Single Cell RNA-Seq Analysis of Pancreatic Cells

Data Curator: Rizky
Programmer: Ally
Analyst: Reshma
Biologist: Aneeq

INTRODUCTION (Shared)

The pancreas plays an integral role in energy homeostasis via hormone regulation and digestive enzyme secretion (Baron, 2016). It consists of acinar, duct cells, and islets, the latter of which contains a number of endocrine cells such as alpha, beta, gamma, delta and epsilon cells. This organ has been implicated in severe human disease including diabetes and pancreatic cancer due to its dysfunctions. However, due to its complex composition of cells with varying gene expression levels, much is still unknown about these endocrine cells, in particular, beta islets for diabetes. Yet, previous research has only produced gene expression profiles using bulk mixtures. In their paper, Baron et. al. (2016) employed the use of a single cell sequencing technique called InDrops to sequence over 12,000 cells from 2 mouse and 4 human pancreatic samples. This allowed them to detect and systematically elucidate the pancreatic cell types with their functions in order to understand their contributions to disease status.

Pancreatic studies have been predominantly done in mice as an appropriate model system (Herreros-Villanueva, 2012). Despite the importance of inter-species relatedness in anatomical structure being of high significance, there have been notably differing reports on the organization of these islets. This emphasizes the importance of conducting this single cell study in order to identify variations between species as well as inter-species differences.

In this study, we attempt to replicate the primary findings of the Baron et al. paper with Salmon and Seurat packages to characterize the diversity of cell types as well as its subpopulations in the pancreas (2016).

DATA (Data Curator)

Run	Bytes	GEO_Accession	Organism	Age	Sex	BMI
SRR3879604	20.98 Gb	GSM2230758	Homo sapien	51	Female	21.1
SRR3879605	14.64 Gb	GSM2230758	Homo sapien	51	Female	21.1
SRR3879606	13.69 Gb	GSM2230758	Homo sapien	51	Female	21.1

Table 1. Sample metadata used for downstream salmon alevin analysis.

Three samples of interest were accessed by referencing the Baron et al. (2016) paper for a GEO accession number link. From this link, we navigated to the SRA Run Selector link to pull the run data associated with the 51 year old female subject with a BMI of 21.1 as depicted in **Table 1**.

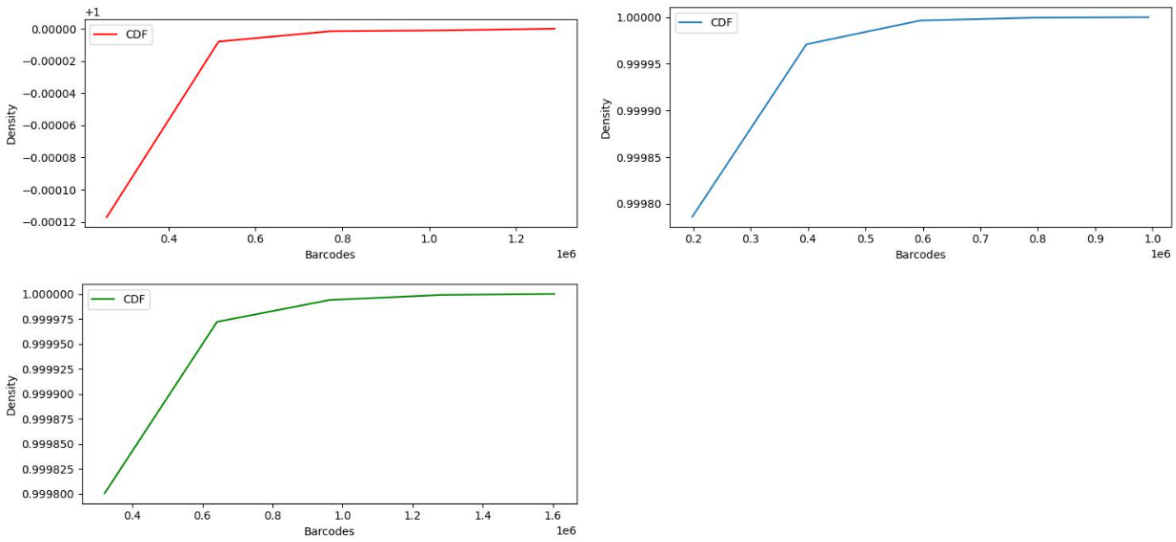


Figure 1. Cumulative distribution plots of demultiplexed barcodes (no UMIs). The X-axis denotes number of barcodes while the Y-axis denotes the density distribution. Red denotes SRR3879606, blue denotes SRR3879605, green denotes SRR3879604.

Extracting barcodes from SRA run data using python

In order to extract informative barcodes from the single cell run files, we used python to parse each unzipped run file. By reading through the file line by line and extracting the barcode and assigning it to a dictionary key, where the values were the barcode counts, we were able to generate a dictionary containing unique barcodes and their respective frequencies. Herein, we take advantage of an inherent property of python dictionaries, where every key entry in a dictionary is unique, with no duplicates.

Whitelisting barcodes based on frequency in python dictionary

Running salmon alevin for this particular recreation of Baron et al. work required a list of whitelisted barcodes. Generating such a list required some sort of filtration step to only include informative barcodes that appear with high frequency. In order to better visualize the distribution of reads among the unique extracted barcodes before we set some sort of filtration threshold, we created cumulative distribution plots of the demultiplexed barcodes for the three samples (**Figure 1**). Admittedly, the distribution was not as informative as we had hoped. As such, we arbitrarily went with a filter that only included barcodes within the 95th percentile of frequencies– i.e. our whitelisted barcodes were expressed more frequently than 95% of the other barcodes.

Using salmon alevin to generate UMI counts matrix

Using these whitelisted barcodes we ran the salmon alevin command. For sample run SRR3879604, there were a total of 5.64e8 reads with an overall mapping rate of 47%. This is in contrast to SRR3879605 and SRR3879606, which had 3.93e8 and 3.68e8 total reads respectively. Not only were the total number of reads lower but so too were the mapping rates. The mapping rate for SRR3879605 was 40.6%, while the mapping rate for SRR3879606 was 41.2%.

Setbacks and adjustments

Initially we encountered several setbacks such as transcript to gene map formatting issues and improper indexing due to neglecting to include the 'genbank' argument when running the 'salmon index' command. However, after working out all these issues we were able to successfully map the run data and generate the UMI counts matrix. Note that the UMI counts matrix was not used for downstream analysis due to time constraints.

METHODS (Shared)

Quality Control and Filtering

The precomputed UMI matrix generated from the salmon software was imported with the R package, tximport (version 1.22.0) (Soneson, 2015), creating a data frame with rows as features and columns as samples, or in this case, cells. Due to the use of Ensembl Gene Identifiers as row names in the features of this matrix, the Ensembl gene IDs were mapped to their respective symbols using the Bioconductor R package, BiomaRt (version 2.48.3) (Durinck, 2009) with the 'hsapiens_gene_ensembl' dataset (human genome version GRCh38.p13), excluding those that either do not map to this genome or are duplicated. After mapping, the matrix was made into a Seurat object with the Bioconductor R package, Seurat (version 4.1.0) (Hao, 2021), with a filter of 3 cells per gene and 50 features per cell.

In determining the number of features to filter at this step, the data was visualized for its complexity (adopted from "Single-cell RNASeq Demo", 2020) as observed in **Figure 2**. Cells in this plot were ranked by the number of genes per cell and thus, its distribution on the plot is a measure of sample quality, noting that there are more cells with higher gene counts. Since the lower inflection point is at approximately 50 genes per cell, 50 was used for the minimum features. The minimum of 3 cells per gene is arbitrary as long as it is non-zero, but to capture genes that are highly expressed in certain cell types, 3 should be sufficient. These numbers were set low to preserve the bulk of the samples for further downstream filtering of damaged, dead, or otherwise low-quality cells. The following quality control as well as downstream dimensionality analysis and cell clustering were adapted from a tutorial ("Seurat - Guided Clustering Tutorial", 2020) from the original authors (Satija, 2015).

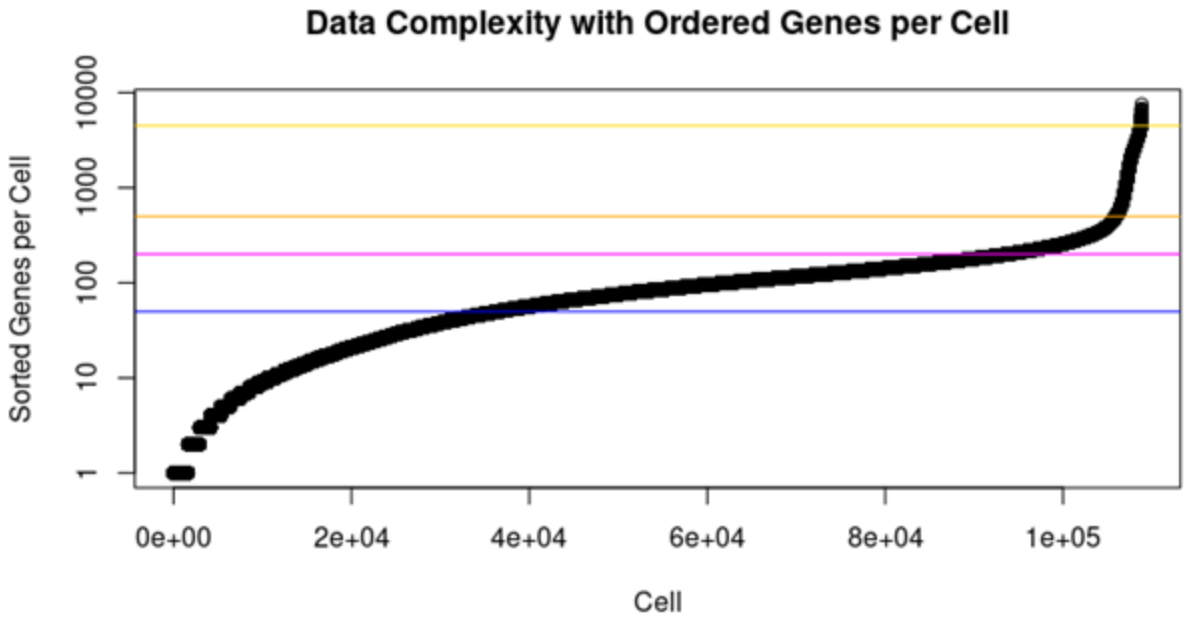


Figure 2. Data complexity with ordered genes per cell. The x and y axis represents the cells in our dataset (108832) and the sorted genes per cell, respectively. Blue line denotes the minimum cells per gene filter used at the first inflection point of approximately 50. Magenta denotes the default filter of 200 provided by the tutorial. The orange and gold line represents the features filtered at 500 and 4500, respectively.

The purpose of the next set of filters was to remove low quality cells, often marked by lower gene counts or extremely high feature counts, which can result from potential cell doublets. Lower quality cells are also marked by higher percentage of MT contamination, as it is an indication of cell stress, damage, or death (Ilicic, 2016). However, since pancreatic alpha and beta cells have naturally occurring mitochondrial gene activity due to its role in metabolism (Maechler, 2010 and Medini, 2021), the mitochondrial percentage filter will take this into consideration. Likewise, capturing cells that have high feature counts is also important as that likely corresponds to presence of unique genes expressed in the cell.

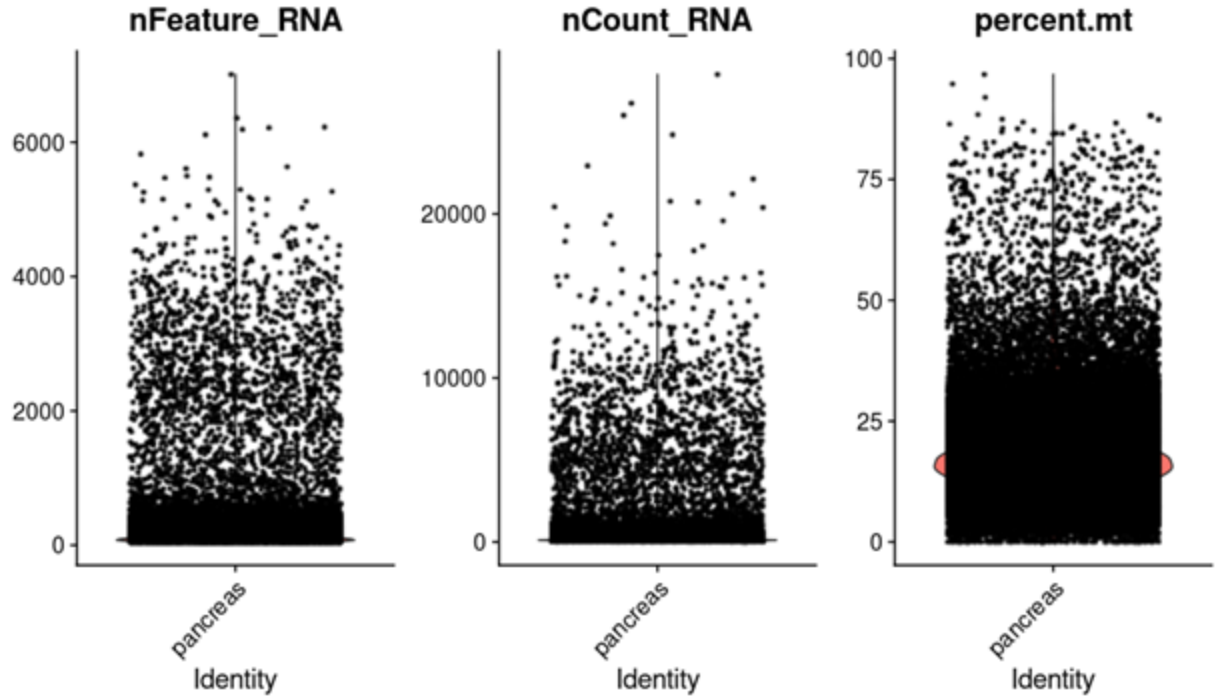


Figure 3. Violin plot for QC metrics of Seurat object. Each subplot shows the distribution of cells that correspond to the number of features (nFeature_RNA), counts (nCount_RNA), and percentage of genes that map to the mitochondrial genome (percent.mt).

In the standard preprocessing workflow of selection and filtration of cells, QC metrics showing the distribution of features, counts, and the percentage of cell counts that match to mitochondrial genome (MT) were implemented and visualized as a violin plot (**Figure 3**) and a scatterplot (**Supplemental Figure 2**). The percent of MT contamination was manually added with the Seurat function, PercentageFeatureSet(). After visual inspection, most of the lower quality cells are bulked at less than 500 features, but most appear to be below 4500, so the feature cutoff was set to > 500 and < 4500 features. According to the violin plot, most cells have up to approximately 30-40% mitochondrial genomic material, so in order to avoid potential loss of healthy pancreatic cells, the %MT will be filtered on < 30%.

After filtering out unwanted cells, the counts matrix was normalized before comparison using a 'LogNormalize' method implemented in Seurat with a default scale factor of 10,000. Feature selection was then performed with Seurat function FindVariableFeatures() with a method of 'vst' to account for the mean variance relationship that is inherent in single-cell data. This function returns a variable feature of 2000 features per dataset by default. Post normalization, the data was scaled as a necessary step prior to dimensionality analysis. Scaling was performed using Seurat function ScaleData(), which will shift the mean expression of the genes across cells to 0 with a variance of 1 across cells. This step is meant to give equal weight to all data in downstream analyses so that highly expressed genes do not dominate.

Dimensionality Reduction Analysis

Linear dimensionality reduction with principal component analysis (PCA) was then performed using RunPCA() from Seurat. Visualization of cells and features were performed using VizDimReduction and DimPlot to name a few methods (**Supplemental Figure 4**). To determine the dimensionality of the dataset, the first 20 principal components were plotted in an elbow plot (**Figure 4**) as well as a JackStrawPlot (**Supplemental Figure 5**). The JackStrawPlot uses a JackStraw procedure that permutes 1% of the data (by default) and reruns PCA to aid in the identification of significant PCAs that have a strong enrichment of p-values. Based on the Elbow Plot, the observed 'elbow' is around PC10, inferring that the majority of variance or signal can be explained by the first 10 principal components. This is confirmed by the JackStrawPlot, which has a sharp dropoff of p-values after PC10.

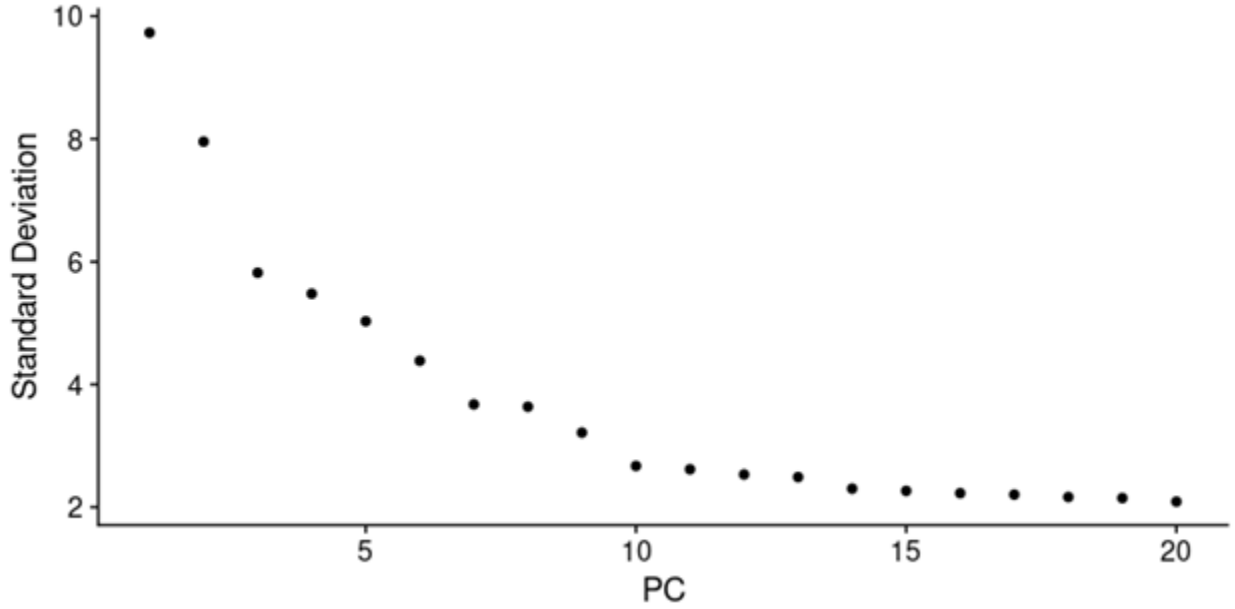


Figure 4. Elbow Plot of the first 20 PCAs in determining the dimensionality of the dataset.

Cell Clustering with UMAP

Cell clustering was then performed using FindNeighbors() followed by FindClusters() from Seurat. The former uses K-Nearest Neighbor (KNN) graph to determine the Euclidean distance in the PCA space using the previously defined dimensionality of the first 10 PCs. To visualize these clusters in two-dimensions, a non-linear dimension reduction technique with Uniform Manifold Approximation and Projection (UMAP) was performed using RunUMAP() from Seurat.

Identifying marker genes and cell type for each cluster

Seurat package was utilized to identify the differentially expressed genes defining each of the 13 (0-12) clusters provided in the sample RDS file. The function *FindAllMarkers* was used with 0.25 threshold to obtain all the markers without filtering out significant markers.

List of identified cell types with their respective markers provided by Baron *et al.* (2016) was referred for assigning the cell type. Since the chosen threshold for log2FC was 0.25, we further assume that if a known cell type's gene marker with minimum log2FC of 0.25 is detected, the corresponding cluster(s) could be labeled as such cell type. Therefore, all the known markers were searched in the DEG list and the clusters were labeled with the corresponding cell type.

DAVID for gene enrichment analysis

The provided clustered sample data was first filtered in R to select ten genes from each cluster with the lowest adjusted p-values. Ties in values were included to allow for the maximum number of genes to be selected. Subsequently this gene list was exported to DAVID and functional annotation clustering performed using default settings to identify the different cell types.

RESULTS (Programmer/Analyst/Biologist)

	Features	*Cells*
Unfiltered Dataset	60233	108832
Post Mapping with Duplicates	39470	108832
Post Mapping without Duplicates	39456	108832
Seurat Object (>3 cells per gene, >50 genes per cell)	21347	71126
Filtered for low quality cells (500 < features < 4500, % MT < 30%)	21347	2608
Features after Variance filtering		
2000 (by default)		
Total Number of Clusters		
9		

Table 2. Summary of number of genes and cells in each step of the preprocessing steps and the resulting number of clusters from the Seurat object.

Table 2 shows the number of features and cells in each step of the preprocessing steps with Seurat. With the unfiltered dataset, we have 60,233 features (genes) and 108,832 cells. After mapping to their gene IDs, 20,763 genes were filtered out, leaving 39,470 features, and when accounting for 14 duplications, 39,456 features were left before creating the Seurat object. After applying a filter for a minimum of 3 cells per gene and 50 genes per cell in creating the Seurat object, "pancreas", another 18,109 features and 37,706 cells were filtered, leaving 21,347 features and 71,126 cells. Cells were then filtered for low quality cells at > 500 and < 4500 features, leaving a total of 2,608 cells for downstream analysis. Before downstream analysis, the data was also filtered for variance, but as the default procedure in Seurat returns 2,000 most highly variable features for each dataset (Hao et al, 2021), this was the number that was returned in this study (**Supplemental Figure 3**).

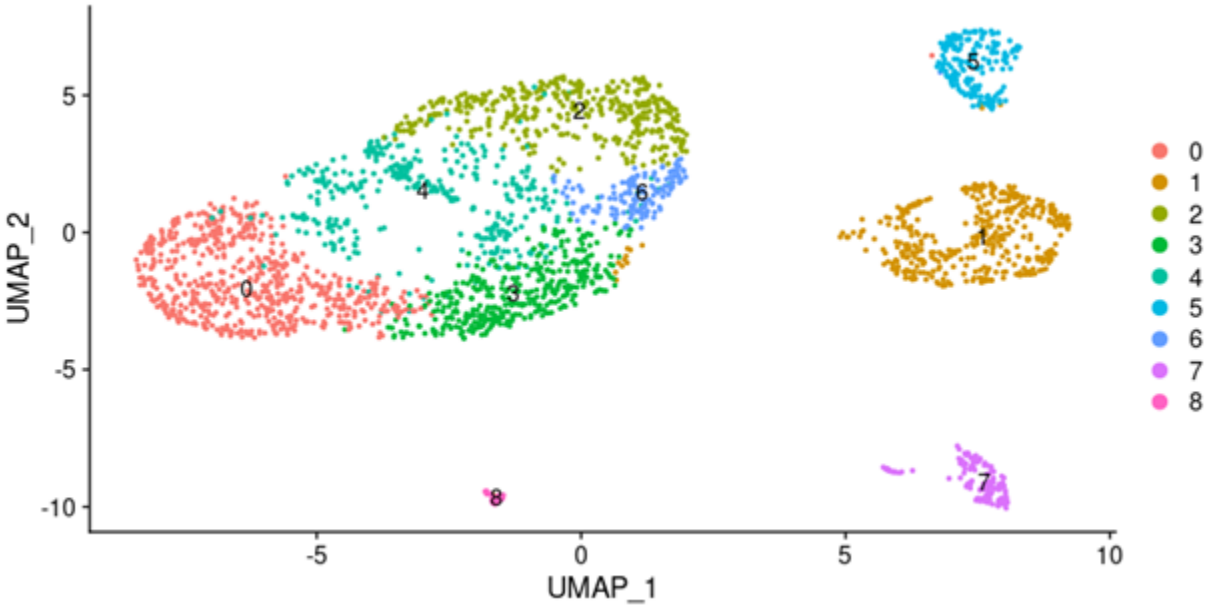


Figure 5. UMAP visualization of PCA clustering. A total of 9 clusters were identified.

After data normalization, scaling, and dimensionality reduction analysis with PCA, a preliminary UMAP was plotted and showed a total 9 clusters (**Figure 5)**. The relative proportions of cell numbers in each cluster are shown as a pie chart in **Figure 6**. Cluster 0 has the highest proportion of cells at 23.35% while the lowest, cluster 8, make up only 1.07% of the cells.

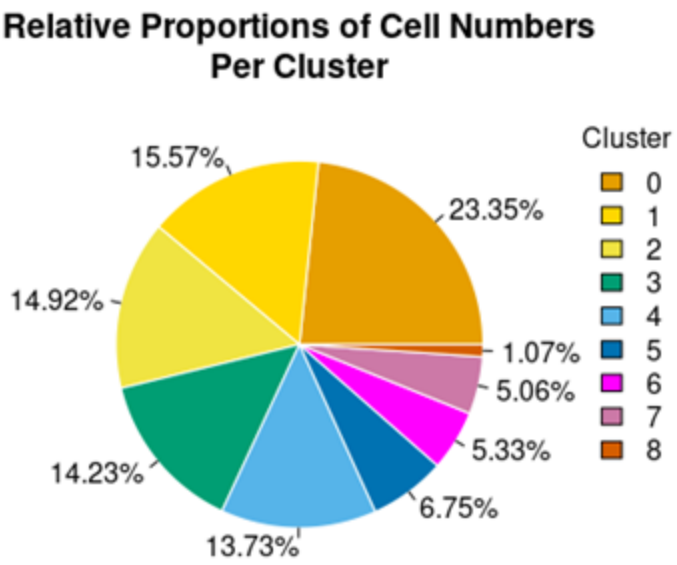


Figure 6. Pie chart of the relative proportions of cell numbers for each identified cluster from the UMAP.

The Seurat function FindAllMarkers() compared genes in each cluster with all the clusters, identifying 6,487 DEGs as potential gene markers for all clusters. The top 2 DEGs with highest log2 fold change for each cluster were reported as cluster-defining genes in **Table 3**. The top 10 DEGs with highest log2 fold change are also represented using a Heatmap in **Figure 7**.

	Cluster	Gene	Avg. Log 2 FC	Adj. P-value
1	0	SP100	1.388	3.027×10^{-118}
2	0	PRRG3	1.363	2.583×10^{-83}
3	1	INS	1.757	0
4	1	DLK1	1.805	3.272×10^{-224}
5	2	FN1	1.950	3.480×10^{-219}
6	2	COL1A1	1.626	1.504×10^{-171}
7	3	REG1B	3.566	0
8	3	REG1A	3.545	0
9	4	TTR	2.375	0
10	4	GCG	2.245	0
11	5	KRT19	2.734	3.505×10^{-282}
12	5	CXCL1	3.051	1.818×10^{-191}

13	6	EEF1A2	1.644	0
14	6	INS	2.183	9.497×10^{-290}
15	7	ACER3	2.602	1.842×10^{-128}
16	7	AL022322.2	2.588	1.289×10^{-87}
17	8	GC	2.094	0
18	8	TTR	2.880	1.260×10^{-277}
19	9	COL1A1	4.494	0
20	9	COL3A1	4.187	0
21	10	CRP	2.980	0
22	10	KRT18	2.744	0
23	11	ALDOB	3.967	0
24	11	PRSS2	3.953	4.735×10^{-166}
25	12	ACP5	5.644	0
26	12	AC007192.1	5.127	2.519×10^{-247}

Table 3. Top 2 gene markers for each cluster based on log2 fold change.

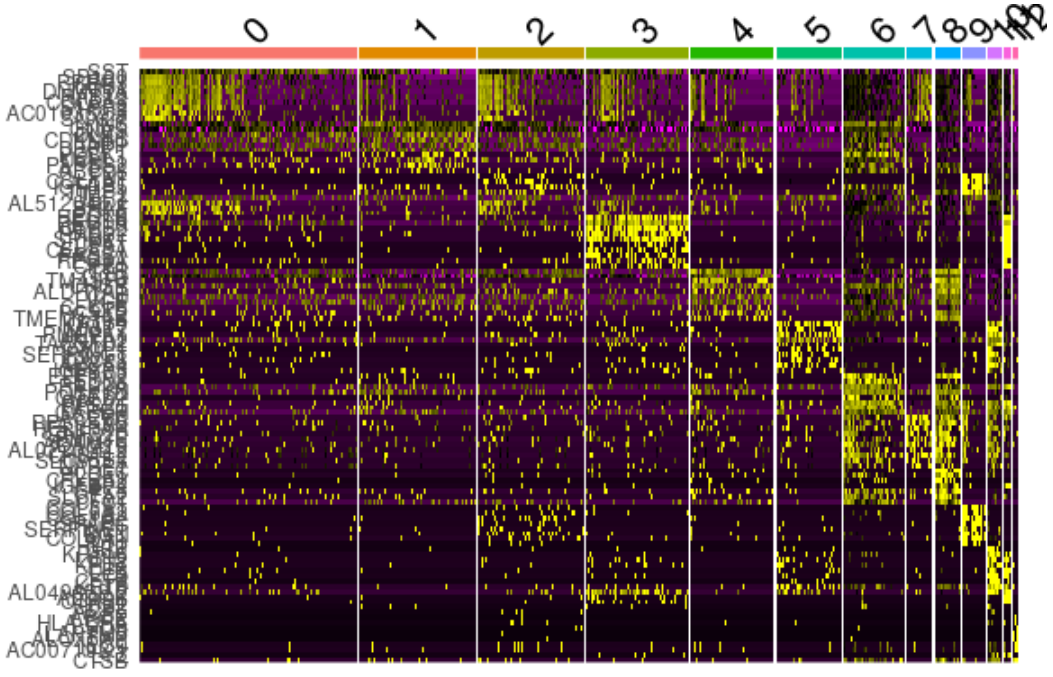


Figure 7. Top 10 gene markers with the highest log2 fold change for each of the 13 cell clusters.

Labeling cell clusters based on expression of marker genes

In order to study the distribution of the gene markers, UMAP figures were produced for each marker to help determine cell types (**Figure 8**). The UMAPs show that some known markers are highly expressed in multiple clusters while some other markers show low expression in all clusters (**Table 4)**.

The results show zero clusters for cell types epsilon, quiescent stellate, activated stellate, endothelial, cytotoxic and mast (**Table 4**). UMAP was used to visualize the clustered cell with labeled cell types (**Figure 9**), except for cluster 7 and 9 because there are no known markers found (**Table 4**).

A heat map was generated for the log normalized UMI counts for the top 2 marker genes of each cluster (**Figure 10)**. It does not suggest clear distinct cell types, which is consistent with **Figure 7** using log2FC as expressions, because some markers are highly expressed in multiple clusters while some markers could not represent the whole cluster well.

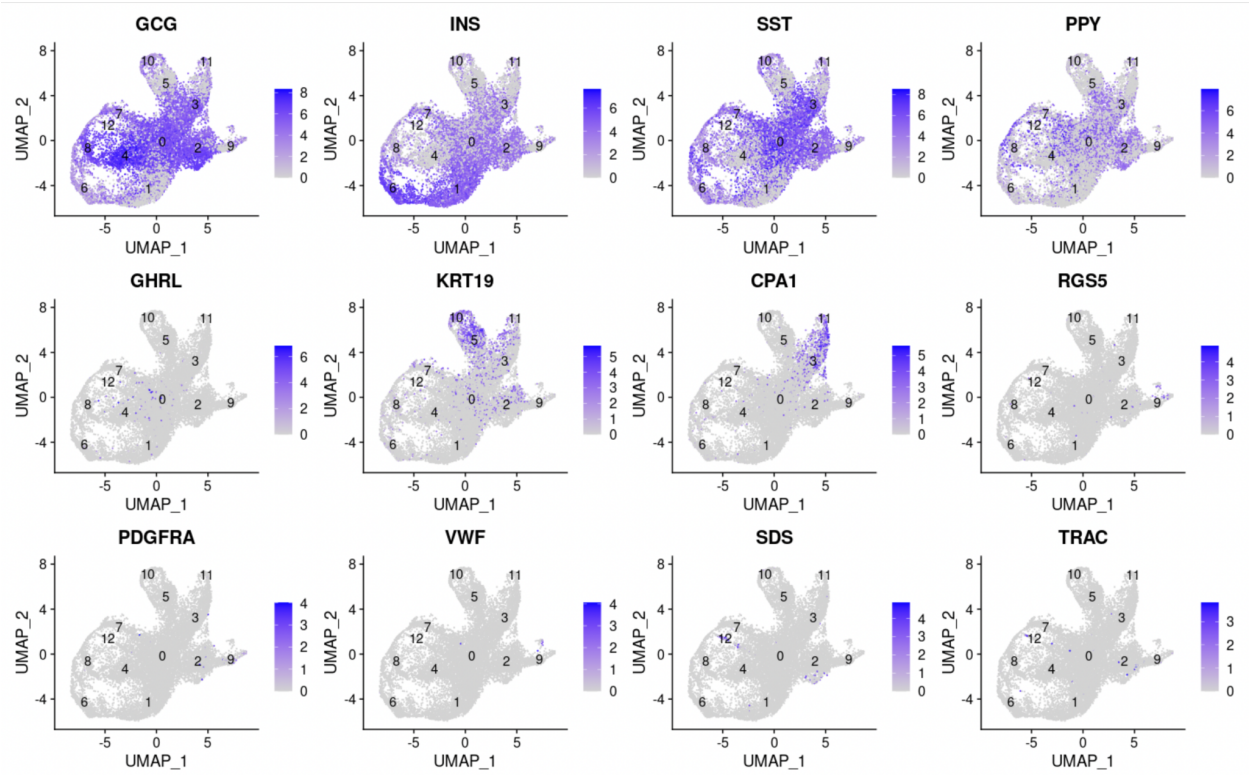


Figure 8. UMAPs for all known marker genes representing expression levels in different clusters. The darker the purple, the greater the expression of the marker gene.

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

Table 4. Known markers and their corresponding clusters identified for each cell type.

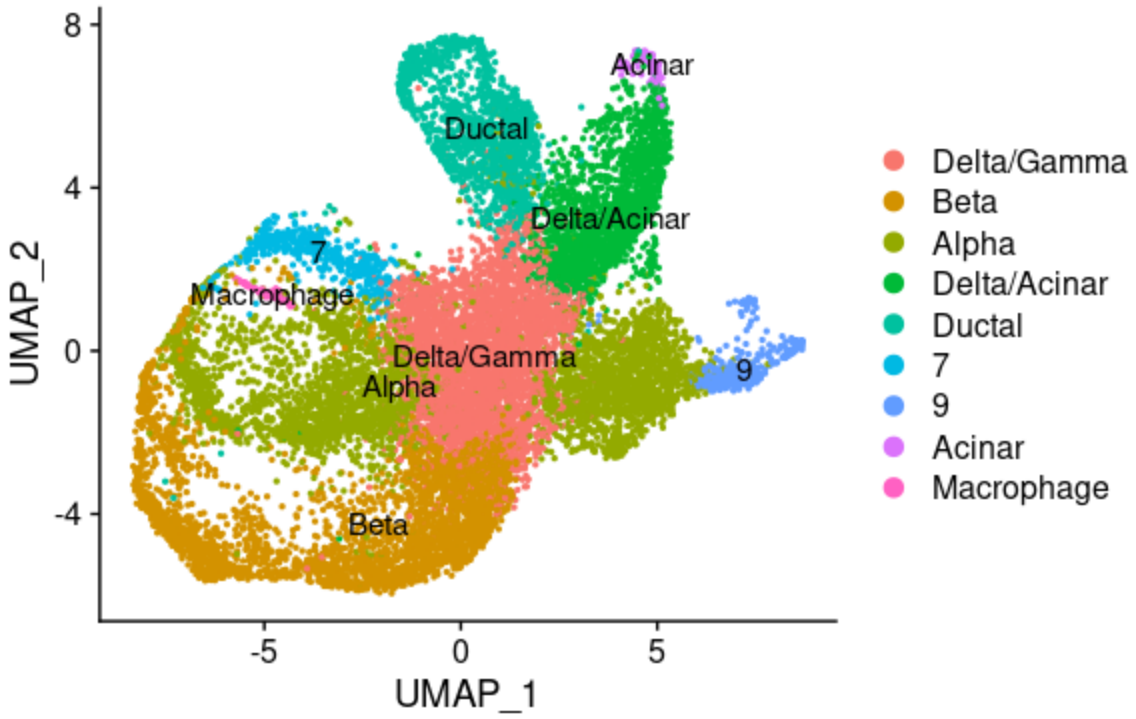


Figure 9. UMAP for clustered cells with labeled cell types. 7 and 9 are still cluster IDs which couldn't be labeled by known markers.

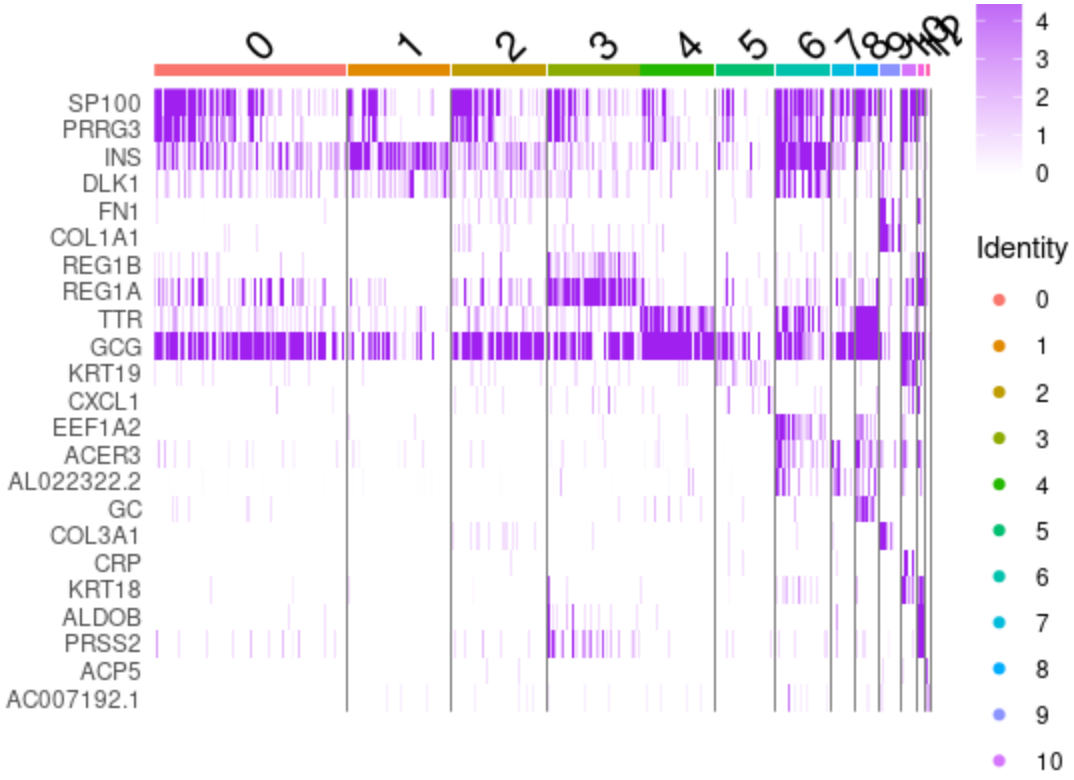


Figure 10. Heatmap of log normalized UMI counts for top 2 marker genes in each cluster.

A total of 1079 genes were selected for gene enrichment analysis using the default settings of DAVID [2],[3]. Although only 12 clusters were present, more genes were selected due to the allowance of tied values. Functional annotation clustering resulted in the identification of 158 clusters. Of these, the top 10 are summarized below in **Table 5**.

S. No.	Cluster Term	Count	p-value	Adj. P-value
1	Extracellular space	198	1.00×10^{-26}	3.70×10^{-24}
2	Protein digestion and absorption	29	5.10×10^{-11}	1.60×10^{-08}
3	Cytosolic ribosome	20	2.40×10^{-09}	1.50×10^{-07}
4	Heparin binding	18	2.50×10^{-06}	1.30×10^{-04}
5	Transmembrane helix	349	1.00×10^{-04}	9.00×10^{-04}
6	Unfolded protein binding	21	5.10×10^{-06}	4.90×10^{-04}
7	Proteolysis	46	2.00×10^{-06}	4.70×10^{-04}

8	ECM receptor interaction	24	6.40×10^{-09}	1.00×10^{-06}
9	Diabetic cardiomyopathy	37	3.40×10^{-08}	3.60×10^{-06}
10	Viral entry into host cell	16	8.20×10^{-05}	9.30×10^{-03}

Table 5. Top functional annotation clusters identified by DAVID.

DISCUSSION (shared)

In determining the cutoff for the cells and features for the Seurat object, if a filter of 200 features were used (magenta line in **Figure 2**), it appears that approximately half of the cells in the dataset would theoretically have been lost. However, this plot does not provide the full picture. When comparing the violin plots of features and mitochondrial percentage at minimum features of 50 and 200 (**Figure 3, Supplemental Figure 1**), there is a clear tradeoff in filtering for the lower and higher features. Filtering out more cells does leave less cells with an increased mitochondrial contamination, but it also has less cells with a percent.mt of < 5. Likewise for features, at a higher filter, there is a clearer demarcation of lower quality cells and the rest, but there is also an overall marked decrease in cells along the spectrum of features.

In our attempt to replicate the 14 clusters of cells in the original study from this dataset, our UMAP only yielded 9 clusters. A great variability lies in how we filtered the data. In order for a Seurat object to be created, there cannot be any duplications of identifiers in the data, and so 14 of such duplicates were removed. It is unclear which set of counts amongst the duplicates were used for further downstream analysis. Since our clusters are less than that of Baron et al. (2016), it is very possible that too many cells were filtered out that can help provide more information in the dimensionality of the dataset. As mentioned earlier, if more cells were allowed to pass the filters, we may capture more unique expression levels, but also more lower quality cells. It is also quite possible that the filter was not set high enough when filtering for features, so that truly highly expressed cells were not captured.

There is also the possibility of an incomplete dataset as the study indicates that there were mice samples involved. In mapping the dataset to mouse gene symbols, the ensembl query yielded no matches (data not shown, see Programmer code), indicating the absence of mouse data, and possibly more, in this dataset. These can also simply be omitted from our data source.

Alternatively, filtering out mitochondrial contamination at < 30% might be too restrictive as alpha and beta cells, alpha cells in particular, are not clearly separated from the other clusters (**Figure 9**), especially with non-identifiable markers. It just simply implies that there is not enough information to call them. Interestingly, this is very similar to the 1% of cluster 8 in the pie chart. Despite consisting of the smallest % of the pie chart, the p-value in the JackStrawPlot is fairly significant, implying that this cluster of cells may be highly expressed and serve a specific purpose in the pancreas, although smaller in number.

Functional annotation clustering using DAVID found that the most significant cluster had almost 200 genes and was associated with extracellular space. This is likely associated with duct cells as they form ducts in the pancreas through which the exocrine secretions of the acinar cells pass through. However, the annotation term with the most number of genes associated with it was the transmembrane helix. This can be understood to be a result of the secretory nature of the cells of the pancreas. Clusters related to proteolysis, protein digestion and absorption were also observed. These are as a result of the enzymes secreted by the acinar cells that produce proteases among other digestive enzymes. The cytosolic ribosome cluster is likely a result of contamination in the dataset. Correct filtering of the dataset would remove this as it is indicative of stressed or dead cells.

Heparin binding proteins (HBPs) have been known to have complex interactions with the pancreas and are associated with normal functioning of the pancreas- primarily involved in molecular transport and cellular movement[4]. This explains its presence as an important cluster as identified by DAVID. Furthermore, mast cells are also known to be rich in heparin, thus the presence of this cluster indicates the presence of mast cells in the pancreas. The extracellular matrix (ECM)-receptor interaction cluster can be associated with active pancreatic stellate cells as these cells have been shown to secrete a variety of ECM components[14]. Diabetic cardiomyopathy is a secondary disease that occurs in people affected with diabetes that leads to heart failure. The associated annotation cluster is indicative of alpha and beta cells as they are responsible for the production and secretion of glucagon and insulin whose correct secretion prevents the onset of diabetes.

Overall, gene enrichment analysis performed using DAVID confirmed the presence of exocrine (acinar and duct cells), some endocrine cells (alpha, beta and pancreatic stellate cells) and mast cells. We are unsure what the annotations of unfolded protein binding and viral entry to cells indicate.

CONCLUSION (shared)

Our overall attempt to replicate the Baron et al. study (2016) resulted in finding 9 clusters compared to their 14. The big challenge with this hinges on the approaches to filter the data. If the data is filtered too much, as is most likely in our case, the downstream analysis will not have enough differentiating information to get the same number of clusters. It is also important to note that the Seurat Bioconductor package was used in our analysis, but the authors of the original study made their own pipeline to obtain these results, so there are nuances in each step. The approach will be different.

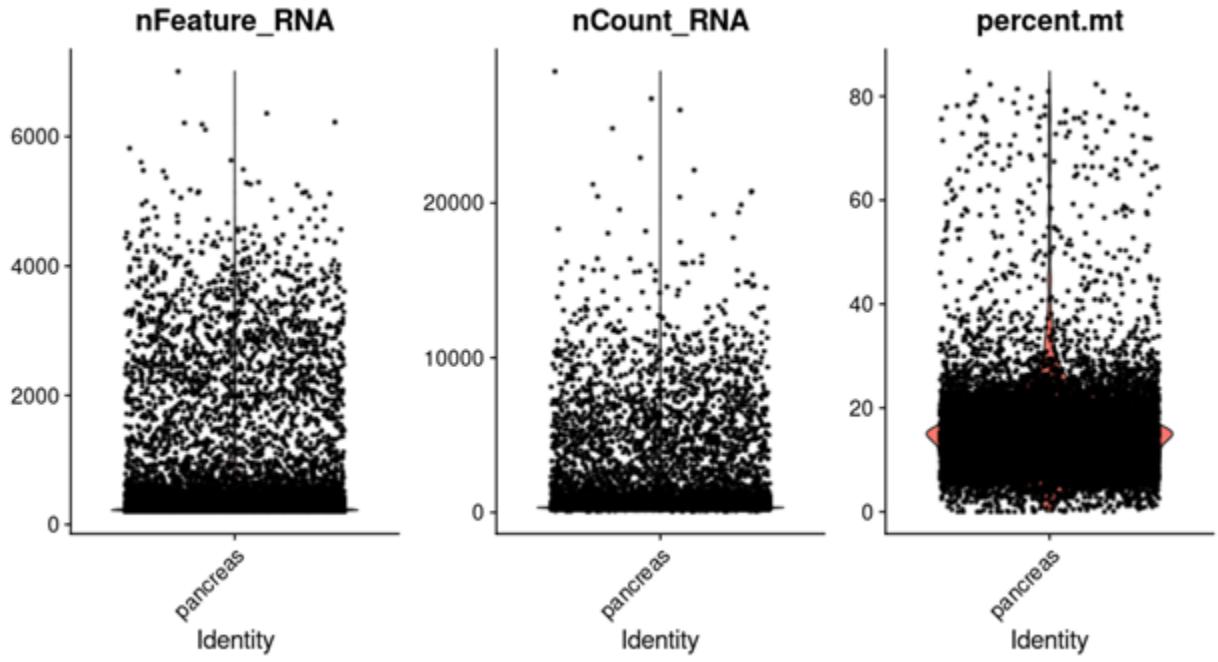
In trying to identify the cells with the help of functional annotation clustering, we were only able to identify 6 of the 12 cell types identified by Baron et al. This is likely due to the fact that the clustering was performed on the sample dataset and not using the results from the Analyst. Rerunning the analysis using the correct data would most probably result in better clustering and identification of more cell types. However, the major cell types were still identified in our analysis with the only exception of delta and gamma cells.

REFERENCES

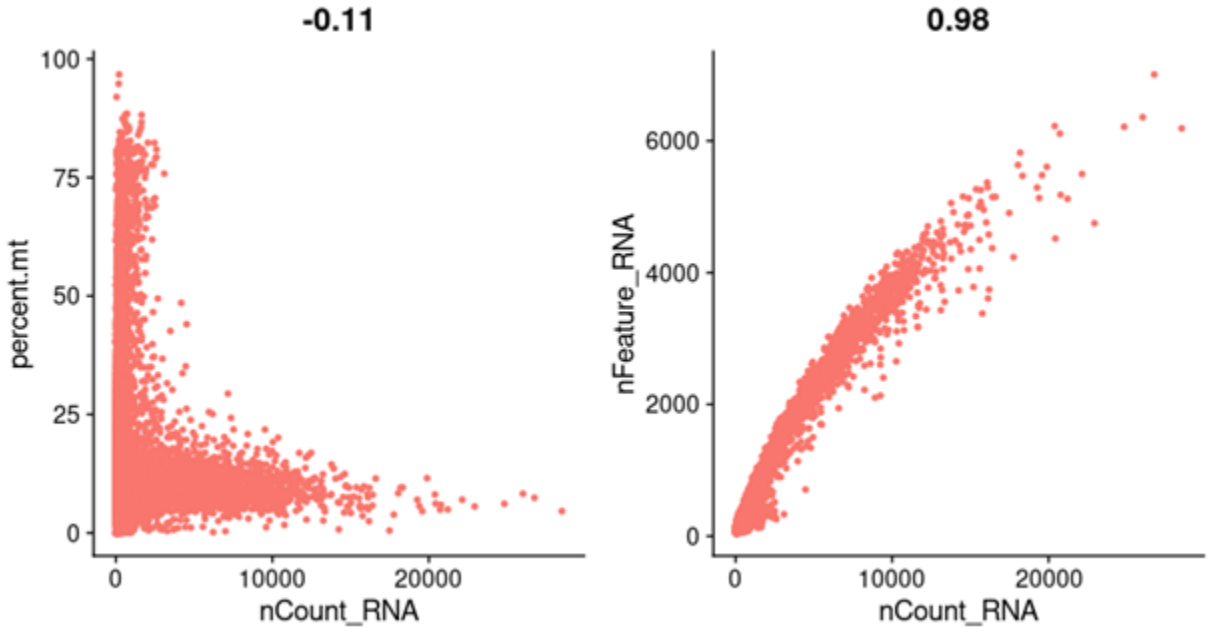
1. Herreros-Villanueva, M., Hijona, E., Cosme, A., & Bujanda, L. (2012). Mouse models of pancreatic cancer. *World journal of gastroenterology*, *18*(12), 1286–1294. https://doi.org/10.3748/wjg.v18.i12.1286
2. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc. 2009;4(1):*44-57.
3. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res. 2009;37(1):*1-13.
4. Nunes, Q. M., Su, D., Brownridge, P. J., Simpson, D. M., Sun, C., Li, Y., ... & Fernig, D. G. (2019). The heparin-binding proteome in normal pancreas and murine experimental acute pancreatitis. *PloS one*, *14*(6), e0217633.
5. Soneson, C., Love, M.I., Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research, 4*(1521) Doi:10.12688/f1000research.7563.1
6. Durinck, S., Spellman, P.T., Birney, E., Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols, 4*(1184-1191).
7. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zagar, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E.P., Jain, J., Srivastava, A., Stuart, T., Fleming, L.B., Yeung, B., Rogers, A.J., McElrath, J.M., Blish, C.A., Gottardo, R., Smibert, P., Satija, R. (2021) Integrated analysis of multimodal single-cell data. *Cell*. Doi:10.1016/j.cell.2021.04.048 url: https://doi.org/10.1016/j.cell.2021.04.048
8. Single-cell RNA-seq Demo (10X Non-Small Cell Lung Cancer) (2020). Retrieved from http://barc.wi.mit.edu/education/hot_topics/scRNAseq_2020/SingleCell_Seurat_2020.html
9. Seurat – Guided Clustering Tutorial (Compiled April 17, 2020). Retrieved from https://satijalab.org/seurat/archive/v3.1/pbmc3k_tutorial.html
10. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., Regev, A. (2015) Spatial reconstruction of single-cell gene expression data. Nature Biotechnology, 33 (495-502) doi:10.1038/nbt.3192 url: https://doi.org/10.1038/nbt.3192
11. Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., Teichmann, S. A. (2016) Classification of low quality cells from single-cell RNA-seq data. *Genome Biology, 17*:1 (29). https://doi.org/10.1186/s13059-016-0888-1
12. Maechler, P., Li, N., Casimir, M., Vetterli, L., Frigerio, F., & Brun, T. (2010). Role of mitochondria in beta-cell function and dysfunction. *Advances in experimental medicine and biology*, *654*, 193–216. https://doi.org/10.1007/978-90-481-3271-3_9
13. Medini, H., Cohen, T., Mishmar, D. (2021). Mitochondrial gene expression in single cells shape pancreatic beta cells' sub-populations and explain variation in insulin pathway. *Sci Rep 11***,** 466. https://doi.org/10.1038/s41598-020-80334-w
14. Buchholz, M., Kestler, H.A., Holzmann, K. *et al.* Transcriptome analysis of human hepatic and pancreatic stellate cells: organ-specific variations of a common transcriptional phenotype. *J Mol Med* **83,** 795–805 (2005). https://doi.org/10.1007/s00109-005-0680-2

15. 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

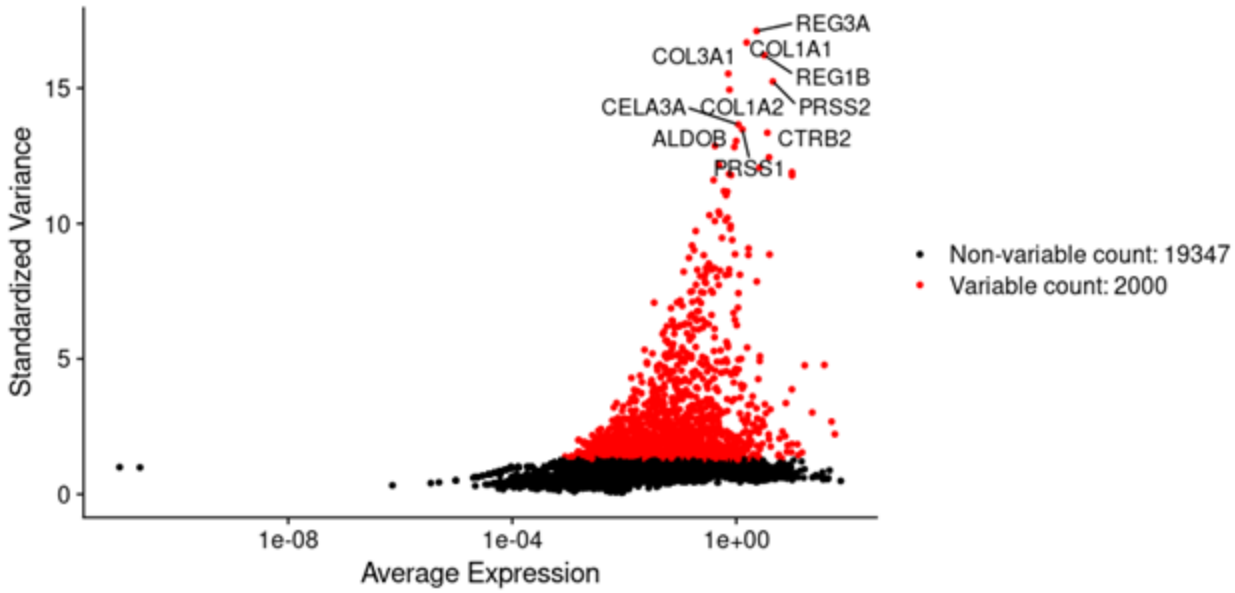
SUPPLEMENTARY FIGURES



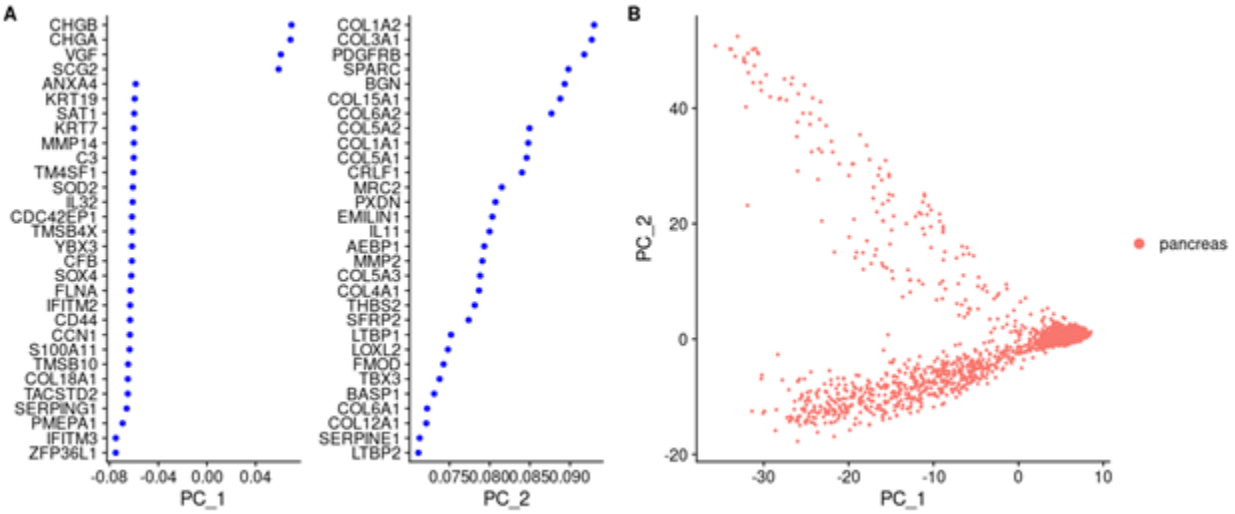
Supplemental Figure 1. Violin plots of QC metrics with a pre-filter of 200 or more features in creating the Seurat object.



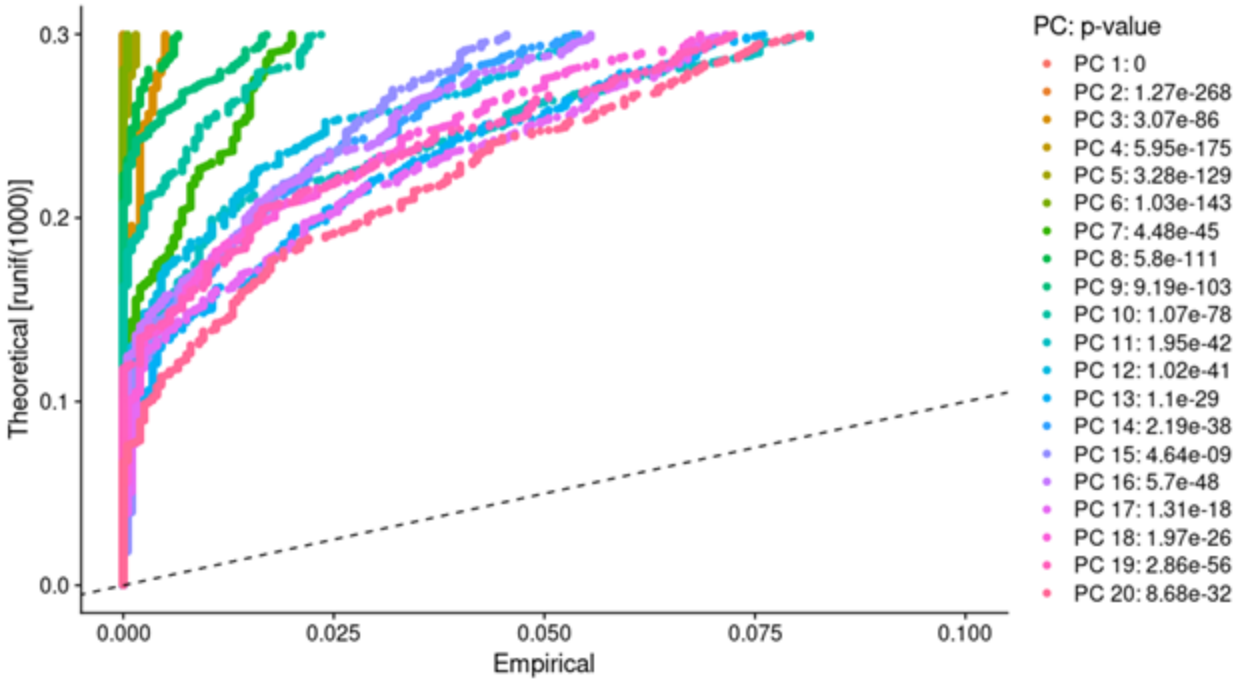
Supplemental Figure 2. Scatterplot showing feature-to-feature relationship between MT contamination and features with counts, respectively.



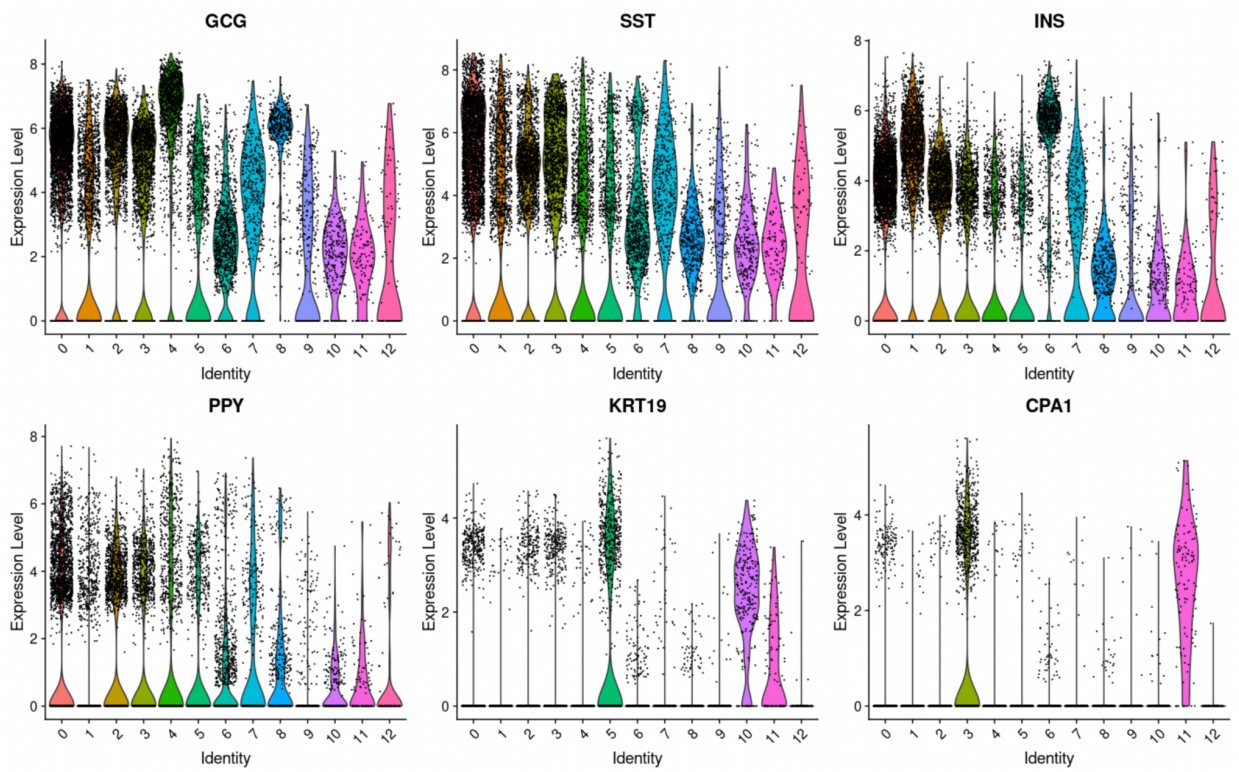
Supplemental Figure 3. Scatter plot of the standard variance of our dataset after normalization with the top 10 most highly variable genes labeled to their respective points.

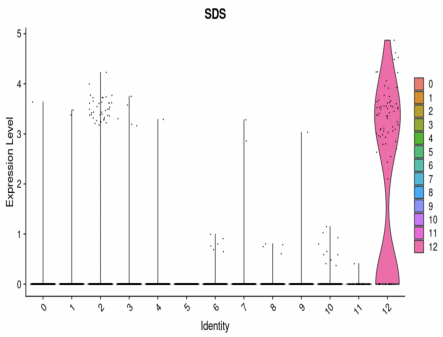


Supplemental Figure 4. Visualization of PCA results with VizDimReduction (A) and DimPlot (B), respectively.

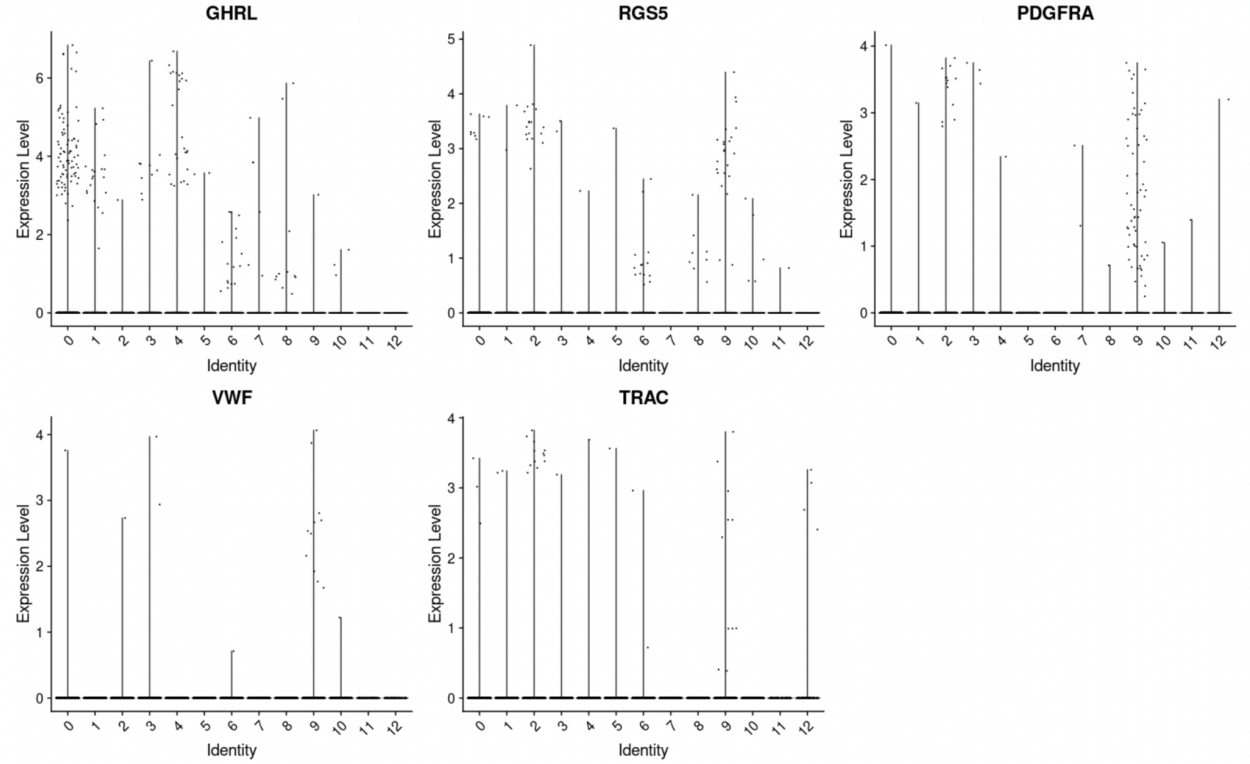


Supplemental Figure 5. JackStrawPlot of the first 20 PCA to determine an appropriate dimensionality for the data.





Supplemental Figure 6. Violin plots for different cell types to compare the clusters



Supplemental Figure 7. Violin plots for different cell types to without any consistent cluster