which allowed for more reliable evidence of either of two groups to be considered significant, which in turn allowed marker genes to be assigned to clusters matching their cell type. PanglaoDB, a database of RNA-sequenced samples from humans and mice, acted as a resource for the correspondence between a gene expression marker and its cell type. The top ten marker genes from each cluster were found through reference against PanglaoDB, as well as analysis of their log2FC, and these genes were indicative of the top up and down regulated genes for each cluster. Using the Seurat package, FeaturePlots, heat maps, and UMAP visualizations of the clusterings were produced, and the cell types were determined using inference of these plots and the expression levels of the marker genes as indicated.

Results:

Table 1. Cell and Gene Counts before and after filtering



Filtering	Cells	Genes
Before Filtering	19915	16885
After Filtering	3404	10986

Before filtering, there were 19,915 cells and 16,885 genes present in the UMI matrix (Table 1). After filtering as presented in the methods, there were 3,404 cells and 10,986 genes ready for further analysis. The pie chart in figure 7 shows the initial results from the clustering analysis, in which 12 cell clusters were identified.

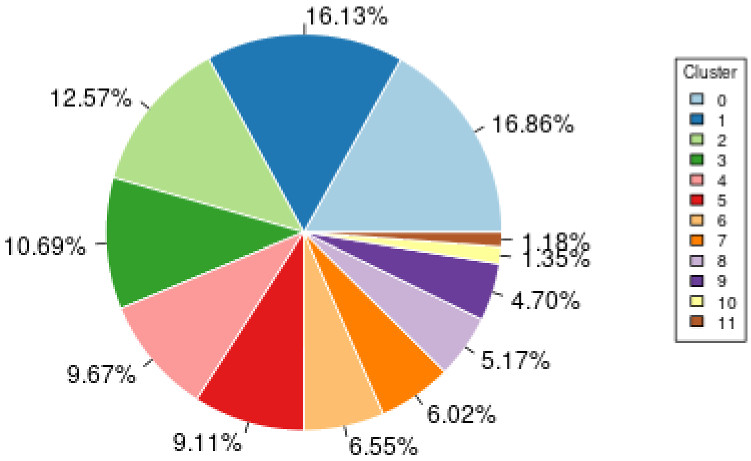


Figure 4. Pie chart displaying the percentage of cells assigned to each of 12 clusters following k-nearest neighbors clustering.

There were 6,619 genes found to be markers using the FindAllMarkers() function in Seurat. These genes were found to reside in twelve clusters, respectively, and a heatmap was created to further visualize this distribution, as seen in Figure X.

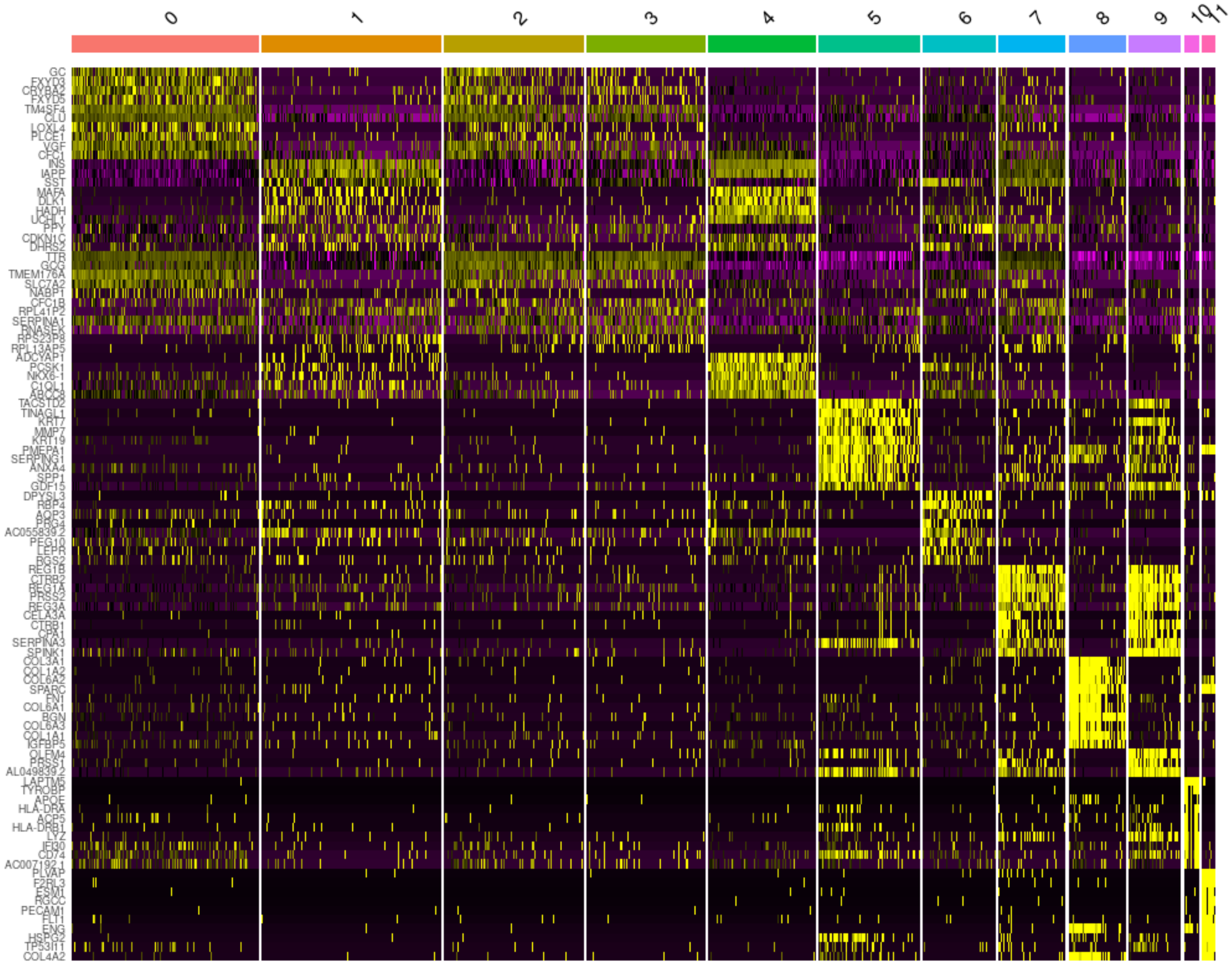


Figure X. Clustered heatmap generated displaying the top 10 differentially expressed marker genes within each cluster. Created using data regarding log normalized UMI counts previously found.

Figure X assisted in determining cell types further downstream. This heatmap visualized the bands created from differentially expressed genes between each cluster. Some of the clusters had overlapping areas of expression, for example, clusters 7 and 9 shared similar areas of expression. Using this to infer cell types, it could be determined that these clusters most likely had similar cell types, despite having different groupings of genes that were previously formed. To further assist in cell typing, the original list of marker genes used in Baron et al. was downloaded from the Supplementary section of the paper.

Next, using the FeaturePlot() and RunUMAP() functions in the Seurat package, a plot of each kind was made. The UMAP projection was made first, in which the clusters that had been found were made easily identifiable through a color legend. This allowed for the visualization of the clusters distinctly. The FeaturePlot that was then created using the UMAP projection allowed for visualization of individual marker gene expression among the separate clusters. The UMAP of all clusters, alongside the FeaturePlots of KRT19, is shown in Figure XI.

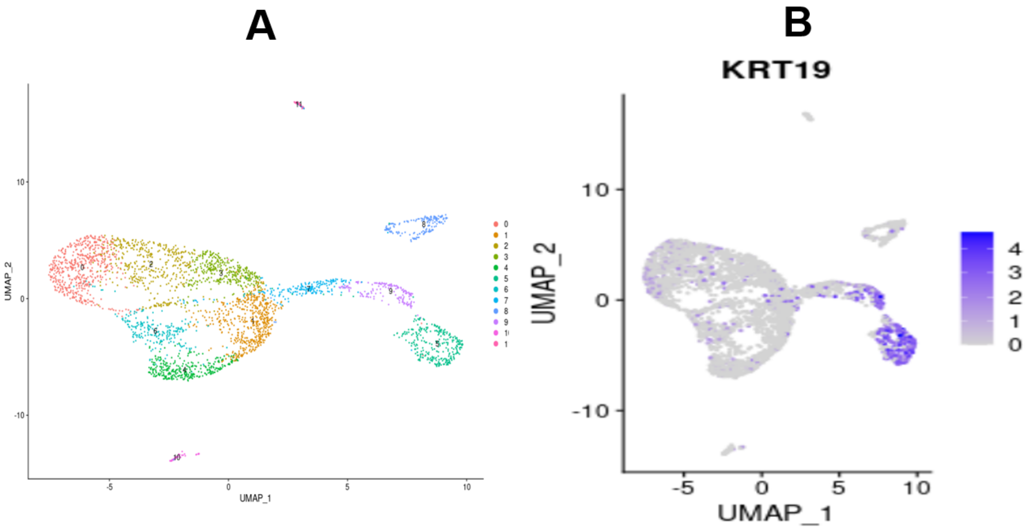


Figure XI. The UMAP projection of all clusters **(A)** and the FeaturePlot of the KRT19 Marker Gene. **(B)**

As seen in Figure XIB, the FeaturePlot of the KRT19 marker gene had higher expression mostly concentrated in the areas corresponding to clusters 5 and 9, as seen in the UMAP projection in Figure XA. This also correlates with the heatmap in Figure X, where there is overlap between the banded areas in cluster 5 and 9, indicating once again at a similarity between the cell types in these clusters, and that these cell types express similar genes.

Using the marker genes from Baron et al, violin plots were made to find expression levels of certain genes in their respective cluster, further narrowing the cell types. The violin plots, as seen below in Figure XII, allowed for further accurate identification of the cluster in which some respective marker genes were highly expressed. As seen in Figure XII, the marker gene KRT19 again produced high expression in clusters 5 and 9

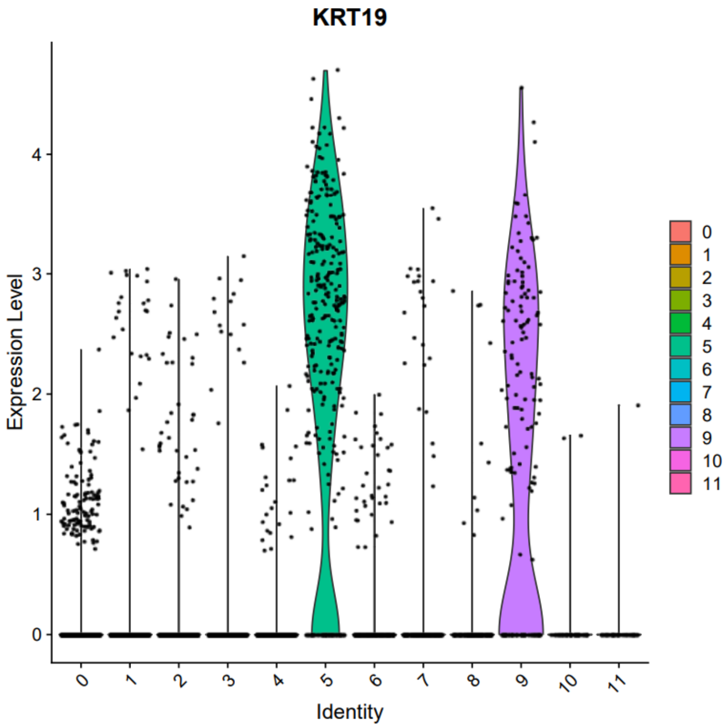


Figure XII. The violin plot produced for expression of KRT19. High expression in clusters 5 and 9.

There were, however, certain cell types that were more difficult to assess, as they spanned across multiple clusters that had many other genes also associated with them. For example, cluster 1 had a multitude of genes sparsely expressed, instead of the strong bands of visible expression seen in Figure X. For this reason, another technique was utilized to determine the cell type. PanglaoDB, a useful database of RN-sequencing data for human and mice samples, was used to further narrow the cell types. The top 10 marker genes expressed in cluster 1 were analyzed, and it was found that two were found to be well represented in alpha cells in PanglaoDB. The SST marker gene, although widely distributed among the clusters, had the highest expression in cluster 1, as seen in Figure XIII, thus determining that this cluster was comprised of alpha pancreatic cells.

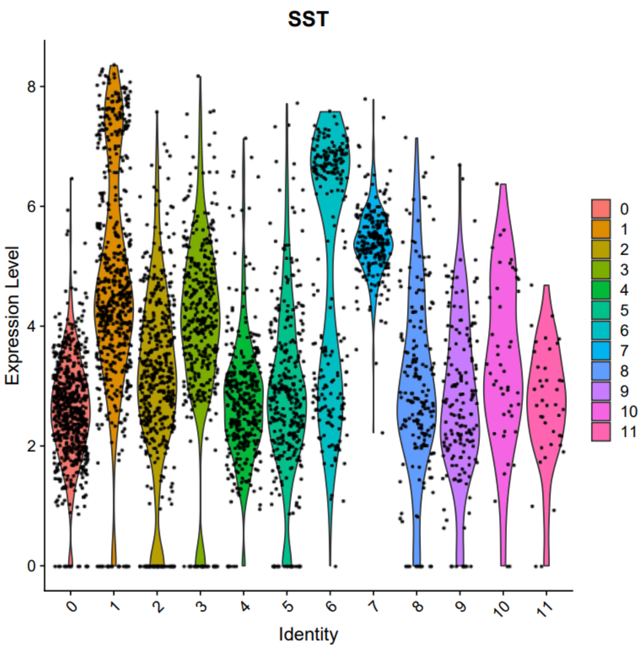


Figure XIII. Violin plot for SST, indicating significant expression in cluster 1.

Some cell types, such as the Cytotoxic T-cells which express CD8, CD3, and TRAC, respectively, were not found to be present in the sample. This was also true of the Mast and Epsilon cell types. The cell types that were identified were further validated using the already produced violin plots, UMAP projection, and FeaturePlots. A new UMAP projection was produced with the cell types indicated.

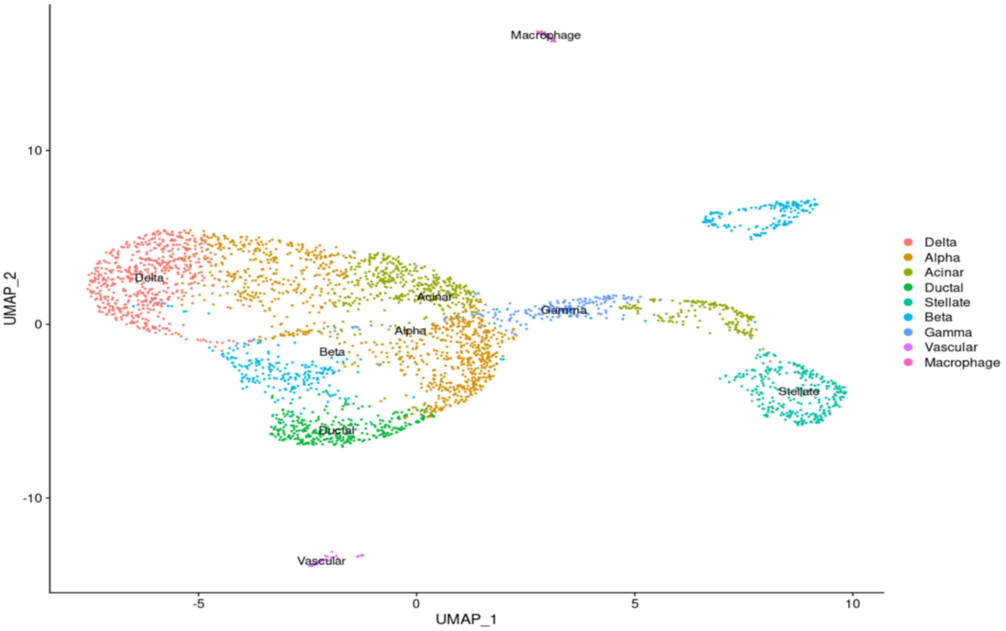


Figure XIV. UMAP projection indicating the cell types associated with each cluster.

To ensure the cell clusterings and typing produced desirable results, a heatmap was produced with the newly associated cell types. The highly enriched genes in each cluster were again indicated by bands across the respective cluster. This heatmap can be seen in Figure XV.

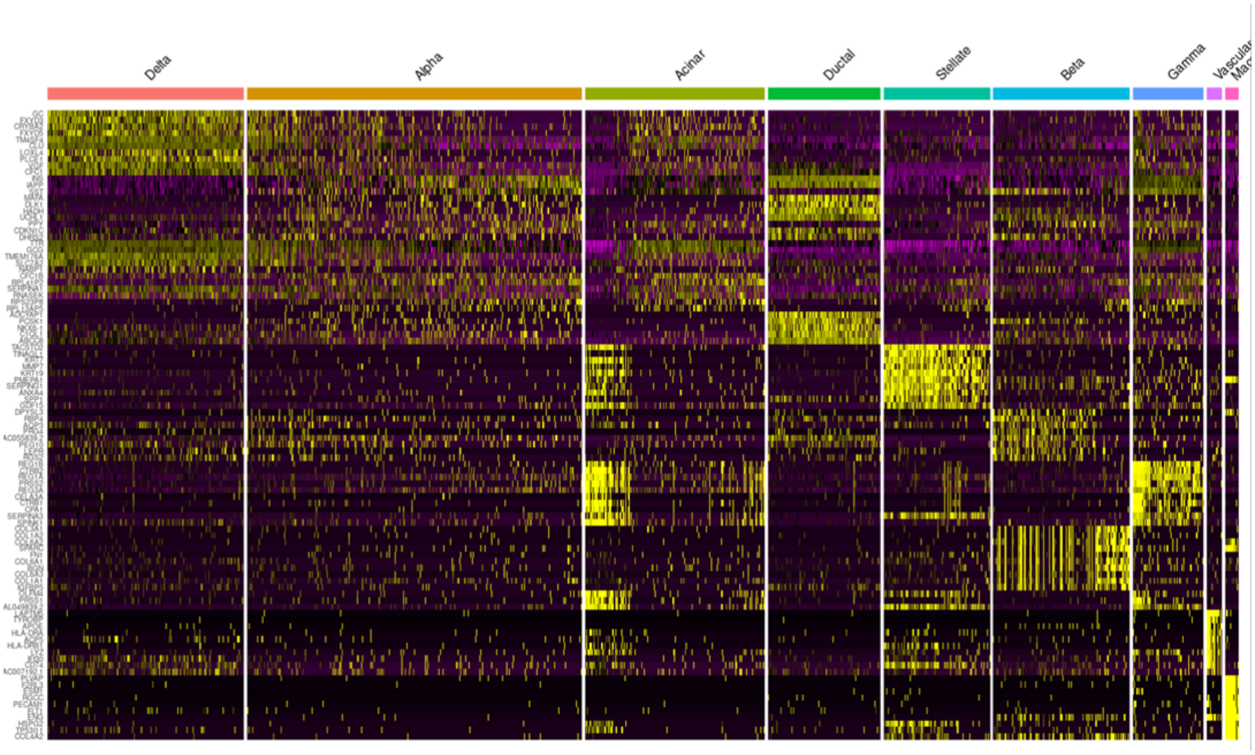


Figure XV. Heatmap of the ten most significantly enriched gene across the clusters, now indicating cell type.

Table 1. Gene set enrichment analysis of the marker gene clusters

		No. of marker genes	**GO Terms**		
Cluster	**Label**		**Biological process**	**Molecular function**	**Cellular component**
0	Delta cells	980	supramolecular fiber organization (GO:0097435)	nucleoside diphosphate kinase activity (GO:0004550)	lysosome (GO:0005764)
1	Alpha cells	141	SRP-dependent cotranslational protein targeting to membrane (GO:0006614)	RNA binding (GO:0003723)	cytosolic large ribosomal subunit (GO:0022625)

2	Acinar cells	214	carbohydrate homeostasis (GO:0033500)	amyloid-beta binding (GO:0001540)	endoplasmic reticulum lumen (GO:0005788)
3	Ductal cells	122	aerobic electron transport chain (GO:0019646)	oxidoreduction-driven active transmembrane transporter activity (GO:0015453)	organelle inner membrane (GO:0019866)
4	Stellate cells	573	protein targeting to ER (GO:0045047)	pyrophosphate hydrolysis-driven proton transmembrane transporter activity (GO:0009678)	endoplasmic reticulum membrane (GO:0005789)
5	Beta cells	783	neutrophil degranulation (GO:0043312)	cadherin binding (GO:0045296)	focal adhesion (GO:0005925)
6	Gamma cells	608	cytoplasmic microtubule organization (GO:0031122)	ATP-activated inward rectifier potassium channel activity (GO:0015272)	trans-Golgi network (GO:0005802)
7	Vascular cells	61	proteolysis (GO:0006508)	peptidoglycan binding (GO:0042834)	secretory granule lumen (GO:0034774)
8	Macrophage	843	extracellular matrix organization (GO:0030198)	cadherin binding (GO:0045296)	focal adhesion (GO:0005925)
9	Unknown	867	neutrophil mediated immunity (GO:0002446)	cadherin binding (GO:0045296)	focal adhesion (GO:0005925)
10	Unknown	376	neutrophil activation involved in immune response (GO:0002283)	cadherin binding (GO:0045296)	focal adhesion (GO:0005925)
11	Unknown	1051	regulation of cell migration (GO:0030334)	cadherin binding (GO:0045296)	cell-substrate junction (GO:0030055)

Using the list of marker genes, we performed gene set enrichment analysis using Enrichr. Top GO terms for molecular function,cellular component, and biological process were used to explain the function of each of the cell type clusters.

Discussion:

There were a few differences found between our reproductive study, and the original paper by Baron et al. Firstly, Baron et al. used tSNE, while this study used UMAP to generate clusterings, which allows for better visualization of distance between clusters. UMAP is better for this purpose than tSNE, which produces more inconsistent distances. This could have been a reason for any differences that occurred when determining cell clustering. It also discourages direct comparison of the plots, as the difference in dimensionality of the two functions can cause misguided inferences.

When comparing the heatmaps between Baron et al. and our reproduction, it can be noted that the same marker genes were used with the intention of increasing accuracy between results. However, there were genes not found to be represented in our final heatmap. These would be the genes that were expressed typically by cytotoxic T-cells, Epsilon, and Mast cell types. When viewing the heatmap generated by our study, there are distinctly marked banding for a variety of cell types, such as those that expressed marker genes KRT19 and CPA1. However, some of the cell types, like alpha cells which expressed marker genes such as SST, were not so distinct in their bandnig and required a handful of techniques to determine their cell type. The difference in these two studies in terms of the cell types represented could be caused due to a limitation of the samples available for reproduction. In this study, nine separate and distinct pancreatic cell types were identified, as opposed to the fourteen identified by Baron et al. These differences could have been due to upstream analysis, as well the difference in visualization techniques used between the studies.

Finally, gene set enrichment analysis helped us find the functions for each cluster. Also, using PanglaoDB Augmented 2021 in enrichr, we were able to ascertain the cell types labeled as unknown for cluster 9,10, and 11. These were found to be acinar, macrophages, and endothelial cells respectively. Carbohydrate homeostasis (GO:0033500) is associated with acinar cells in cluster 2 and this fits since pancreatic enzymes are associated with glucose homeostasis. It was interesting to see that amyloid-beta binding (GO:0001540) was associated with this cluster as well. Miklossy et al. talks about beta amyloid deposits in the pancreas in type 2 diabetes and maybe this GO term could be related to that. The GO terms for cluster 3 ductal cells makes sense since ductal cells are involved in transport of enzymes to the digestive tract. Rest of the enrichment analysis results match the respective cell type labels too.

Conclusion:

To conclude, we found 12 marker gene clusters and successfully labeled 9 of them based on cell type. Baron et al. reported 15 clusters and this difference could be due to the subset of data that we analyzed. There were also differences in upstream analysis and methodology for generating and labeling clusters between the original paper and our study. But despite this, we were able to reproduce the results and overall, our findings were similar to those reported in Baron et al.

References:

1. Seurat – Guided Clustering Tutorial (n.d). https://satijalab.org/seurat/archive/v3.1/pbmc3k_tutorial.html
2. Introduction to Single-cell RNA-seq - ARCHIVED. Harvard Chan Bioinformatics Core (n.d). https://hbctraining.github.io/scRNA-seq/lessons/04_SC_quality_control.html
3. Miklossy J, Qing H, Radenovic A, Kis A, Vileno B, Làszló F, Miller L, Martins RN, Waeber G, Mooser V, Bosman F, Khalili K, Darbinian N, McGeer PL. Beta amyloid and hyperphosphorylated tau deposits in the pancreas in type 2 diabetes. Neurobiol Aging. 2010 Sep;31(9):1503-15. doi: 10.1016/j.neurobiolaging.2008.08.019. Epub 2008 Oct 23. PMID: 18950899; PMCID: PMC4140193.