BF528 Project 4

Single Cell RNA-Seq Analysis of Pancreatic Cells

Marina Natividad Avila(Data curator), Vishwa Talati(Programmer), Brad Fortunato(Analyst), Kyrah Kotary(Biologist)

Abstract

The pancreas' crucial role in energy homeostasis leads to the dysfunction of it to be of clinical relevance and with a wide range of conditions involving it. Its unique construction where the 5% of the cells responsible for hormone secretion live among the other 95%. This leads to conventional gene expression detection methods to yield inaccurate results. The present work uses a novel scRNA approach to cluster cells and do differentially expressed analysis on them.

Introduction

Single cell RNA seq is an approach of isolating single cells, capturing their transcripts, and generating sequencing libraries in which the transcripts are mapped to individual cells which allows assessment of fundamental biological properties of cell populations and biological systems at unprecedented resolution. The pancreas is a vertebrate-specific organ with a central role in energy homeostasis achieved by secreting digestive enzymes and metabolic hormones and its dysfunction is most notable in type 1 (T1D) and type 2 diabetes mellitus (T2D), pancreatitis, and cancer. Significant efforts have been made to replace beta cells and to produce insulin-secreting beta cells in vitro from pluripotent cells. However, these efforts are limited by an incomplete understanding of the gene expression landscape of adult beta cells. Thus, Baron *et al.* (2016) focused on pancreatic cells where a droplet-based, single-cell RNA-seq method was used to determine the transcriptomes of over 12,000 individual pancreatic cells from four human donors and two mouse strains. The inDrop method provides a systematic approach for capturing thousands of cells without pre-sorting and uses high-throughput droplet microfluidics that barcode the RNA from thousands of individual cells, implementing a sensitive linear amplification method for single-cell transcriptome library construction. Cells were divided into 15 clusters that matched previously characterized cell types. Detailed analysis of each population separately revealed subpopulations within the ductal population, modes of activation of stellate cells, and heterogeneity in the stress among beta cells.

The present project attempts to replicate the findings of the paper for a single human individual using current methodology and packages. The goals include processing the barcode reads of single cell RNAseq dataset, performing cell-by-gene quantification of UMI counts, performing quality control of UMI counts matrix, analyzing the UMI counts to identify clusters and marker genes for distinct cell type populations and finally providing a biological meaning to the clustered cell types followed by identification of novel marker genes associated with them.

Methods

Data Preprocessing

Samples SRR3879604, SRR3879605 and SRR3879606 obtained by Baron *et al.* (2016) were retrieved from NCBI GEO: GSE84133 and correspond to a 51 year-old female patient with a BMI of 21.1. The barcode fastq.gz files had already had the adaptor sequence removed. For each sample, the number of reads per distinct barcode were calculated. Code is available at https://github.com/BF528/project-4-project-2-glass-bottom-boat/blob/master/Barcodecounter.py. The logic behind the barcode whitelist selection; the barcodes considered to have sufficient counts to be included in the analysis, came from Melsted *et al.* (2021). Barcode counts were rearranged in ascending order and an empirical cumulative distribution function (ecdf) was plotted (Figure X). Salmon (Patro *et al.* 2017) was used to create a reference index from the ENSEMBL reference transcriptome 37 (Frankish *et al.*, 2018). It was then estimated where the first derivative of the empirical function d/dx=0, which represents a change in the behavior of the graph. Counts equal or bigger than the thresholds were selected and the whitelists for the three samples were merged. Reads were aligned to the index. UMI (unique molecular identifier) count matrix was later generated using the alevin tool from the salmon software.

Creating Seurat Object

To analyze the UMI counties matrix created by alevin, the tximport() function of 'Tximport' package (version 1.16.1) was used to return UMI counts matrix where the values represent the number of molecules for each feature (i.e. gene; row) that are detected in each cell (column). The CreateSeuratObject() function from the 'Seurat' package (version 4.0.1) was used to create object which served as a container for data and analysis of single-cell dataset.

Mapping gene symbols and filtering low quality genes

The UMI matrix generated had Ensembl Gene Identifiers as row names which were mapped to gene symbols using 'EnsDb.Hsapiens.v79' package (version 2.99.0) . For preprocessing, the percentage of counts originating from a set of mitochondrial genes (MT-) was calculated using the PercentageFeatureSet() function. Visualization of the QC matrix was done using the Violin Plot (**Fig. 2**) for nFeature_RNA, nCount_RNA and percent.mt. Next, cells that have unique feature counts over 2,500 or less than 200 and >5% mitochondrial counts were filtered. To visualize the feature-feature relationships, the FeatureScatter() function was used, providing two distinct scatter plots that can be shown in **Fig. 3**.

Normalization and Feature Selection

After pre-processing (removing unwanted cells from the dataset), normalization was done "LogNormalize" method using NormalizeData() function that normalizes the feature expression measurements for each cell by the total expression, multiplies this by a scale factor (10,000 by default), and log-transforms the result. This is followed by calculating the subset of features that exhibit high cell-to-cell variation in the dataset using FindVariableFeatures() function which

returns 2000 features per dataset. ‘Vst’ method was used for this purpose that first fits a curve to predict the variance of each gene as a function of its mean, by calculating the local fitting of polynomials of degree 2 and then standardizes the feature values by subtracting the observed mean and expected variance followed by dividing the expected standard deviation of a feature from the fitting. This method accounts for the mean variance relationship that is inherent to single cell RNA-seq and thus more reliable.

Scaling the Data and Linear Dimension Reduction

Next, linear transformation (‘scaling’) that is a standard pre-processing step prior to dimensional reduction techniques like PCA was done using the ScaleData() function which shifts the expression of each gene, so that the mean expression across cells is 0 and scales the expression of each gene, leading to a variance across cells of 1. This step gives equal weight in downstream analyses, so that highly-expressed genes do not dominate. PCA is thereafter performed on scaled data using RunPCA() function which is visualized using different ways like DimPlot, VimDimLoadings (**Fig. 5**), among others.

Dimensionality Determination

A resampling test inspired by the JackStraw procedure was used to determine dimensionality which approximately took 11 minutes to run where we randomly permuted a subset of the data (1% by default) and rerun PCA, constructing a ‘null distribution’ of feature scores, and repeated this procedure. We thus identified ‘significant’ PCs as those who have a strong enrichment of low p-value features. The JackStrawPlot() function **(Fig. 6**) was used as a visualization tool for comparing the distribution of p-values for each PC with a uniform distribution (dashed line). Another alternative for this was the Elbow plot that was visualized using the ElbowPlot() function (**Fig. 7**) that ranked principle components based on the percentage of variance explained by each one.

Cell Clustering

For clustering the cells, a KNN graph was constructed based on the euclidean distance in PCA space, and refined the edge weights between any two cells based on the shared overlap in their local neighborhoods (Jaccard similarity). This step was performed using the FindNeighbors() function, and taking the previously defined dimensionality of the dataset (first 10 PCs) as input. Next, modularity optimization techniques such as the Louvain algorithm (default) or SLM were applied to iteratively cluster cells together, with the goal of optimizing the standard modularity function using the FindClusters() function with 0.5 resolution.

Non-Linear Dimension Reduction

To explore the dataset, nonlinear dimensionality reduction techniques, such as tSNE and UMAP of Seurat which place similar cells together in low-dimensional space. The same PCs were used as input as in cluster analysis.

Gene set enrichment analysis on marker genes

Gene IDs (in the form of ENSG00000000000) were mapped to their respective gene symbols using the SynGO geneset analysis tool (Koopmans et al, 2019). The clusters were filtered to

include only genes with adjusted p-value < 0.05. The genes in each cluster were sorted based on adjusted p-value, from smallest to largest. Top predicted cell types were determined for each cluster with PanglaoDB using the top 10 genes in each cluster (Franzen et al, 2019). Each of the filtered and sorted clusters was analyzed using Enrichr (Chen et al, 2013).

Labeling clusters to cell types

Data analysis took place with the tool R. Packages used for data analysis include; Suerat, BiocManager, limma, tximport, and dplyr. First, cell data was read into R using 'Tximport' and normalized using Seurat's NormalizeData() function. Gene markers were located and clusters were assigned using Seurat's FindAllMarkers() function; 13 clusters were located as opposed to the 15 found in the reference paper. Cluster ID's were assigned to the 13 clusters; "Alpha", "Beta", "Delta", "Gamma", "Epsilon", "Acinar", "Ductal", "Quiescent Stellate", "Activated Stellate", "Endothelial", "Macrophage", "Mast", "Cytotoxic".

Visualization of Cell Data

Seurat's Dimplot/Umap function was used to create a scatter plot featuring the previously mentioned clusters (**Fig. 9**). The top n = 5 (determined by weight log fold change) were sequestered and their expression plotted on a heatmap visualization (**Fig. 10**).

Labeling of cell types and Novel marker genes

Markers associated with certain clusters (i.e., GCG with the alpha cell cluster), were parsed an put into a table format (**Table 4**). The same was performed for occurrences of novel marker genes found in the cell data (**Table 5**).

Results

Pre-processing statistics

Fig. 11 shows the graphs used to select the whitelist cutoffs. To start, 1.324 billion barcodes were processed and 1,324,770,031/1,327,837,961 were used (99.99%). There were a total of 4,251,176 unique barcodes and 426,619 were whitelisted (same number of analyzed cells), 13.16% of reads were thrown away because of noisy barcoding. The index contained 233,652 targets and 0 decoys. The mapping rate was 43.01% and a total of 28,331,267 UMI were created after deduplication. Salmon alevin skipped 53,109 barcodes due to no mapped read. The total run time was 3 hours and 6 minutes.

Pre-processing UMI counts matrix

The Alevin data when imported through 'tximport' had 60232 genes in total. After creating the object using SEURAT using the filter for removing genes with less than 3 cells and minimum features 200, we got 29663 genes in total. Since the object had ENSEMBL identifiers as row names, we mapped them to gene symbols and got 29663 distinct genes.

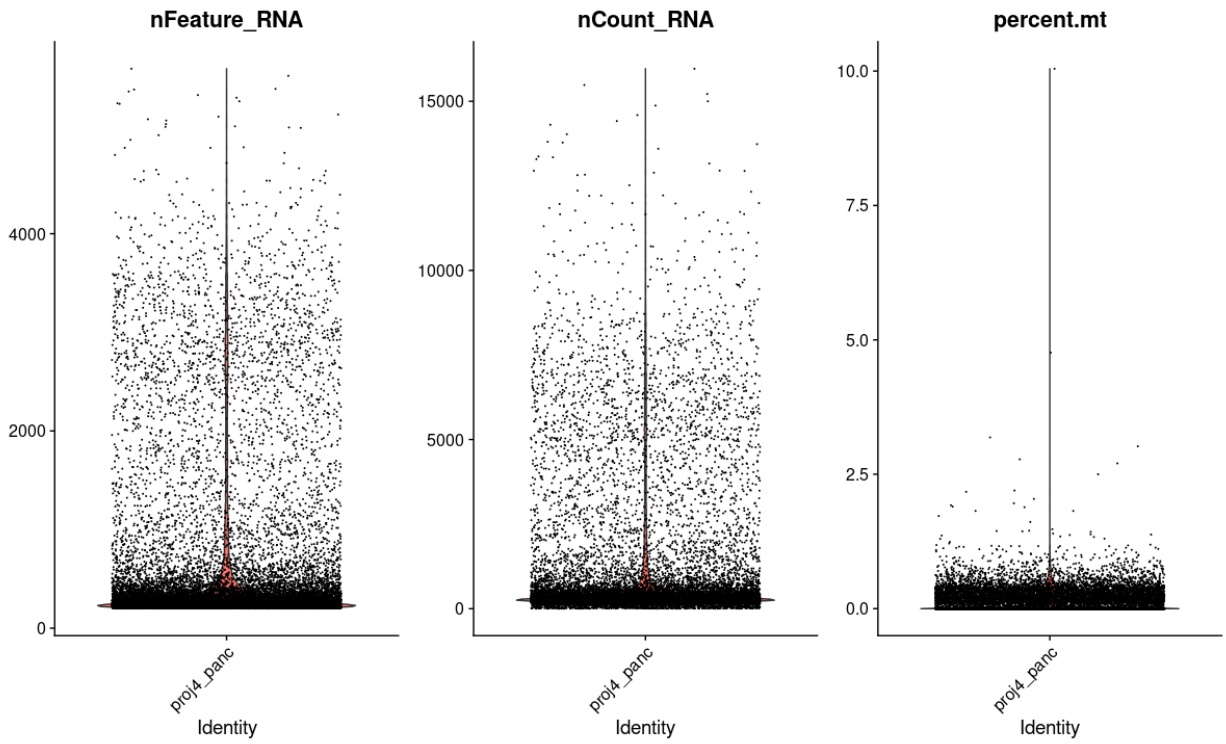

Figure 2: Violin Plot for quality analysis of counts matrix where nFeature_RNA is the number of genes detected in each cell, nCount_RNA is the total number of molecules detected within a cell and percent.mt is the percentage of mitochondrial genes for each cell.

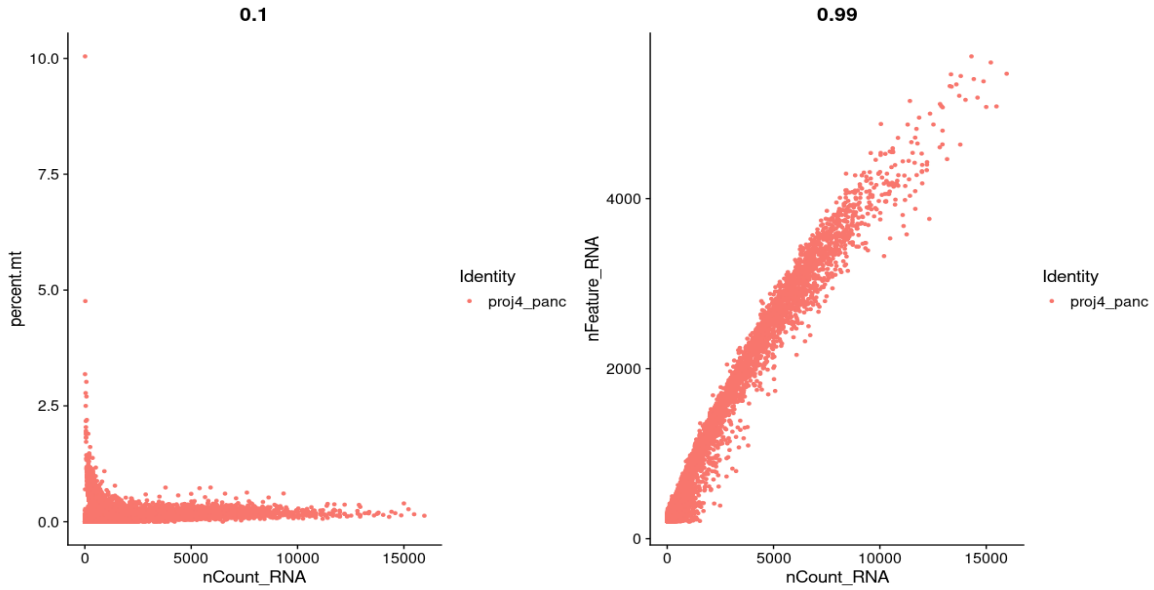

Figure 3: Scatter plot to visualize feature- feature relationships of nCount_RNA with percent.mt and nFeature_RNA

Due to high percentage of mitochondrial genes, the data was of low quality. To filter this, we subset it by removing cells with greater than 5% mitochondrial genes. Cells with fewer than 200 genes and greater than 2500 genes were also filtered. Thus, we got 27833 genes as a result.

Figure 4: Plot for variable features with labels where 2000 high variable features are highlighted in red and selected for further downstream analysis out of which top 10 are labelled in the plot.

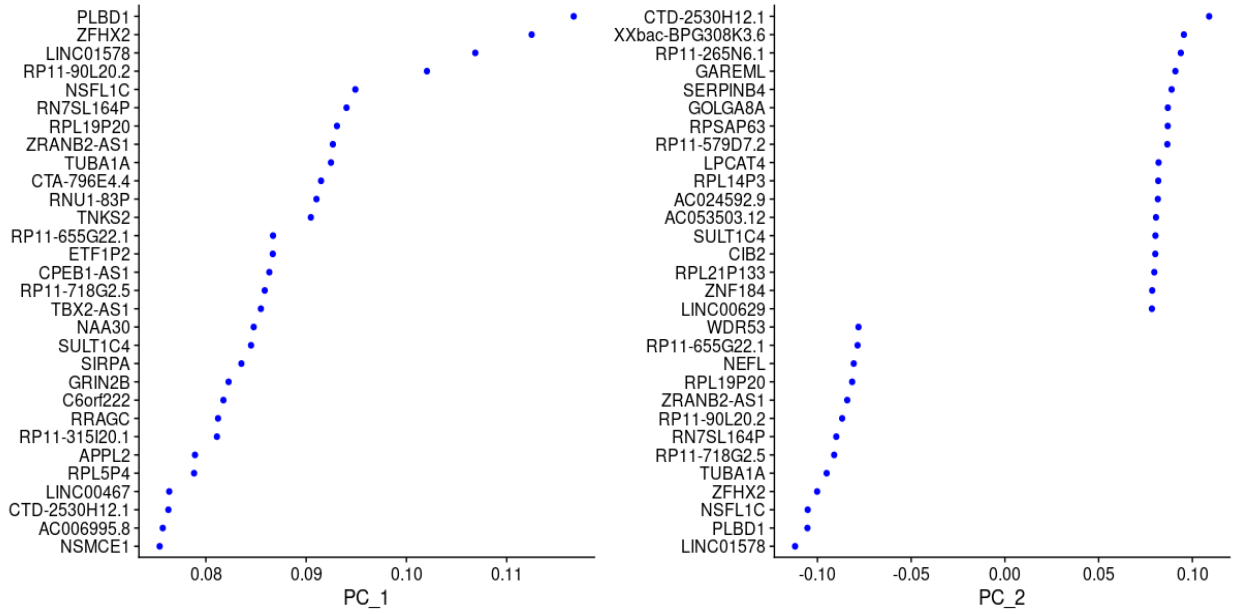

Figure 5: Scatter plots for gene subsets in PC_1 and PC_2 where PC_2 has both positive and negative gene groups.

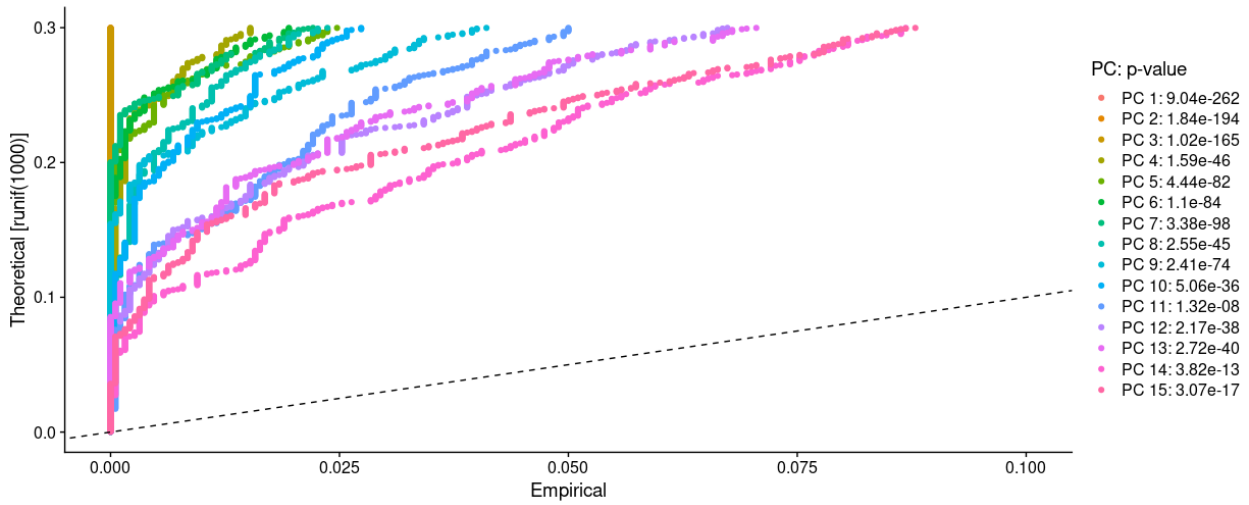

Figure 6: JackStraw Plot for each Principal component where the p value of each PC is displayed. All PCs show strong deviation from the expected statistic (dotted line) if there was no significance. They all show a strong true signal and no structure.

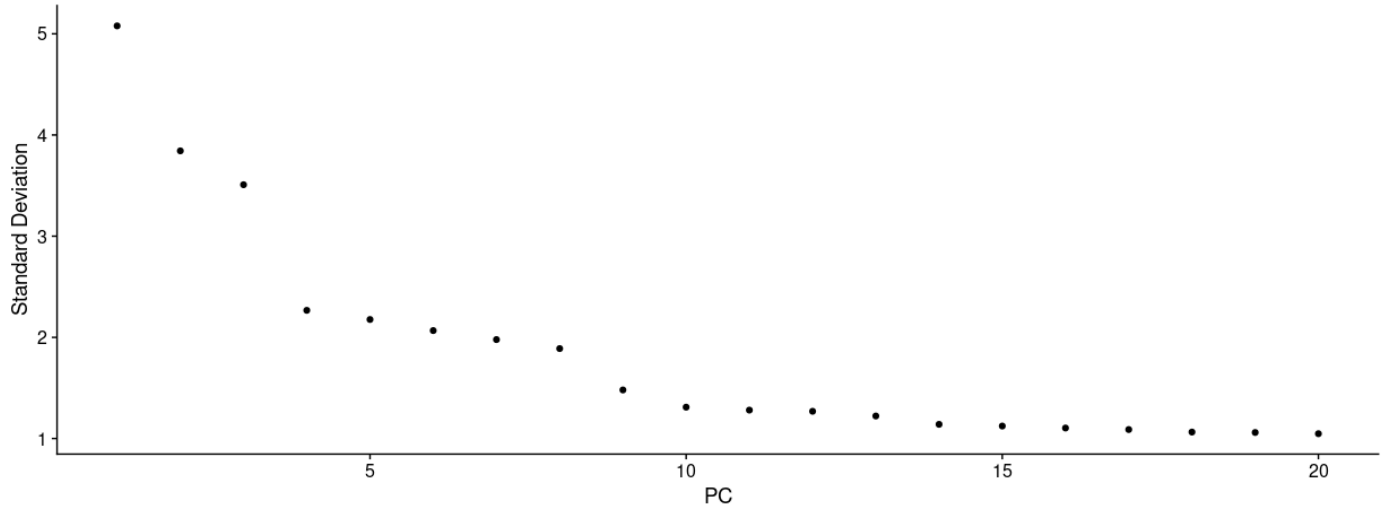

Figure 7: Elbow plot displays percentage variance of each PC where the elbow is formed between PC_5 and PC_10.

From the above figure, a noticeable decrease was found from PC_4 to PC_6 followed by that from PC_8 to PC_10. For the remaining, there is a steady decrease. This indicates that adding PC_5 and PC_9 will provide additional information to the clustering algorithm. However, this won't provide substantial information about the variation of cell populations. To find the clusters, UMAP reduction was used with 0.5 resolution. This gave us 13 clusters in total.

Our results did not exactly replicate the results from Baron *et al* (2014), as we identified 13 clusters of genes whereas Baron et al had genes grouped into 14 clusters. Baron et al. also provided parkers for each pancreatic cell type they identified. Only 5 of the marker genes in their analysis appeared in our top gene lists for our clusters. Because of this, cell types could

not be assigned to all of our clusters based on the Baron et al marker genes, so cell types were also predicted using PanglaoDB (Tables 1 and 2). One of the Baron *et al.* (2016) marker genes appeared in more than one of our clusters (INS), and one of our clusters (cluster 0) had more than one of the marker genes. Several of the marker genes provided by Baron et al. did not appear in any of the top marker gene lists for our clusters. The pathways, GO terms, and disease risks associated with our clusters varied, although many reported pathways and functions were related to protein digestion (Table 3).

Table 1. Cell Types Assigned to Marker Genes from Baron et al, with genes in red indicating those that were found in the top gene lists for the clusters.

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**

Table 2. Top Genes for Each Cluster, With Cell Type Assignments

Cluster	Top differentially expressed genes	Cell Type according to Baron et al.	Top Predicted Cell Type with PanglaoDB
0	SST, ND4, FP671120.6, ND1, FP671120.7, INS, IAPP, ND2, CYTB, ATP6	Beta or Delta	Endothelial
1	REG1A, CTRB2, REG3A, REG1B, PRSS2, AL049839.2, SERPINA3, SPINK1, FP671120.6, FP671120.7	N/A	Enterocytes
2	AC027290.2, GCG, EEF2, TTR, H3-3B, MYL6, SCGN, RPL11, RPL32, SCG2	Alpha	Endothelial
3	SLC7A2, FXYD5, FXYD3, TTR, GC, CRYBA2, ALDH1A1, TM4SF4, IRX2, EGFL7	N/A	Endothelial
4	IGF2, MAFA, INS, AC132217.2, C1QL1, IAPP, PCSK1, HADH, DLK1, GNAS	Beta	Endothelial
5	GRN, TMSB10, KLF6, PKM, ACTB, CD59, RPS16, RPS19, HSPB1, RPL28	N/A	Neurons
6	XACT, RPS6KA5, AL022322.2, AL162458.1, ALDH6A1, PATJ, TOR1AIP2, AL590822.2, ACER3, NFIA	N/A	Fibroblasts

7	FTL, CTSD, ND6, ND2, ND5, ND1, TMSB4X, IFI30, AC007192.1, FP671120.6	N/A	Endothelial
8	PSME4, TYRO3, PHKB, ARSB, CTIF, OASL, SCUBE3, SCUBE1, SEMA4C, RAD9A	N/A	Germ Cells
9	NOTCH3, MMP11, COL1A1, SPARC, PDGFRB, FN1, IGFBP5, FMOD, PXDN, COL5A1	Stellate	Fibroblasts
10	IL32, ALDOB, DUOXA2, CTRB2, SERPINA3, AL049839.2, PRSS2, SPINK1, C3, ALB	N/A	Enterocytes
11	FLT1, RGCC, VWF, PASK, CD93, F2RL3, PODXL, PLVAP, ACVRL1, PLPP3	Vascular	Endothelial
12	TTC27, PPP1R12B, DIMT1, PCDH11Y, FZD3, ZNF175, POLR1G, PIK3CA, TTLL9, PPHLN1	N/A	Germ Cells

Table 3. Predicted Pathways, GO Terms, and Disease Implications for our 13 Clusters.

Cluster	Pathways	GO Biological Process	GO Molecular Function	Disease Risk
0	Electron transport chain, Oxidative phosphorylation	negative regulation of protein oligomerization	Hormone Activity	Type 1 diabetes

1	Protein digestion and absorption, Activation of matrix metalloproteinases	Response to peptide	Peptidoglycan binding	Pancreatitis
2	Translation, Cytoplasmic Ribosomal Proteins, Lysosome	Neutrophil degranulation	Beta-galactosidase activity	Intestinal malabsorption, Maturity-onset diabetes
3	Cytoplasmic ribosomal proteins	Cotranslational protein targeting	RNA binding	Anemia
4	Translation, Cytoplasmic Ribosomal Proteins	SRP-dependent cotranslational protein targeting to membrane	RNA binding	Anemia
5	Endosomal/vacuolar pathway, Phagosome,Tight junction	Antigen processing	RNA binding	hereditary nephrotic syndrome, Malignant neoplasm of prostate, psoriasis
6	Serotonin receptor, electron transport chain, Alzheimer disease	Membrane protein ectodomain proteolysis	Sulfuric ester hydrolase activity	Diseases of pancreas, mitochondrial diseases
7	Electron transport chain, oxidative phosphorylation	Antigen processing and presentation of exogenous peptide antigen	Aspartic-type peptidase activity	Sepsis, Logical memory (delayed recall), Central retinal vessel vascular tortuosity
8	Downstream signaling events of B cell receptors, oxidative phosphorylation	Extracellular matrix disassembly	Arylsulfatase activity	Lipoprotein disorders, Central retinal vessel vascular tortuosity

9	Cytoplasmic ribosomal proteins	SRP-dependent cotranslational protein target	RNA binding	Anemia, Serum thyroid-stimulating hormone levels
10	Cytoplasmic ribosomal proteins	SRP-dependent cotranslational protein target	RNA binding	Anemia, Age-related macular degeneration, Adenocarcinoma
11	ECM-receptor interaction	Extracellular matrix organization	Activin binding	Chronic heart disease, Neoplasm Metastasis
12	Plasma membrane estrogen receptor signal	Antigen processing and presentation of exogenous peptide antigen	G-protein beta/gamma-subunit complex binding	Type 1 diabetes, lung cancer

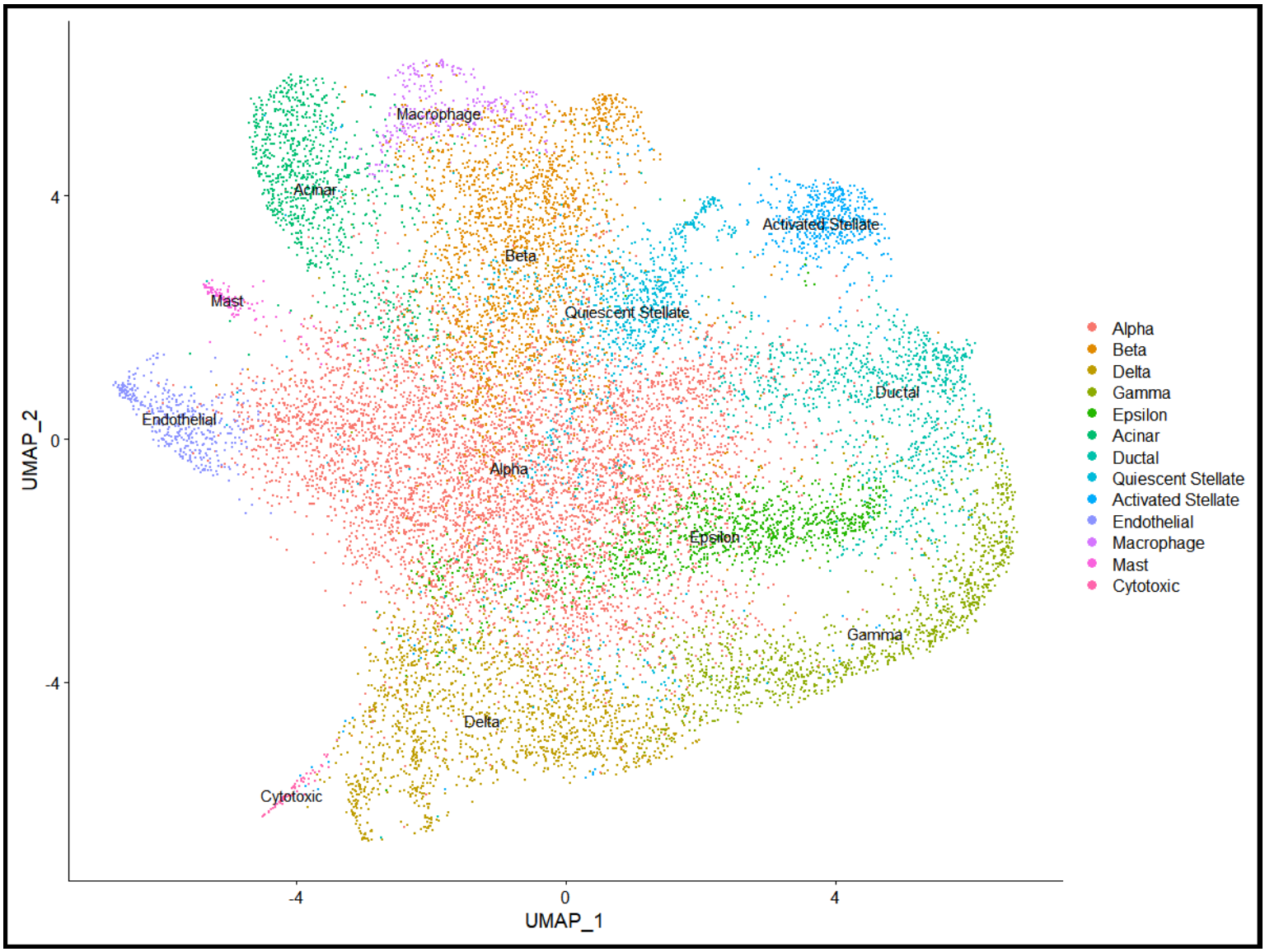

Figure 9: DimPlot using 'umap' reduction. Each color represents a different cell cluster, clusters are labeled by cell type. Each cell type presents a characteristic profile.

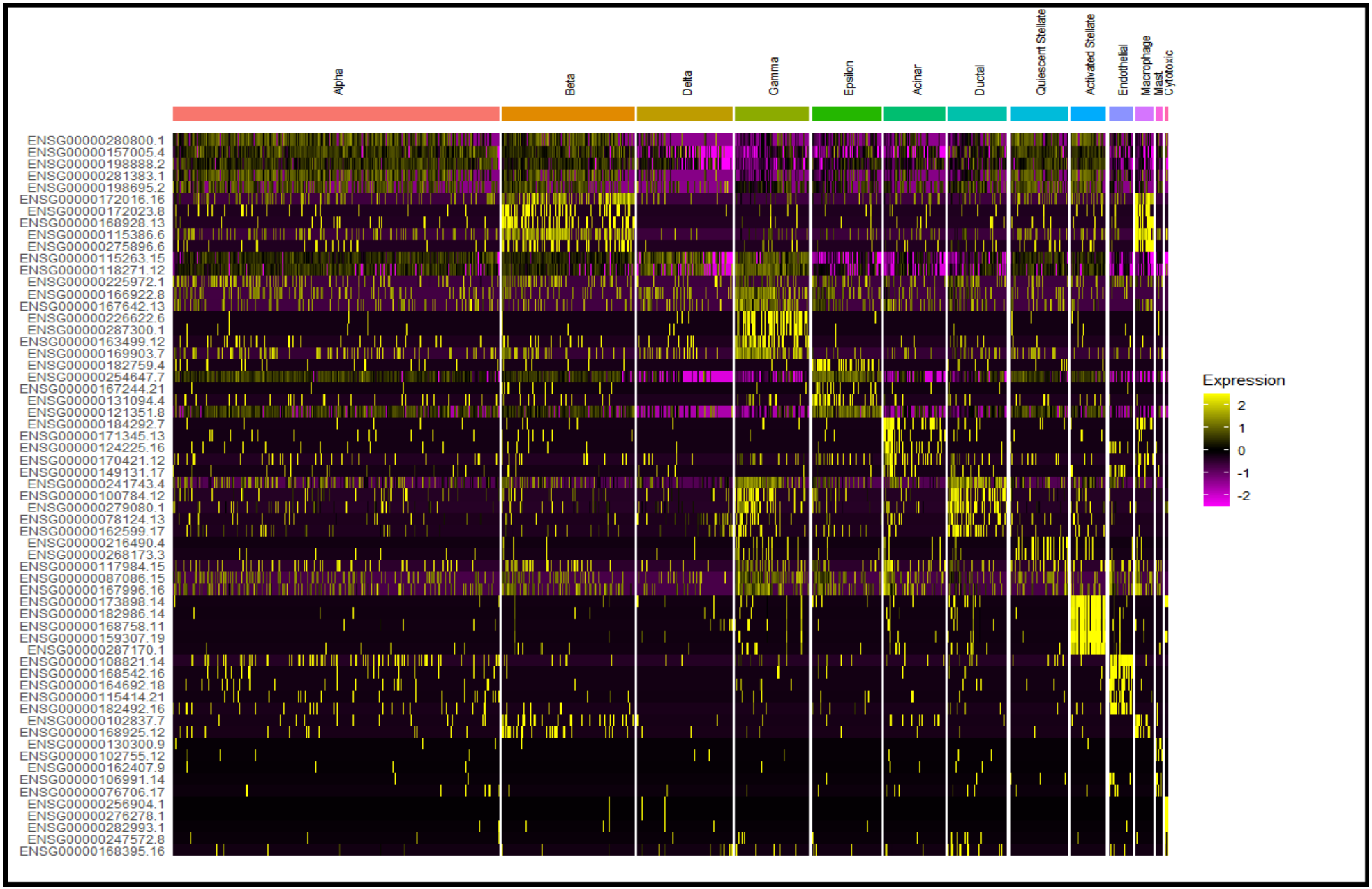

Figure 10: DimPlot using 'umap' reduction. Each color represents a different cell cluster, cell types are labeled in clusters.

Table 4. Labels of clusters as a cell type based on marker genes. Clusters, cell types markers, and cell types not listed were not located in the dataset.

P Value	avg_log2FC	pct.1	pct.2	Adjusted P Value	Cluster	Cell Type Marker	Cell Type
4.29E-66	0.944593862	0.831	0.948	1.27E-61	2	GCG	Alpha
8.47E-184	0.936546439	0.996	0.932	2.51E-179	3	GCG	Alpha
1.24E-240	0.417954403	0.976	0.813	3.67E-236	0	INS	Beta
0	1.688828195	0.986	0.859	0	4	INS	Beta
0	0.62660539	0.979	0.853	0	0	SST	Delta
1.37E-05	0.47478687	0.354	0.898	0.407467645	12	SST	Delta
6.22E-136	0.47992527	0.617	0.445	1.85E-131	0	PPY	Gamma
3.13E-10	0.31093738	0.599	0.499	9.30E-06	8	PPY	Gamma
7.14E-236	3.120015453	0.374	0.027	2.12E-231	10	CPA1	Acinar
2.08E-294	2.459586469	0.319	0.042	6.17E-290	5	KRT19	Ductal
0	2.766091559	0.267	0.002	0	11	VWF	Endothelial

Table 5. Novel marker genes.

P value	avg_log2FC	pct.1	pct.2	Adjusted P Value	Cluster	Gene
1.41E-299	0.836531	0.835	0.638	4.19E-295	Alpha	ENSG00000281383.1
5.64E-305	1.648525	0.358	0.083	1.67E-300	Beta	ENSG00000275896.6
7.40E-93	0.524637	0.279	0.1	2.19E-88	Delta	ENSG00000280138.1
0	2.824092	0.461	0.021	0	Gamma	ENSG00000287300.1
1.04E-168	0.977244	0.278	0.057	3.08E-164	Epsilon	ENSG00000286190.2
8.87E-128	1.403429	0.388	0.121	2.63E-123	Acinar	ENSG00000273259.3
1.47E-212	1.65319	0.348	0.068	4.36E-208	Ductal	ENSG00000285796.1
0	2.738088	0.413	0.053	0	Quiescent Stellate	ENSG00000268173.3
0	1.48081	0.68	0.05	0	Activated Stellate	ENSG00000287170.1
2.61E-13	0.502091	0.593	0.407	7.74E-09	Endothelial	ENSG00000233927.5
4.68E-22	0.821553	0.29	0.108	1.39E-17	Macrophage	ENSG00000285976.2
1.14E-260	3.225769	0.345	0.009	3.39E-256	Mast	ENSG00000261371.6
3.14E-77	1.305619	1	0.144	9.32E-73	Cytotoxic	ENSG00000288663.1

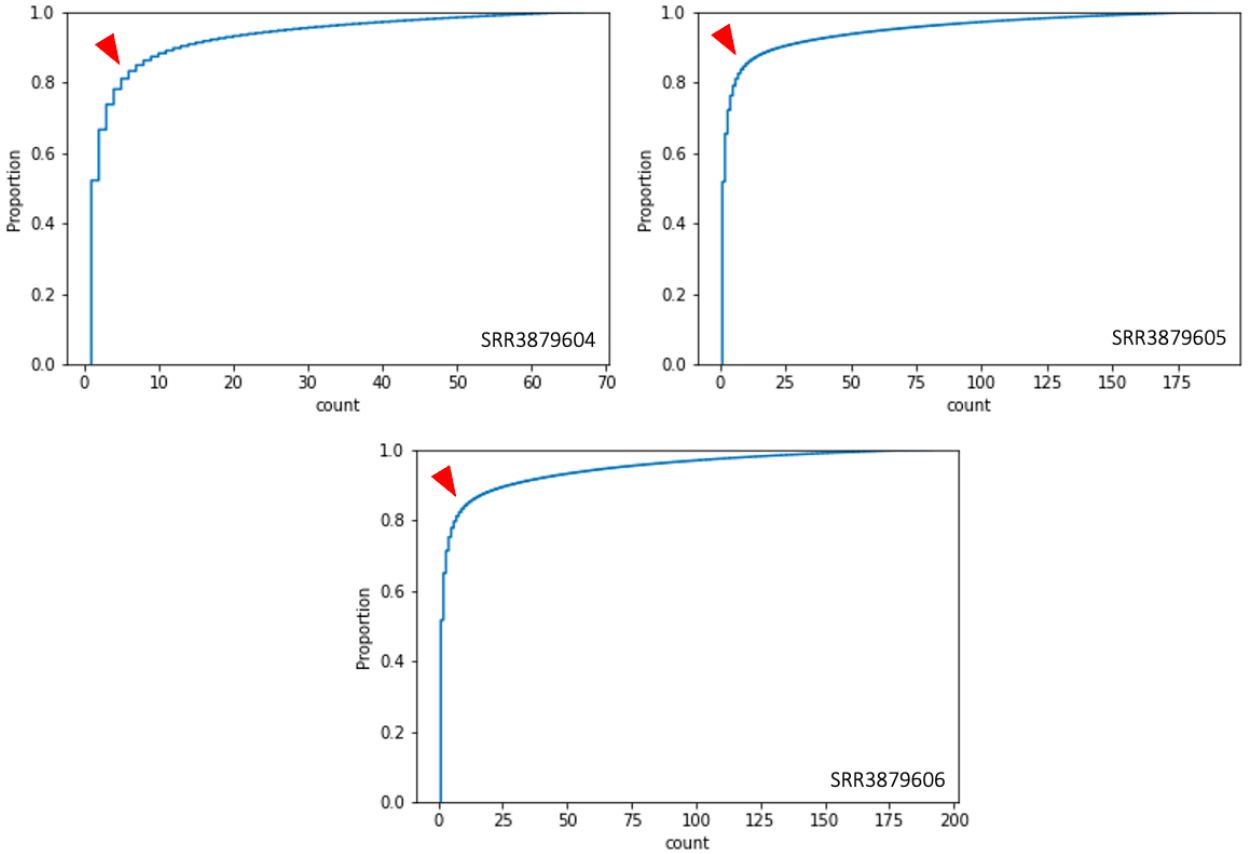

Figure 11. Cumulative distribution plot of barcode counts. The red triangle represents the empirical position where d/dx=0 which was selected as the threshold for whitelist selection. Count thresholds were 8, 13, 12 for samples SRR3879604, SRR3879605 and SRR3879606 respectively.

Discussion

Gene expression studies like bulk RNA-seq mask the heterogeneity present within a sample. This makes the characterization of tissues like the pancreas where multiple cell types share a spatial niche to be inaccurate since changes in expression of the minor cell population is likely to be filtered out or unable to reach statistically significance during quality control. scRNA allows for the aggregation of cells through deconvolution methods. We were able to successfully determine the transcriptomes of thousands of pancreatic cells. When comparing our results to those of Baron et al. (2016), only 13 of the 14 cell clusters were detected. This could be due to the fact that only one human subject was analyzed, decreasing the power of the analysis and many DE transcripts were unlikely to reach statistical significance.

Of the 13 cell clusters we identified, 5 had marker genes in common with the Baron et al. (2016) analysis. The rest had to be computationally predicted. This led to identification of certain cell types that are not found in pancreatic tissue. Because the data came from a pancreatic tissue sample, we know these cell types to be incorrect. Many of the clusters had more than 10 genes with highly significant adjusted p-values. It is possible that including more genes in the comparison or filtering differently would have yielded better results.

There was a high amount of variability in the pathways, GO terms, and diseases identified for the clusters. Some of the classifications were expected for pancreatic tissue, such as protein digestion and risk of type 1 diabetes and pancreatitis. However, some did not make sense for a pancreatic tissue sample, including logical memory delayed recall, central retinal vessel vascular tortuosity, and chronic heart disease.

Clustering did not present problems in practice, but in outcome there are several differences between our results and Baron et al. The most notable difference being differing numbers of clusters; While Baron et al. were able to divide their data into 14 clusters, while we were able to only output 13 clusters (the exception being the Schwann cell cluster). Further, as shown in **Table 4**, several of the clusters could not be classified in accordance with their predicted genetic markers. This is to say, their gene markers could not be found in the cell data; exceptions include Epsilon, Quiescent Stellate, Activated Stellate, Macrophage, Mast, Cytotoxic, and Schwann.

As for our heatmap, similarities were found between Baron et al.'s heatmap. The noticeable downward expression curve crossing the heatmap is present in our graphic as well, albeit with much more noise. Our scatterplot was not as lucky; not only is there not much correlation between cluster location between our scatterplot and Baron et al.'s, ours features much more overlap

between individual data points. Both the noise found in our heatmap and the overlapping data in our scatterplot can most likely be attributed to more stringent and targeted filtering of their dataset.

Conclusion

Although we were not able to perfectly replicate the results from Baron et al, our results still show that single cell RNA sequencing data can be useful for identifying cell types and functions of interest. Our analysis, along with that of Baron *et al.* (2016)l, demonstrates the ability of single cell RNA-seq to be used in novel cell subtype discovery.

References

1. Baron, M., Veres, A., Wolock, S., Faust, A., Gaujoux, R., Vetere, A., Ryu, J., Wagner, B., Shen-Orr, S., Klein, A., Melton, D. and 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), pp.346-360.

2. 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)*.

3. Frankish, A., Diekhans, M., Ferreira, A., Johnson, R., Jungreis, I., Loveland, J., Mudge, J., Sisu, C., Wright, J., Armstrong, *et al*., 2018. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1), pp.D766-D773.

4. Hwang, B., Lee, J. and Bang, D., 2018. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, 50(8), pp.1-14.

5. Klein, A., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. and Kirschner, M., 2015. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*, 161(5), pp.1187-1201.

6. Koopmans, F., van Nierop, P., Andres-Alonso, M., Byrnes, A., Cijsouw, T., Coba, M. P., … Verhage, M. (2019). SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. *Neuron*, *103*(2). https://doi.org/10.1016/j.neuron.2019.05.002

7. Melsted, P., Booeshaghi, A., Liu, L., Gao, F., Lu, L., Min, K., da Veiga Beltrame, E., Hjörleifsson, K., Gehring, J. and Pachter, L., 2021. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nature Biotechnology*,.

8. Patro, R., Duggal, G., Love, M., Irizarry, R. and Kingsford, C., 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4), pp.417-419.

9. Olsen TK, Baryawno N. Introduction to Single-Cell RNA Sequencing. Curr Protoc Mol Biol. 2018 Apr;122(1):e57. doi: 10.1002/cpmb.57. PMID: 29851283.

10. Oscar Franzén, Li-Ming Gan, Johan L M Björkegren, PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data, Database, Volume 2019, 2019, baz046, doi:10.1093/database/baz046

11. Seurat - Guided Clustering Tutorial. (2021). Retrieved May 2, 2021, from Satijalab.org website: https://satijalab.org/seurat/archive/v3.1/pbmc3k_tutorial.html

12. Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zagar, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar B. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo,Peter Smibert and Rahul Satija, Integrated analysis of multimodal single-cell data, bioRxiv, 2020, 10.1101/2020.10.12.335331, https://doi.org/10.1101/2020.10.12.335331

13. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences, Charlotte Soneson and Michael I. Love and Mark D. Robinson, 2015, F1000Research, 10.12688/f1000research.7563.1,4,1521

14. Johannes Rainer (2017). EnsDb.Hsapiens.v79: Ensembl based annotation package. R package version 2.99.0.

15. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Steffen Durinck, Paul T. Spellman, Ewan Birney and Wolfgang Huber, Nature Protocols 4, 1184-1191 (2009).

16. Lawrence M, Huber W, Pag\`es H, Aboyoun P, Carlson M, et al. (2013) Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol 9(8): e1003118. doi:10.1371/journal.pcbi.1003118

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

18. Douglas Bates and Martin Maechler (2021). Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.3-2. https://CRAN.R-project.org/package=Matrix

19. Wang X and Cairns MJ (2013). Gene Set Enrichment Analysis of RNA-Seq Data: Integrating Differential Expression and Splicing. BMC Bioinformatics, 14(Suppl 5):S16.

20. Wang X and Cairns MJ (2014). SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing. Bioinformatics, 30(12):1777-9.

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