Single Cell RNA-Seq Analysis of Pancreatic Cells

Data curator - Varun
Programmer - Neha Gupta
Analyst - Rachel Thomas
Biologist -Yichi Zhang
TA- Jackie

Introduction

The pancreas plays a specific role in helping the body to achieve and maintain energy homeostasis through secreting digestive enzymes and metabolic hormones [5]. A majority of the pancreas is composed of two main exocrine cell types: acinar and duct cells. Acinar cells produce digestive enzymes, including amylase, lipase, and peptidases and duct cells secrete bicarbonate [2]. Islets make up the additional 5% of the pancreatic mass. Within the Islet mass lies exocrine tissue and channels that house endocrine cells secreting hormones responsible for glucose homeostasis [7]. Among these endocrine cells there is a wide range of cell types present.

Dysfunction with the pancreas can lead to cancer, and Type 1 and 2 Diabetes. It has been shown that this diverse population of cell types in exocrine regions is crucial for the overall pancreas function[4] Therefore, a complete profiling of the molecular make-up of pancreatic cell types is necessary to gain a deeper understanding of their connection with disease.

Ever improving technological methods has equipped scientists with the tools necessary to dive deeper into the biomolecular world. Single-cell RNA-sequencing is a method that characterizes transcripts at a cellular level using unique barcode sequences. As a result, opportunity to gain new insights into the molecular heterogeneity of cell types within tissue is provided. Using single-cell RNA-seq, Baron et al. classified the cell types within the human and mouse pancreas based on their transcriptome and showed clear separation of distinct cell types using a visualized projection plot. The aim of this study is to replicate the results of Baron et al. using different data processing, clustering, and visualization techniques.

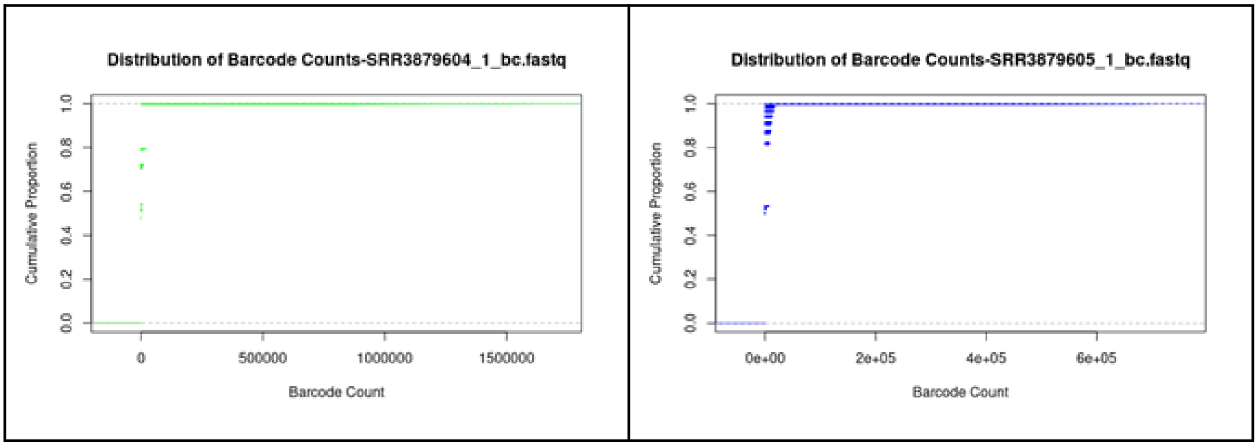
Data:

Barcode Filtering:

The overall dataset was sourced from the GEO database [13] (specific link is as follows: GEO Accession viewer (nih.gov)), focusing on a single female donor who was 51 years old. The dataset included samples of pancreatic islets (sections of the pancreas associated with the endocrine system) from 89 human donors with different blood glucose levels, with and without Type 2 diabetes [13]. Three SRA (Sequence Read Archive) files were extracted containing multiple reads linked with nucleotide “barcodes” (simplified barcode files that linked to the reads were provided in class). The reads were 3’ end tagged, and the barcodes were 19 bases long. Due to the extreme number of reads, and the relatively simple format of the SRA files, a combination of command line-based text processing tools and batch commands were used to extract the barcodes and counts for further analysis (awk, grep, tr, for example).

The first stage of analysis involved the creation of cumulative distribution plots (**Figure #1)** for each file to view the overall distribution of barcode frequencies, using the RStudio (v1.3) Empirical Cumulative Distribution Function (ecdf). A barcode that was not counted many times would imply an erroneous read, or at least one not statistically significantly present in the data to be used for analysis. If such barcodes were present in the data, or if the graphs indicated that it wasn’t uncommon to find barcodes with small counts in the data, then that would indicate contamination, or some other issue with the dataset. Instead, the plots seem to show that a large count is not uncommon(**Figure #1 top right,** especially).

However, as the plots indicate, a few values were discovered between the extremely large counts and extremely small counts. They were different enough from the least counted barcodes to warrant further analysis in later steps. Therefore, for each file, the mean for the overall number of counts was found as the first filter level, again using command-line based text parsing. This reduced even the largest file from having over 1,000,000 rows (1,293,792 distinct barcodes for the first file,1,333,842 for the second, and 1,227,152 for the third) to less than 50,000 rows each. As there were still counts excluded close to the average value (the count number for various barcodes were often repeated in small clusters), the mean was then subtracted (for each file) by the standard deviation of the number of counts. The resulting barcodes were then combined for the “whitelist” of barcodes used in the next step.



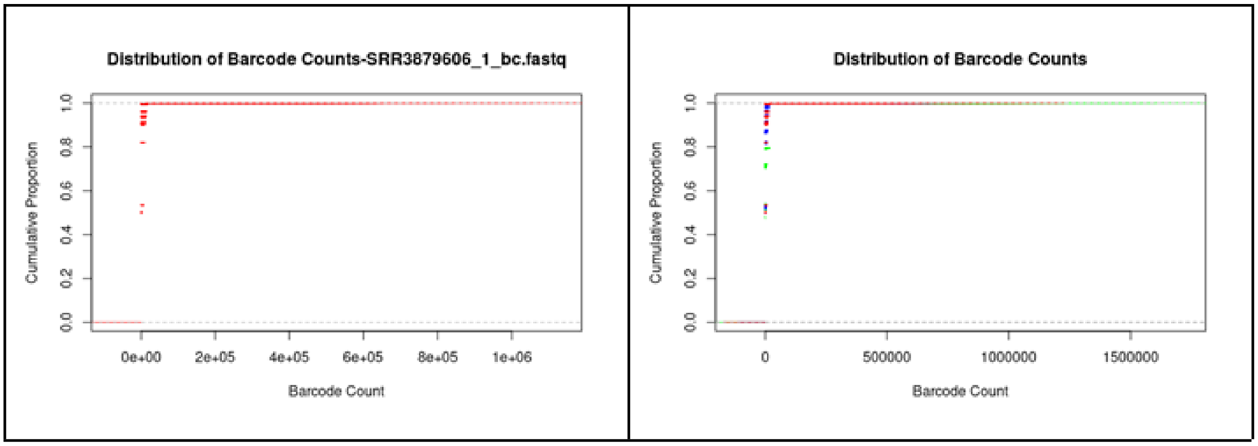


Figure #1--Cumulative Distribution Plots of Barcode Plots: Each plot with a single color is of a single file, with the final plot in the bottom right corner having them overlaid on each other.

Salmon Alevin:

Salmon [17] is a tool for transcript quantification. Alevin [16], a tool using Salmon [17] software, can be used to map and quantify reads, and can produce a cell-by-gene count matrix. The transcriptome (the most current human reference) was sourced from the Genecode [15] website(accession number GSE50244 [14]). The transcript to gene map was also created using annotations from Gencode [15]. Then Salmon [17] was used to produce an index for the transcriptome. The results were then plugged into an Alevin [16] query on the genes, which then produced the (Unique Molecular Identifier) UMI counts matrix. The “whitelist” of barcodes was provided as an input, and was used for cell detection instead of the auto-generated “whitelist” Alevin [16] would use. The resulting matrix could be used for further analysis, as it linked the reads to gene and cell-based datasets.

Methods:

Creating Seurat Object:

After the dataset generation of Pancreatic cells, we will be analysing it using the Seurat package[10]. First, we read the data through the tximport package(version 1.16.1)[11] using the *tximport* function where type argument is used to specify what software was used for estimation. A simple unique molecular identifier (UMI) is returned, where the transcript level information is summarized to the gene-level. The rows in this matrix represent the number of molecules for each gene(feature) whereas columns represent it’s detection in each cell.

Mapping gene symbols and filtering genes:

After reading the data, we created a Seurat object(version 4.0.0). The object stores information about both data(like count matrix) and analysis(like PCA results) for a single cell dataset. We mapped the ensembl gene ids to their respective gene symbols using EnsDb.Hsapiens.v79 package(version 2.99.0)[12] from Bioconductor. For preprocessing, we first calculate the percentage of counts originating from a set of mitochondrial genes using PercentageFeatureSet function. We used the set of all genes starting with MT- as a set of mitochondrial genes. We visualized(figure 2) these counts, features and percent using violin plot and then filtered for cells having unique feature counts over 2500 or less than 200 and >5% mitochondrial counts. Another scatter plot(figure 3) was made to visualize the feature to count and feature to percent.mt relationships.

Normalization:

After filtering the dataset, we normalize the data using NormalizeData() function using the LogNormalize method. This method 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. Next we subset by filtering low variance genes. Using the FindVaribaleFeatures() function we selected 2000 high variable features genes through the 'vst' parameter. Vst method accounts for the mean-variance relationship that is inherent to single cell RNA-seq. A set of dot plots(figure 4) were created to visualize the standard variance and average expression for variable and non variable counts. We applied scaling(linear transformation) prior to dimensional reduction techniques like PCA. ScaleData function was used.

Linear Dimensional Reduction:

Now we did PCA on this scaled data using RunPCA () and this was visualized using VizDimLoadings (figure 5) and DimPlot(figure 6). We implemented a resampling test inspired by the JackStraw procedure(took ~10 minutes for it's execution). We randomly permute a subset of the data (1% by default) and rerun PCA, constructing a 'null distribution' of feature scores, and repeat this procedure. We identify 'significant' PCs as those who have a strong enrichment of low p-value features. The JackStrawPlot function was used to visualize distribution of p-values for each PC with uniform distribution(figure 6). Significant' PCs will show a strong enrichment of features with low p-values. An alternative method, ElbowPlot() function was used to plot as well(figure 7).

Cell Clustering:

For clustering cells, we first construct a KNN graph based on the euclidean distance in PCA space, and refine the edge weights between any two cells based on the shared overlap in their local neighborhoods. This step is performed using the FindNeighbors function, and takes as input the previously defined dimensionality of the dataset. To cluster cells, the Louvain algorithm was applied to group cells together. During clustering, resolution was set to 0.5. The clusters are shown in figure 8.

Using the UMAP technique of Seurat, non-linear dimensional reduction was visualized and explored. The goal of these algorithms is to learn the underlying manifold of the data in order to place similar cells together in low-dimensional space. Same PCs that were used as input to clustering analysis were used as input to UMAP.

Differentially Expressed Features:

Once the data set was clustered, differential expression analysis was performed to biomarkers that define clusters in the data set. The Seurat R package [6] was used to perform differential expression which is based on the non-parametric Wilcoxon rank sum test. This test was chosen because it ranks all cluster values from smallest to largest and does not require a normal distribution of data. Each cluster produced a set of associated markers based on the top up and downregulated genes. The top genes were then used for classifying clusters by cell types.

Cell Type Classification:

Clusters were then labeled as a cell type based on the marker genes identified. Additionally, the supplementary material (Table S2, figure 1) provided in the Baron *et al*. (2016). paper [5] and the Panglao Database [7] were used to match single cell expression data with associated cell types. Figure 9 depicts the various clusters with accompanying cell type. The specific marker genes associated with each cell type cluster are shown in the Figure 10 heatmap.

Visualize Top 5 Marker Genes per Cluster

The top 5 marker genes for each cluster were identified from the differential expression analysis and using the Log2 fold change as the selective parameter. Then, a heatmap was produced to visualize the top 5 significantly expressed genes per cluster.

Find novel marker genes

The novel marker genes for each cluster were identified by using a 0.05 p-adjusted threshold for gene expression and filtering out the genes that were used to identify cell types in the earlier sections. Identified novel genes were then used for gene set enrichment analysis.

Results:

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 27659 genes in total. Since the object had ENSEMBL identifiers as row names, we mapped them to gene symbols and got 26069 distinct genes.

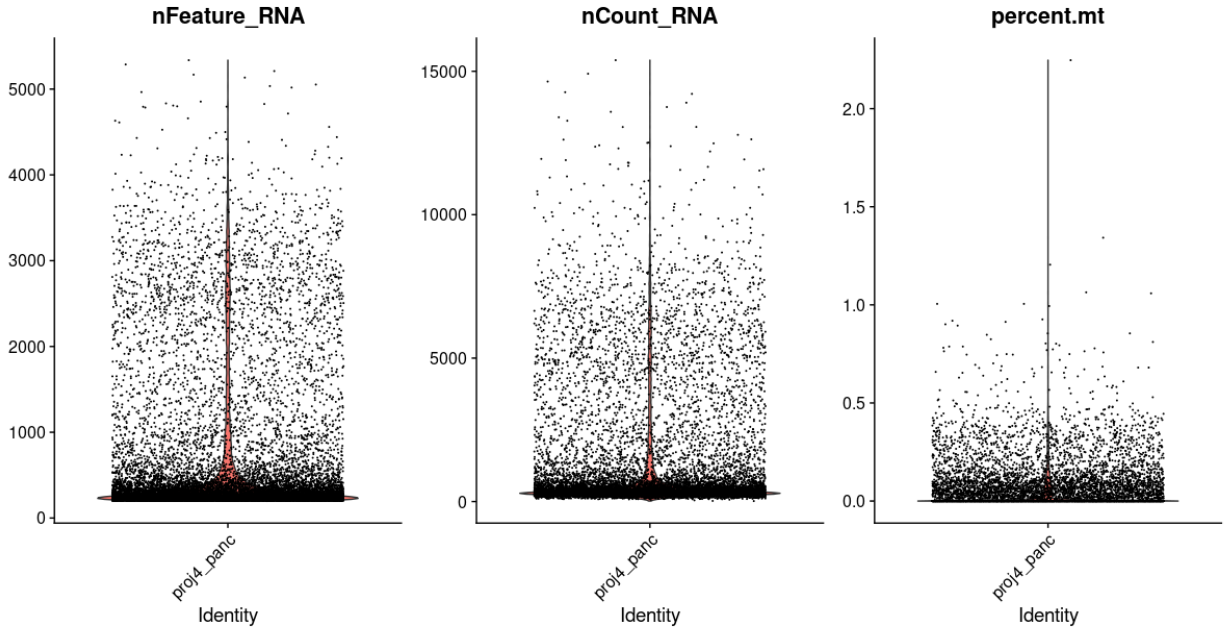


Figure 2: Shown above are violin plots for the number of reads, expressed genes and percentage of mitochondria genes for each cell.

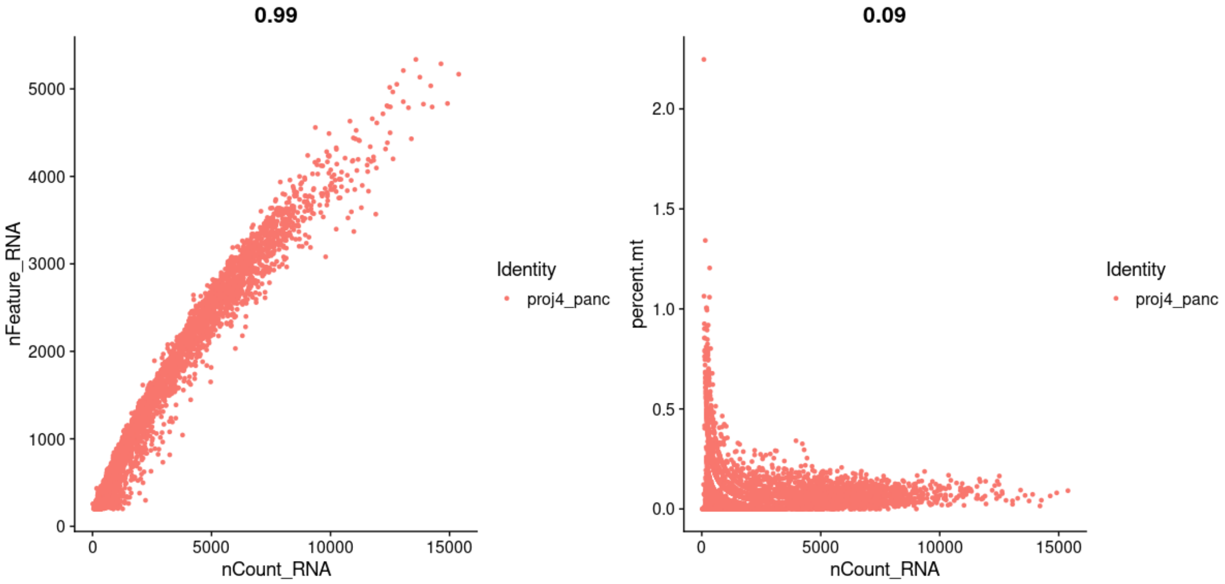


Figure 3: The scatter plots for the reads-expressed genes and reads-mitochondrial genes relationships. This plot is to visualize feature-feature relationships of nCount_RNA with percent.mt and nFeature_RNA.

Due to high percentage of mitochondrial genes, the data was of low quality. To filter this, we created a subset by removing cells with greater than 5% mitochondrial genes. Cells with fewer than 200 genes and greater than 2500 genes were also filtered. Therefore, we got 26069.

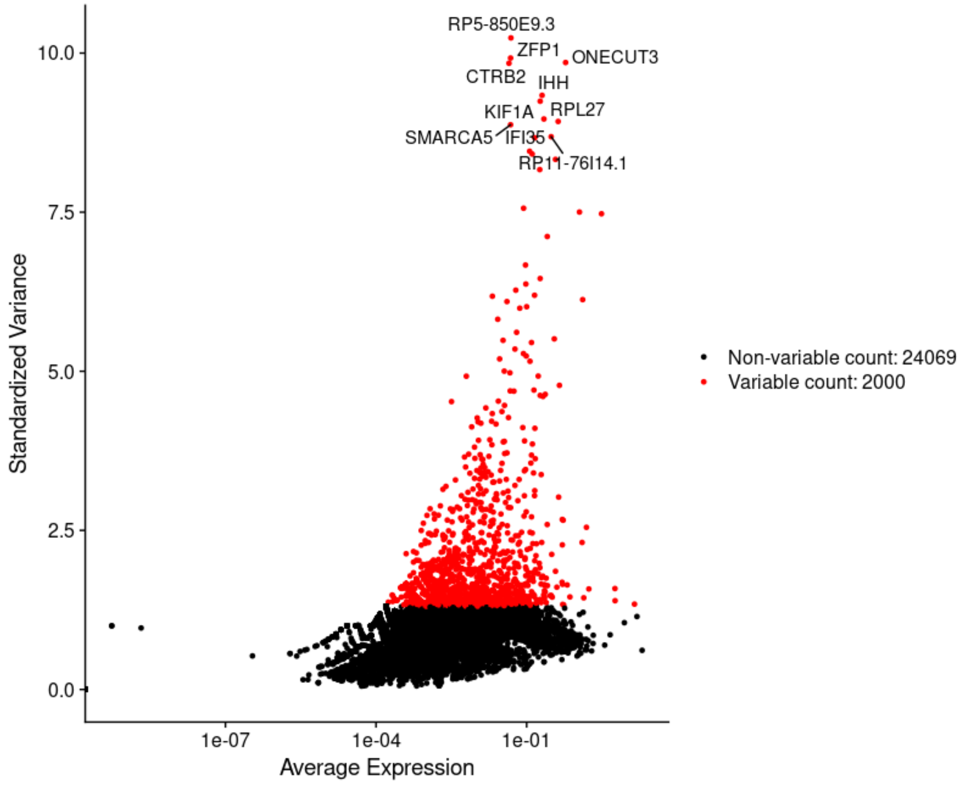


Figure 4: Plot for variable features with labels where 2000 high variable features are highlighted in red and selected for further downstream analysis out of which top 10 are labelled in the plot. Red dot genes were kept for following analyses, the black dots are the genes filtered out.

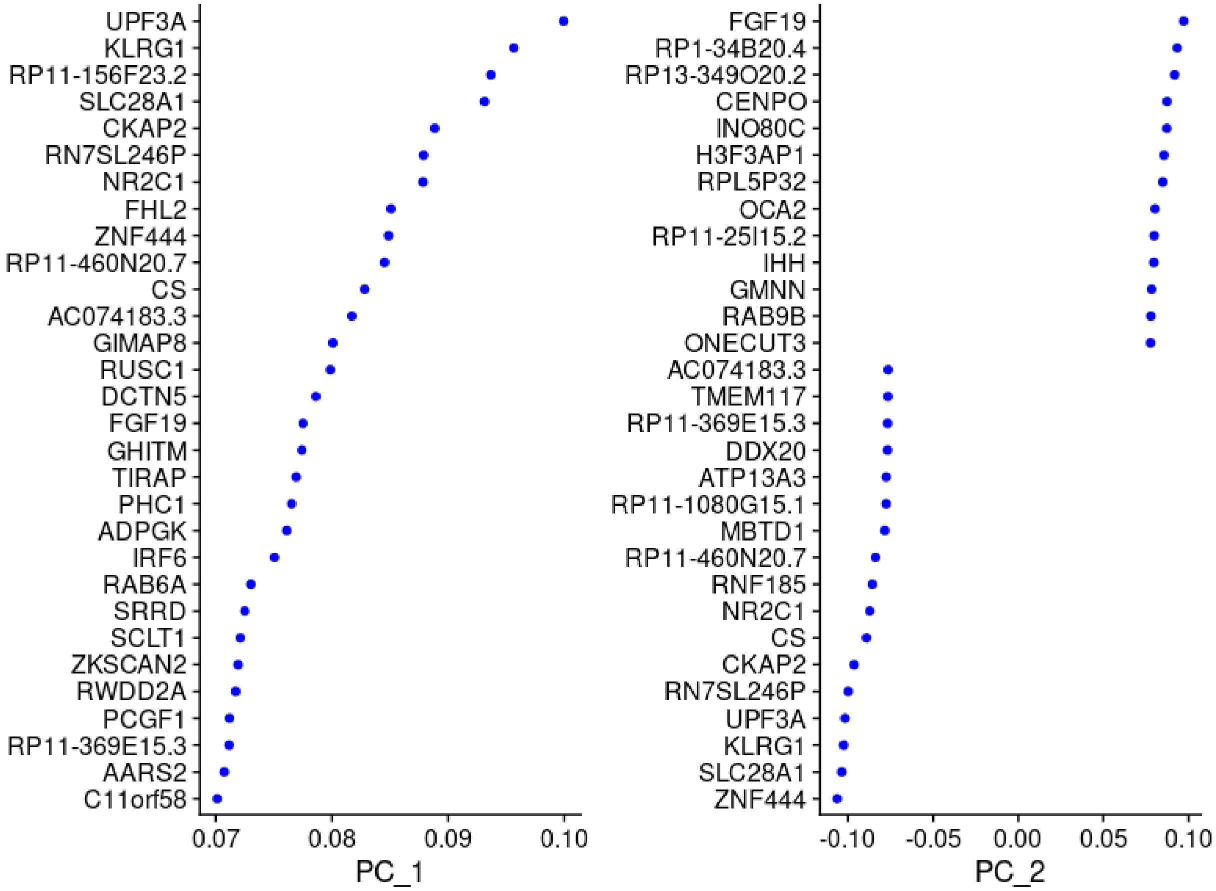


Figure 5: The scatter plot for gene subsets in Pc1 and PC2. PC1 has a positive gene group whereas PC2 has a negative gene group also.

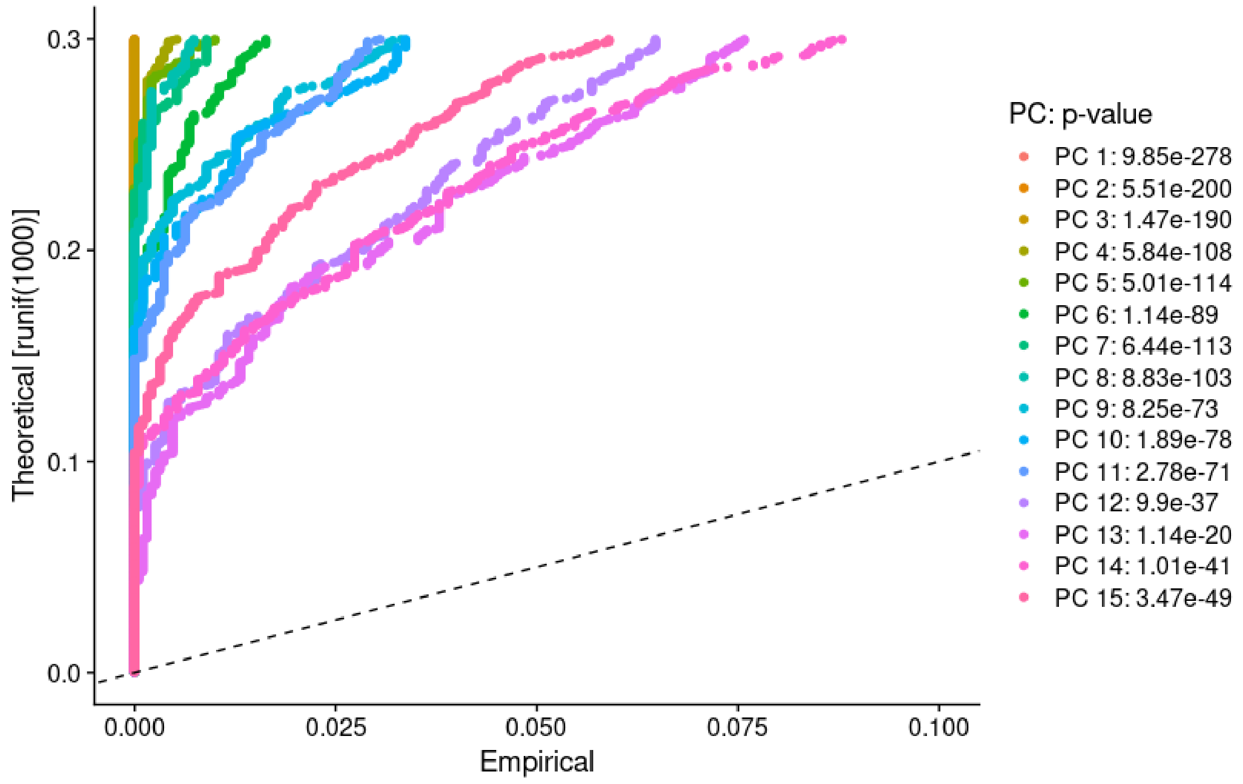


Figure 6: The Jackstraw plot for each PC. The plot displays the p value of each PC formed.

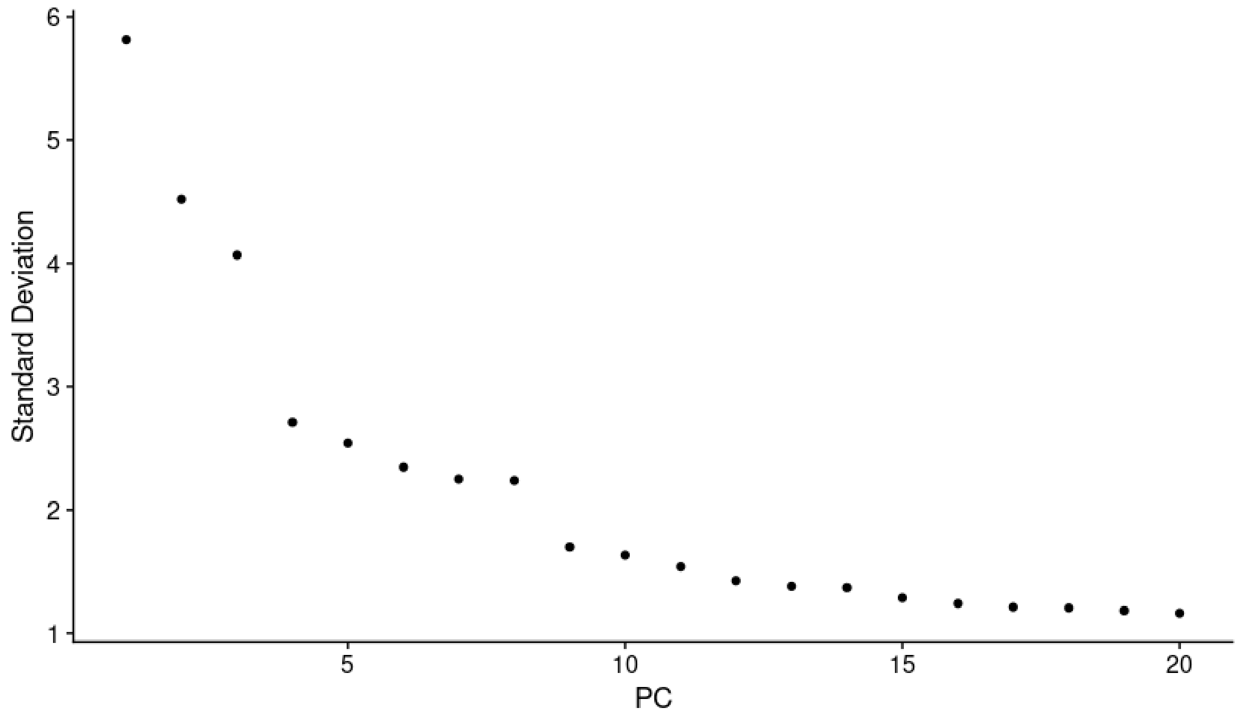


Figure 7: This elbow plot displays standard deviation(y axis) for each PC.

There is a noticeable decrease in standard deviation from principal components from PC3 to PC 4 and PC 8 to PC9. There is a steady decrease in deviation from the remaining PCs. This indicates that adding PC 4 and PC 9 would provide additional information to the clustering algorithm. However, adding the following Pcs to the analysis would not necessari;y add substantial information about the variance of the cell populations. Using RunUMAP, 13 clusters were identified as shown in figure below.

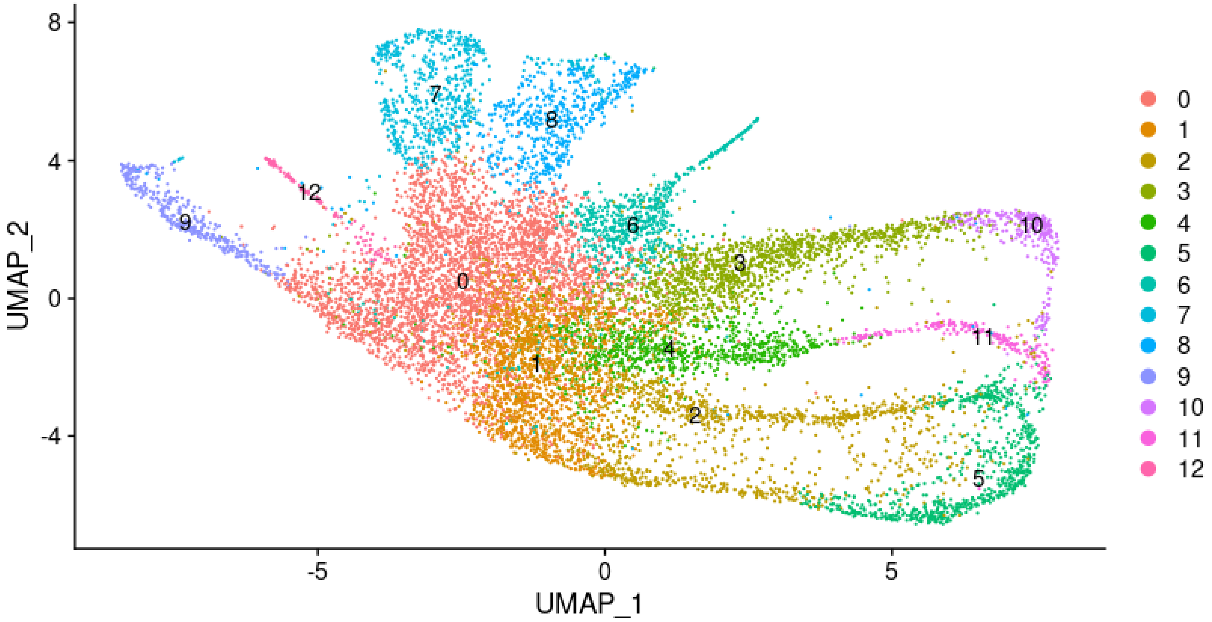


Figure 8: The clustering results in the UMAP method into 13 clusters.

Cluster	**Cell type**	**Marker Gene**
0	Delta	SST
1,6	Beta	INS
2,4,8,10,11	Alpha	GCG
3,11	Acinar	CPA1
5, 10	Ductal	KRT19
7,12	Macrophage	CD68
9	Stellate	PDGFRA, PDGFRB
10,11	Cytotoxic T	C3, C8

Table 1 Marker genes and cell type identifiers for each cluster. Baron *et al*. (2016). Table S2 [5] was used as a reference to label clusters. There were no Gamma, Epsilon, Vascular, or Mast cells identified from differential expression analysis in our sample data set.

Cluster	**Cell Type Identifier**
0	Delta
1	Beta
2	Alpha_1
3	Acinar
4	Alpha_2
5	Ductal

6	Beta_2
7	Macrophage
8	Alpha_3
9	Stellate
10	Cytotoxic T
11	Cytotoxic/ Acinar
12	Macrophage

Table 2: Labeled Cell Types. Clusters with more than one cell type identifier were further analyzed and 1 cell type was chosen to represent each cluster, with the exception of cluster 11 which is represented by two cell types. Alpha and beta cells were most prevalent and were labeled the designated cell type for multiple clusters.

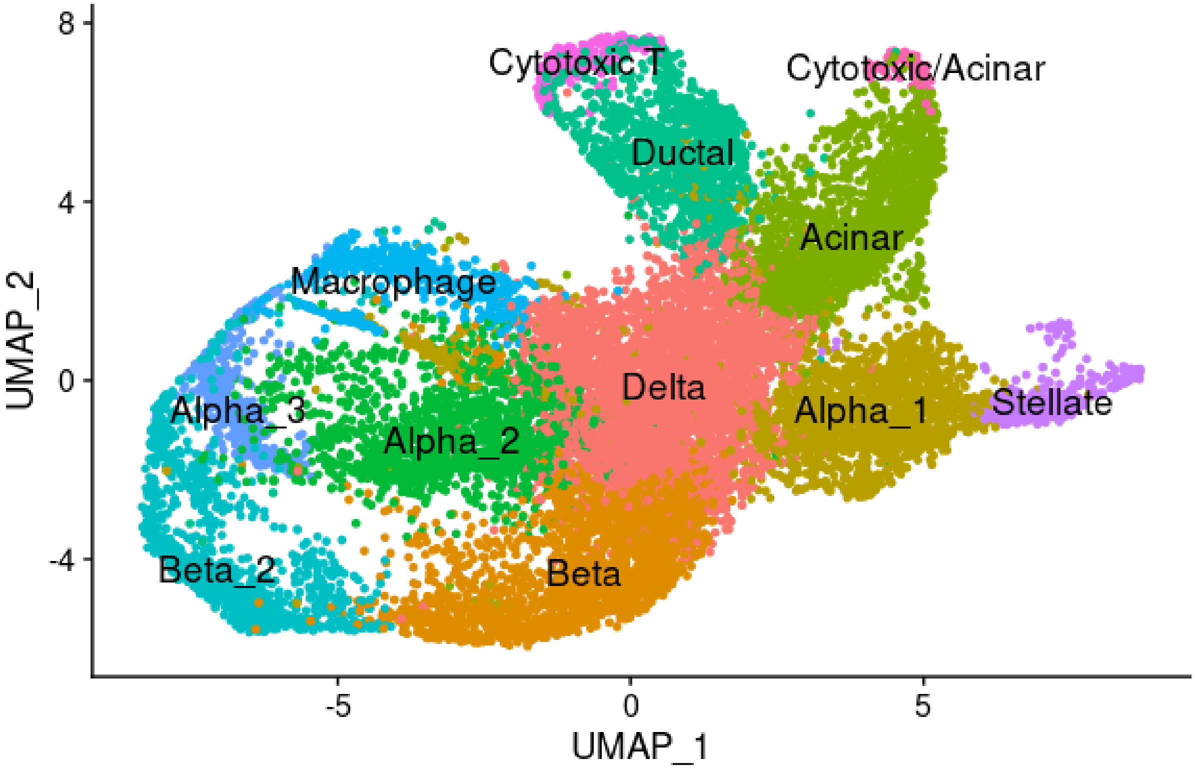


Figure 9 : Identified cell types based on marker genes. Thirteen different cell types were identified based on marker genes from each cluster. The top genes were chosen

and cell types were identified using the Panglao Cell type identification Database as Table S2 of the Baron*et al*. (2016) supplementary data. Each cluster and cell type are distinguished by a different color.

Figure 10 : Top marker gene identified for each cell cluster.

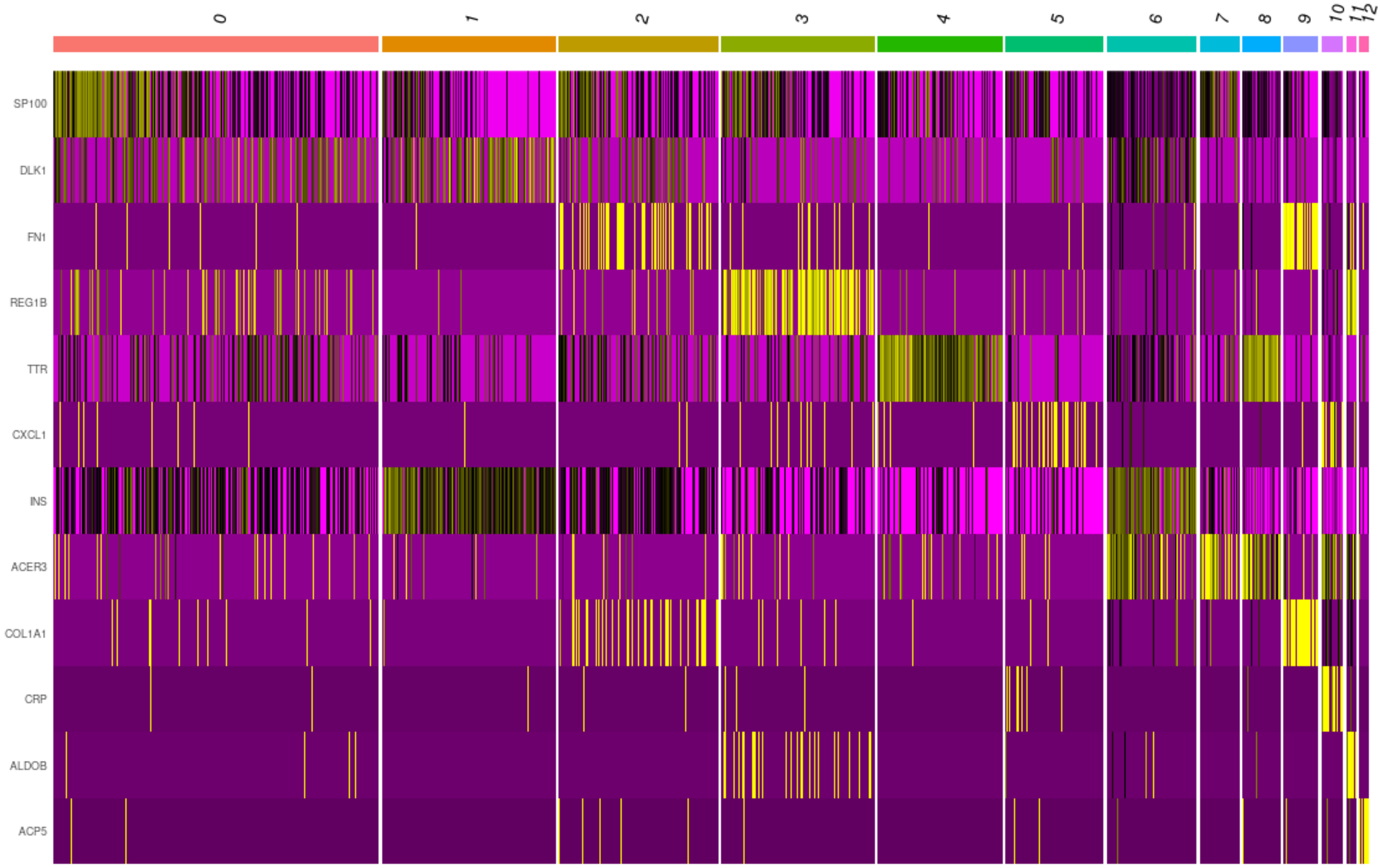


Figure 10a shows the expression of each marker gene relative to the identified cell type clusters. Each clustered cell type is based on log normalized UMI counts of associated genes. (a) top gene from each cluster, (b) top 5 genes from each cluster visualized

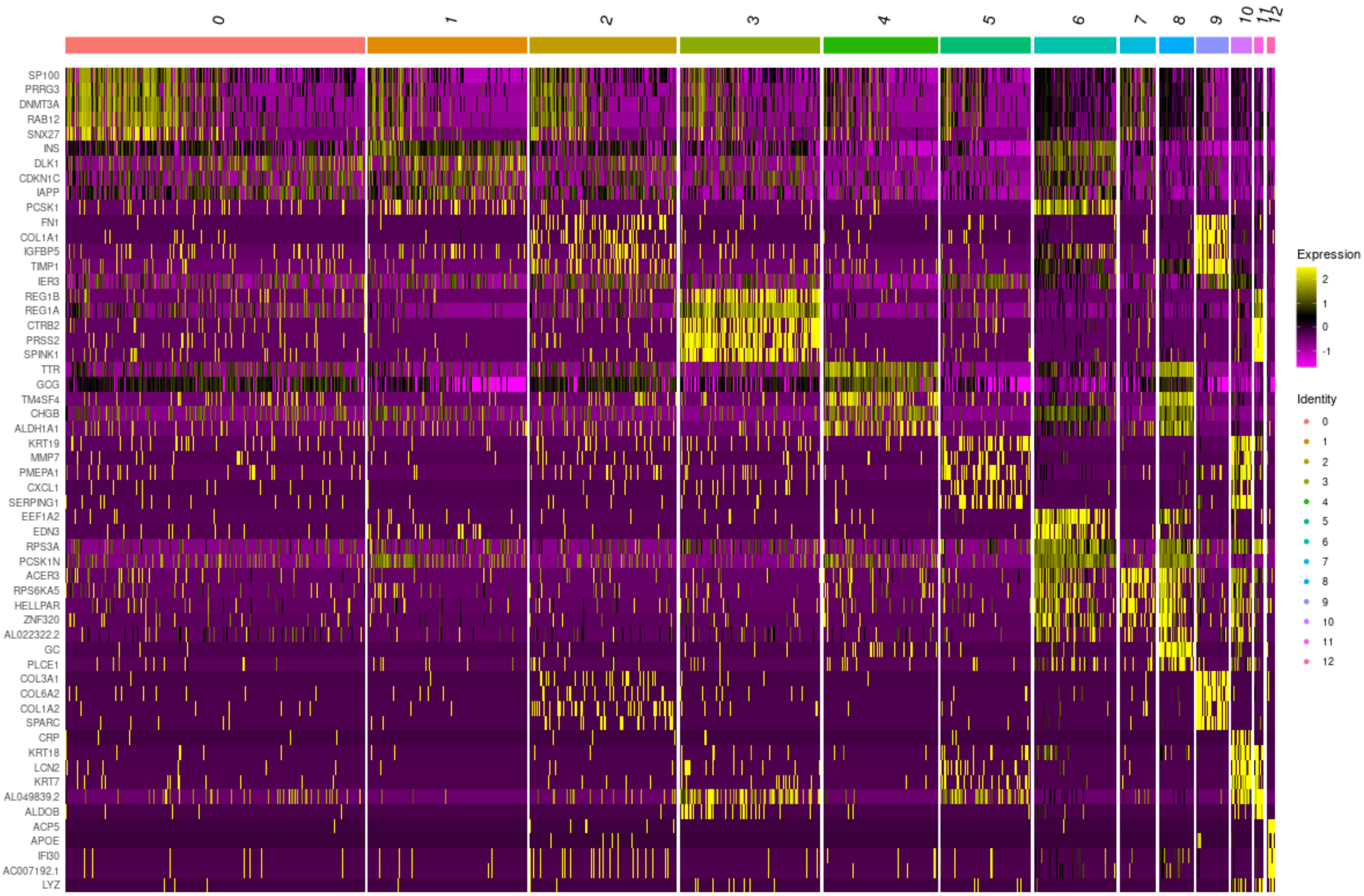


Figure 10b Top 5 marker genes identified for each cell cluster

Gene Set Enrichment Analysis

6487 marker genes were generated from the sample Seurat object with processed, clustered counts. In a filter of adjusted p-value not greater than 0.05 and average log2 Fold Change greater than 1.5, 494 genes were selected in confidence that expressed significant differences. A file containing gene symbols was passed to *Metascape*[8] for interpretation of systems-level studies of enriched biological pathways and protein complexes contained.

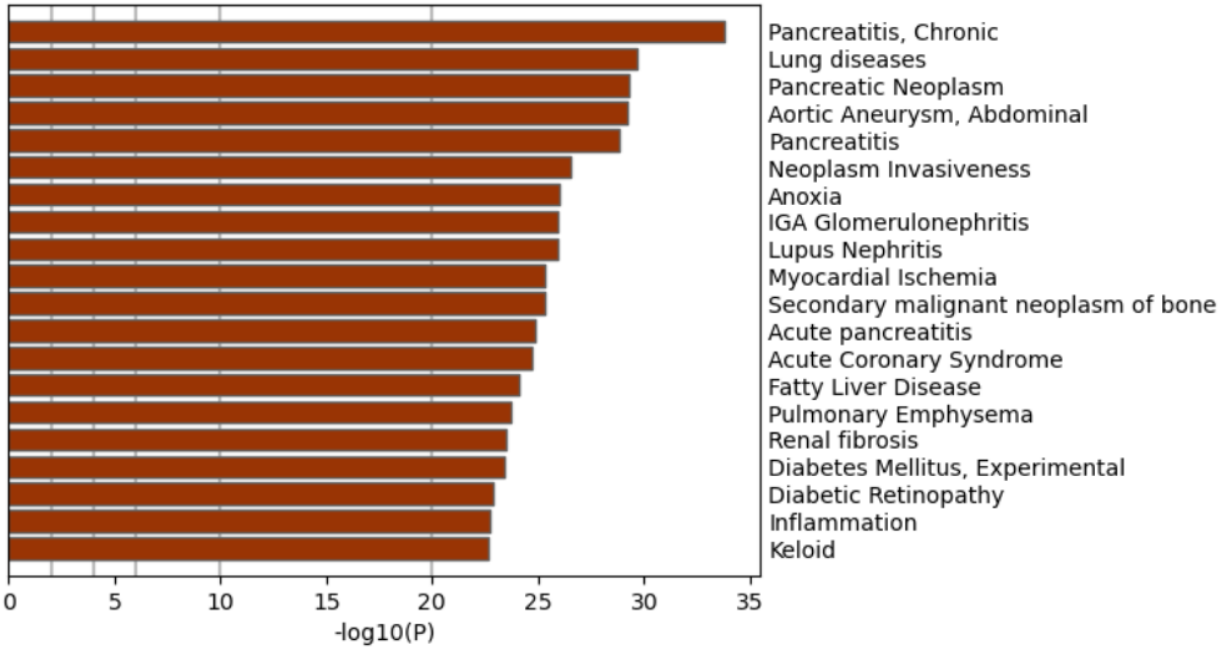


Figure 11: Bar graph of top 20 clusters in DisGeNET for gene functions and roles in biological events, colored by p-values.

For each given gene list, pathway and process enrichment analysis has been carried out with the following ontology sources: KEGG Pathway, GO Biological Processes, Reactome Gene Sets, Canonical Pathways, CORUM, TRRUST, DisGeNET, PaGenBase, Transcription Factor Targets, WikiPathways, PANTHER Pathway and COVID. All genes in the genome have been used as the enrichment background. Terms with a p-value < 0.01, a minimum count of 3, and an enrichment factor > 1.5 (the enrichment factor is the ratio between the observed counts and the counts expected by chance) are collected and grouped into clusters based on their membership similarities. The most statistically significant term within a cluster is chosen to represent the cluster.

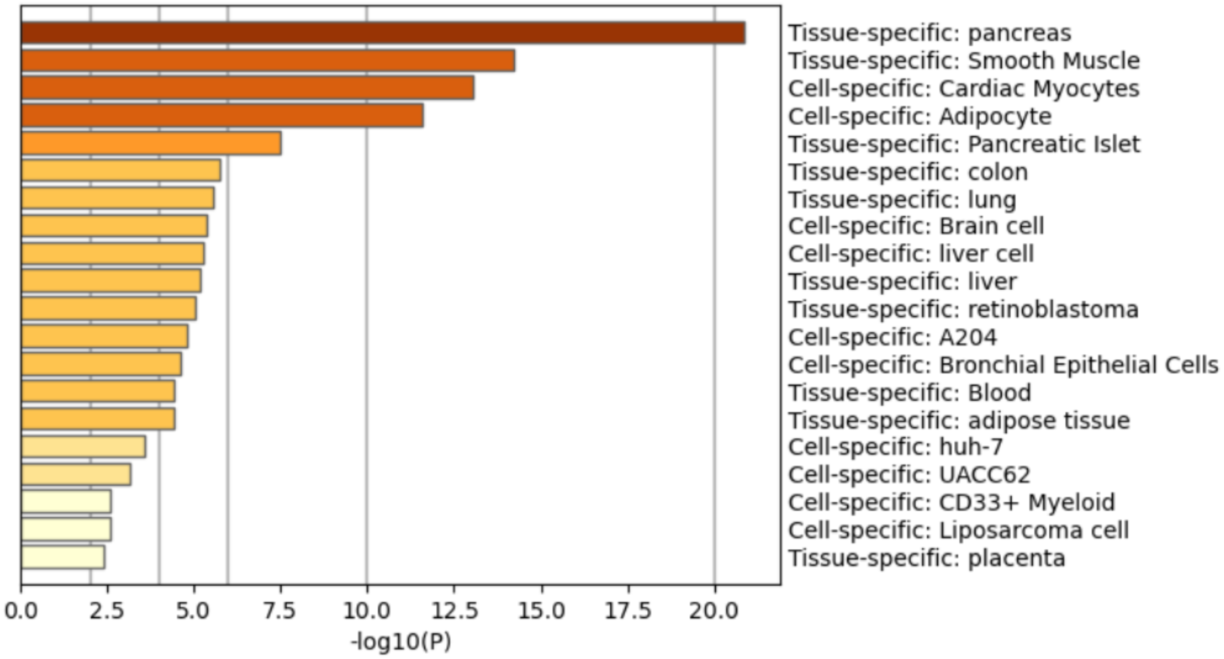


Figure 12: Bar graph of top 20 clusters in PaGenBase for gene functions and roles in biological events, colored by p-values.

pattern genes are defined as a group of genes that exhibit modularized expression behavior under serial physiological conditions Three types of pattern genes are presently attracting significant attention: housekeeping genes, specific/selective genes and repressed genes. Housekeeping genes, which are expressed ubiquitously across tissues under all physiological conditions and developmental stages, are generally believed to maintain basal cellular functions. Housekeeping genes are typically adopted as molecular controls in qualitative or semi-quantitative measurements of gene expression. Specific/selective genes are genes that are preferentially expressed under one or more conditions. Their enriched expression levels are typically considered markers of the initiation or existence of some biological phenomena, such as development, proliferation, differentiation or pathogenesis.

Discussion:

We attempted to recreate the single cell RNA-seq analysis conducted by Baron *et al*. (2016) to verify the cell types they identified. The initial clustered results produced 13 clusters, however, after classification only 8 different cell types remained. There were also cases of cell type overlap in clusters even after filtering. As shown in Table 1, through our clustering analysis we found 8 of the 15 cell types identified by Baron *et al*.

(2016). Table 2, displays our clustering and cell types produced from our analysis. We were able to identify 8 out of 12 cell types identified by Baron *et al*. (2016). These include 3 endocrine cells (alpha, beta, delta,), both exocrine cells (ductal, acinar), 2 immune cells (macrophage, cytotoxic T cells), and stellate cells.

The projection plot produced from our experiment displays very similar cell type clusters, but the arrangement of the clusters and the size vary from the project plot produced in Baron *et al*. (2016). This could be due to the difference data sets used or in plot tools, UMAP vs tSNE projection plots. Processing the UMI counts matrix presented some challenges. One such issue was the over filtering of the data and mapping gene symbols to ENSG ids. After multiple attempts to try and remedy the issue, we elected to use the sample data provided. This could be the reason for the variations in results obtained from our experiment compared to those in the Baron *et al*. (2016) paper. The heatmap produced in Figure 10, shows distinction in gene expression in clusters 3-12, similar to the heatmap shown in the Baron *et al*. (2016) paper. Clusters 0-2 are less defined and have a fair amount of overlap in gene expression throughout.

The gene set enrichment analysis really tells a different story. By searching for gene symbols used in Figure 1 of the reference paper which were used to represent 15 different cell clusters and 13 different cell types in Pancreas and cross validating with marker gene alias in the supplementary table provided by authors related to Figure 1, the current analysis failed to directly identify GHRL gene, stands for epsilon cells, VWF gene, stands for vascular cells, RGS5 gene, stands for quiescent stellate cells, PDGFRA gene, stands for activated stellate cells, SOX10 gene, stands for Schwann cells, and TPSAB1, stands for Mast cells. Only 9 cell types were included before the GO analysis. By choosing a more confidential filter on Fold Change and adjusted p-values, only marker genes representing 6 cell types were included in 404 significantly differentially expressed genes. GO analysis for significant marker genes was quite noisy because most of them contribute to cell cycle and cell proliferation.

However, Figure 12 summarizes genes categorized by tissue-specific or time-specific patterns of genes, as many can be identified as housekeeping genes. In this case only 2 sets relating to Pancreas have genes assigned. This can only support the existence of alpha and beta cells which contribute to 90% components of Pancreatic islets. The top hit of disease pathway in Figure 11 links to chronic pancreatitis. Fibrosis is a prominent feature of chronic pancreatitis and of the desmoplastic reaction linked with pancreatic cancer. While the pathogenesis of fibrosis remains elusive, the activation of stellate cells contribute to pancreatic fibrosis. In healthy tissue, this pathway supports the existence of quiescent (inactivated) stellate cells which compose very small amounts of total cell mass but lack evidence of presence in this study. Following a similar logic, the absence of Mast cells in healthy tissue was expected for Pancreatic Mast cells playing an important role in triggering the local and systemic inflammatory response in the early stages of acute pancreatitis. In Figure 11 and Figure 12, gene sets related to lung tissue and lung disease were identified in spite of lacking Pancreas Mast cells because lung mast cells are not directly involved in the inflammatory response related to pancreatic damage[9].

Conclusion

In conclusion, eight different cell types presented in Figure 1 of the by Baron *et al*. (2016) journal article were identified using RNA-seq analysis and the Seurat R package. Due to data processing difficulties, the sample data set was used for RNA-seq analysis which produced slightly different results than those published in Baron et al. Gene ontology analysis confirmed that alpha and beta cells contribute to 90% of components of pancreatic islets. In terms of sources of error or difficulty for the project, as stated above, processing the UMI matrix and matching the gene symbols to the ENSG ids was a major hurdle to overcome, and the workaround used did alter the group's overall results. Additionally, the time required for processing the initial datasets also affected the overall project; the major issue for the initial data curation was formatting code to automate the parsing of text files for the barcodes. This process took a while to perform, and reduced time that the group had to work on the actual dataset for the final analysis.

References:

1. Kimmel, R.A., and Meyer, D. (2010). Molecular regulation of pancreas development in zebrafish. Methods Cell Biol. *100*, 261–280.
2. Whitcomb, D.C., and Lowe, M.E. (2007). Human pancreatic digestive en- zymes. Dig. Dis. Sci. *52*, 1–17.
3. DJ. The role of gut hormones in glucose homeostasis. J Clin Invest. 2007; 117:24–32. [PubMed: 17200703]
4. Mastracci TL, Sussel L. The endocrine pancreas: insights into development, differentiation, and diabetes. Wiley Interdiscip Rev Dev Biol. 2012; 1:609–628. [PubMed: 23799564]
5. Baron M, Veres A, Wolock SL, et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst*. 2016;3(4):346–360.e4. doi:10.1016/j.cels.2016.08.011
6. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018 Jun;36(5):411-420. doi: 10.1038/nbt.4096. Epub 2018 Apr 2. PMID: 29608179; PMCID: PMC6700744.
7. 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
8. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019 Apr 3;10(1):1523. doi: 10.1038/s41467-019-09234-6. PMID: 30944313; PMCID: PMC6447622.
9. Lopez-Font I, Gea-Sorlí S, de-Madaria E, Gutiérrez LM, Pérez-Mateo M, Closa D. Pancreatic and pulmonary mast cells activation during experimental acute pancreatitis. World J Gastroenterol. 2010 Jul 21;16(27):3411-7. doi: 10.3748/wjg.v16.i27.3411. PMID: 20632444; PMCID: PMC2904888.
10. Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng,Andrew Butler, Maddie Jane Lee, Aaron J Wilk, Charlotte Darby, Michael Zagar, et al.Integrated analysis of multimodal single-cell data.bioRxiv, 2020.

11. Charlotte Soneson, Michael I Love, and Mark D Robinson. Differential analyses for rna-seq: transcript-level estimates improve gene-level inferences.F1000Research, 4, 2015.
12. Johannes Rainer. Ensdb.hsapiens.v79: Ensembl based annotation package. 2017. R Package version 2.99.0.
13. 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.
14. Zhou Y, Park SY, Su J, Bailey K et al. TCF7L2 is a master regulator of insulin production and processing. *Hum Mol Genet* 2014 Dec 15;23(24):6419-31. PMID: 25015099
15. *Genecode*, EMBL-EBI, www.gencodegenes.org/human/.
16. Srivastava A, Malik L, Smith T, Sudbery I, Patro R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. Genome Biol. 2019 Mar 27;20(1):65. doi: 10.1186/s13059-019-1670-y. PMID: 30917859; PMCID: PMC6437997.
17. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017 Apr;14(4):417-419. doi: 10.1038/nmeth.4197. Epub 2017 Mar 6. PMID: 28263959; PMCID: PMC5600148.