Project 4: Single Cell RNA-Seq Analysis of Pancreatic Cells

Alec Jacobsen, Daisy Wenyan Han, Divya Sundaresan, Emmanuel Saake
Biologist Analyst Programmer Data Curator

ENG BF528, Spring 2021

Introduction

The mammalian pancreas is comprised of numerous cell types, each of which hosts countless complex interactions that are vital for the functioning of the human body. By developing a deeper understanding of the gene expression profiles and transcriptional activity of each of these cell types, it therefore becomes possible to elucidate information on the regulation of important dysfunctions that give rise to diseases such as Type 1 and Type 2 Diabetes Mellitus. In their 2016 study, *A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure*, Baron et al. implemented a droplet-based single-cell RNA-seq experiment in order to determine the transcriptomes of both human and mouse pancreatic cells. In addition to matching previously characterized cell types, they were able to detect subpopulations of cells with distinct expression profiles. Baron et al. were therefore able to generate a dataset that can serve as a resource for the discovery and analysis of novel, cell-type specific, differential gene expression analysis.

This study was novel in that the availability of single-cell analysis and the inDrop methodology provided Baron et al. the opportunity to construct single-cell transcriptome libraries on a high-throughput scale. While previous studies have been shown effective in categorizing novel cell types (Li et al., 2016), it had proven challenging to scale up this analysis due to difficulties obtaining samples of human cells, as well as developing a system in which a sufficient number of cells can be captured. The inDrop methodology, developed by Klein et al. in 2015, just one year prior to Baron et al.'s study, provided a solution to both of these issues (Klein et al., 2015).

Baron et al. analyzed the pancreatic cells of four individuals. For our analysis, we will be re-examining the data from one such individual in the study, and evaluating the cellular makeup of the corresponding samples in order to corroborate the initial findings using more advanced, updated techniques.

Data

Samples/ Sequencing Libraries

All data used in this project was made available for us on the Boston University Shared Computing Cluster. A total of thirteen sequencing libraries were available, across four individuals. Using the accession number GSE84133 on the NCBI web platform, the metadata corresponding to the dataset was retrieved and processed to generate a shortlist of samples corresponding to a single 51 year-old female donor. There were three samples used as primary sequence libraries for analysis: SRR3879604, SRR3879605 and SRR3879606.

Counting Reads by Barcode and Whitelist Generation

To process the compressed FASTQ files made available to us, we used the `AWK` command. Awk is a programming command that allows for efficient processing of large data files within shell pipes. Using AWK, we extracted and combined the barcodes, which were further processed to generate a frequency count of each barcode.

These frequency counts were then read into R as a data-frame. The mean of frequency of each barcode was computed and applied as a lower-bound filter, thereby generating a whitelist of barcodes with count greater than the mean. Figure 1 shows the cumulative distributive plot of the number of reads per distinct barcode. Normally the 'knee' method would be used to determine the filtering threshold. However the point of inflection of Figures 1A, 1B and 1C is 0, making it unsuitable for filtering.

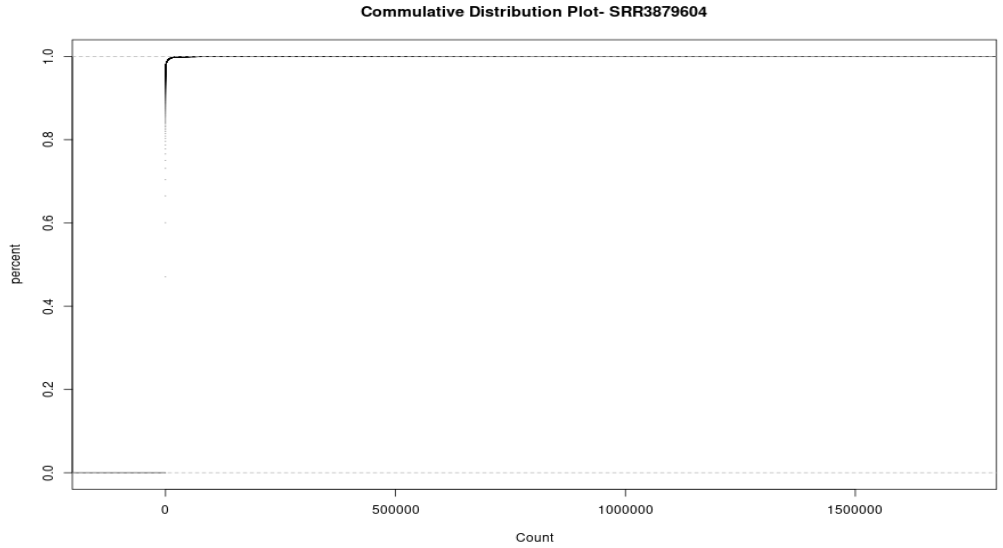


Figure 1A - Cumulative distribution plot of sample SRR3879604

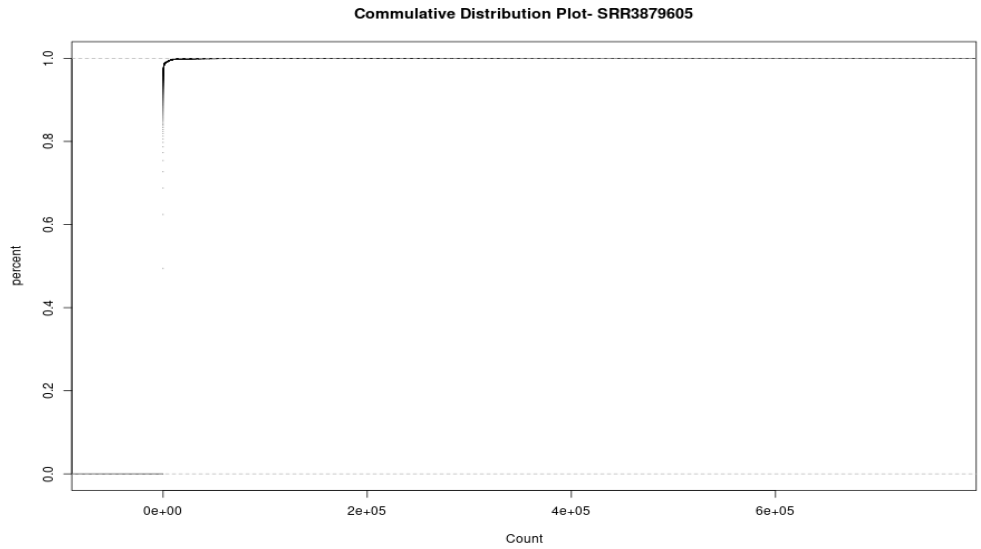


Figure 1B - Cumulative distribution plot of sample SRR3879605

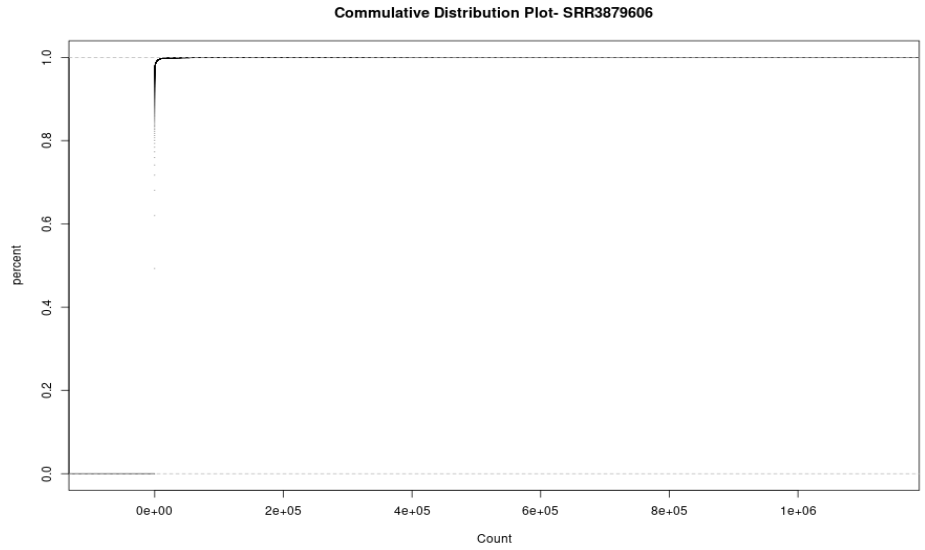


Figure 1C. Cumulative distribution plot of sample SRR3879606

Generation of UMI Counts Matrix using Salmon Alevin

Generation of the UMI count matrix using Salmon was a six step process:

1. Download of gene annotation sequence, transcript sequences and genome sequences from ***gencodegenes.org*** . Details provided in Table 1.
2. Installing Salmon (version 1.1.0) : `conda install --channel bioconda salmon`
3. Extracting the names of genome targets using grep, to be used later for indexing.
4. Concatenate transcriptome and genome reference file for index generation
5. Build indexes using 'salmon index' instruction.
6. Alignment of sequences by running Alevin on index , sample fastq files and whitelist.

Table 1 - List of required sequences for the generation of UMI count matrix

Content (Human)	Region	File type
Comprehensive Gene Annotation	ALL	GTF
Transcript Sequences	CHR	GTF
Genome Sequence (GRCh38.p13)	ALL	Fasta

Salmon Mapping Statistics

The main mapping statistics identified is the mapping rate, with a value of 42.03%. Other secondary numbers were reported in the log files. Table 2 captures a few of these statistics..

Table 2 - Mapping statistics from allevin.log and salmon_quant output files

Variable	Stats_Value	Report Source
Mapping rate	42.03%	salmon_quant
Targets in index	234291	salmon_quant
Discarded reads(Noisy Cellular Barcodes)	13.81%	alevin.log
White-listed barcodes identified	1187046	alevin.log
Unique Barcodes found	4251176	alevin.log

Methods

Data Quality Control Using Seurat

Initial Filtering

The UMI counts matrix generated via Salmon Alevin was then loaded into R as a Seurat object, before undergoing quality control following a tutorial provided by developers of the Seurat package. Filtering out low quality cells was performed using three filters: removal of genes not mapping to known gene symbols, thresholding by percent mitochondrial counts, and setting a minimum and maximum number of features. The raw inputted data contained 60232 unfiltered cells, which were filtered down to 56120 genes by mapping the Ensembl IDs to gene symbols. I originally did not want to filter data based on if there was a mapped symbol or not, but it came down to a decision of either finding MT-genes or using the full geneset.

Ensembl IDs were matched to their respective gene symbols using the EnsDb.Hsapiens.v79 package, available via Bioconductor. Various other packages were attempted for gene symbol matching, such as BiomaRt, and org.Hs.eg.db, but while those allowed us to keep the unmatched IDs, they did not contain any mitochondrial genes. I made the decision that filtering based on mitochondrial genes would be more important later on in downstream analysis. I am aware that of the unmatched ~4000 genes removed from the mapping step, it is possible that we have unknowingly removed some important features. However I believed finding these mitochondrial genes would be more important.

After mapping, we then filtered down to 6028 cells. This was done by making sure the number of features were greater than 200 and less than 4200. We also made sure that the percentage of mitochondrial genes for each cell was less than 20%. This limit of feature number per cell was important because low-quality cells or empty droplets often have few genes, and cell multiplets may exhibit an aberrantly high gene count (Hao et al., 2020). Therefore, based on *Figures 2a*, 2b we determined 200-4200 features per cell seemed appropriate.

The other filter was the percentage of mitochondrial genes. Low-quality or dying cells often exhibit extensive mitochondrial contamination (Hao et al., 2020). A mitochondrial percentage of 20% per cell was chosen as the cutoff, as determined by viewing the violin plot, and scatter plot below in *Figure 2*.

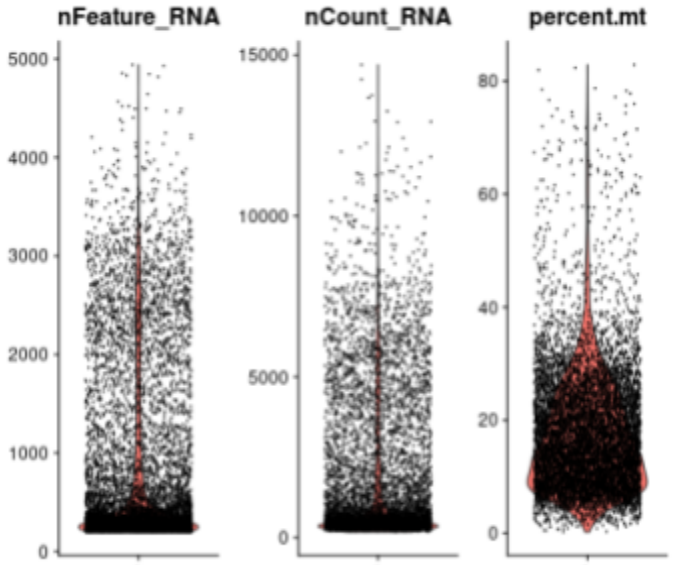


Figure 2A - Distribution of Features, Counts, Percent Mitochondrial DNA for QC

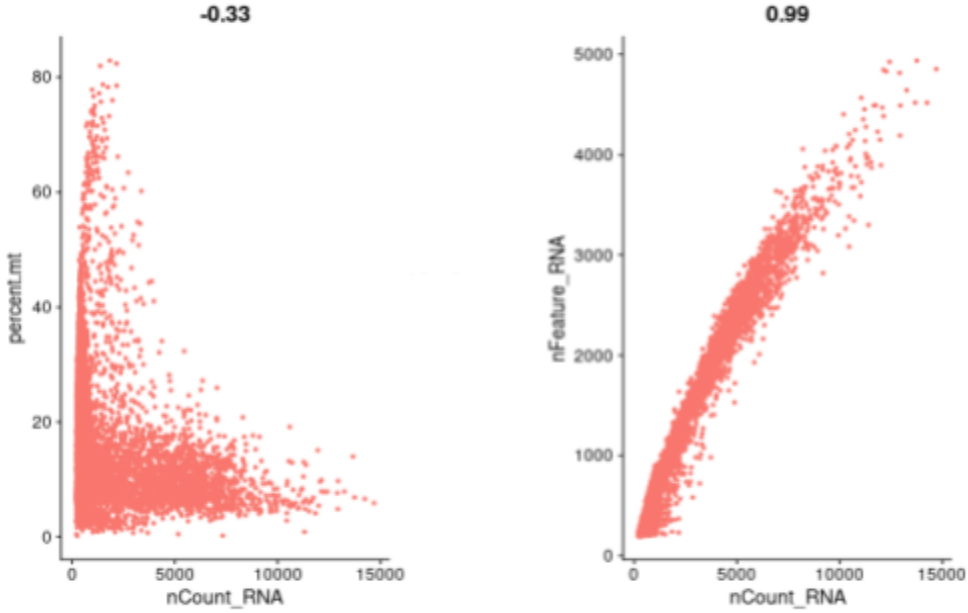


Figure 2B - Scatter plot of Quality Control Metrics
Unfiltered: 60232 Ensembl IDs/cells, After filtering: 6028 cells

Feature Selection

Identification of highly variable features was then carried out to filter low variance genes. The highly variable features are a subset of features that exhibit high cell-to-cell variation in the dataset . Focusing on these genes in downstream analysis helps to highlight biological signals in single-cell datasets. After normalizing the data, variance selection was performed in order to compare feature expression across cells. This is done using the FindVariableFeatures function, contained in the Seurat package. *Figure 3* shows the 2,000 most variable features selected, with the top ten by average Log2 Fold Change, labelled.

The data was then scaled using Seurat's ScaleData command. This shifts the expression of each gene so that the mean expression across cells is 0, and scales the expression of each gene so that the variance across cells is 1. This is a standard pre-processing step prior to dimensional reduction techniques like PCA.This step ensures highly-expressed genes do not dominate in downstream analyses (Hao et al., 2020).

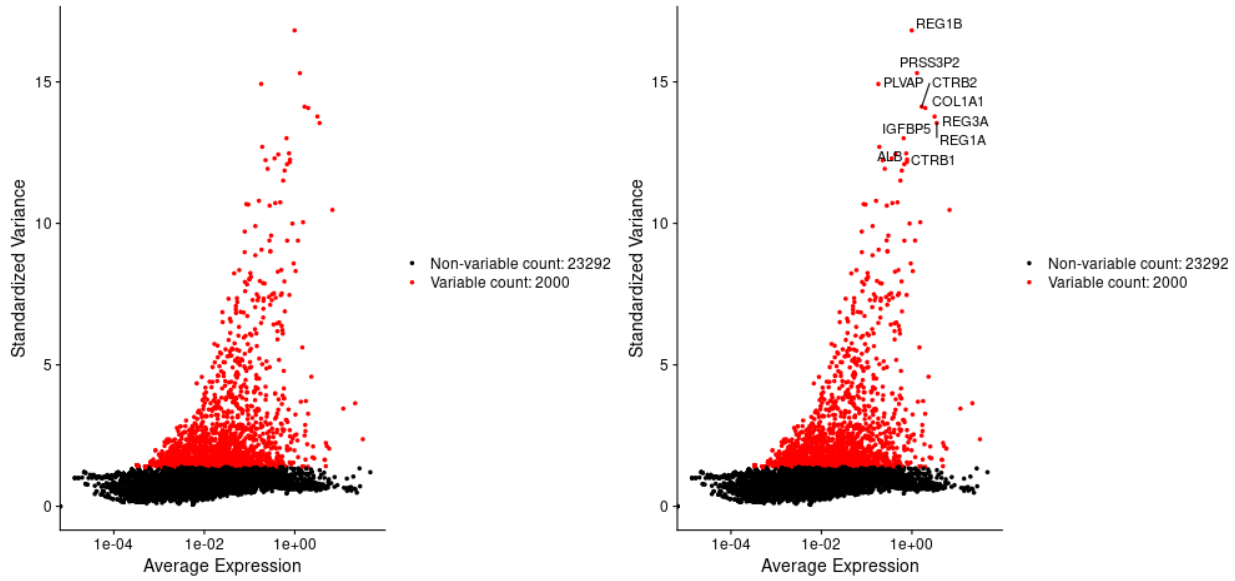


Figure 3 - Feature Selection Scatter top 10 genes labeled

After scaling the data, principal component analysis (PCA) is performed. We then want to determine how many features to include in the dataset. After a PCA we need to determine how many PCs to include this can be done by examining the elbow plot which ranks the principal components based on the percent of variance explained by each component. The top principal components therefore represent a robust compression of the dataset and I have chosen to include 10 as that is where the elbow becomes flat, we therefore believe all the variance is explained by the first 10 PCs shown by *Figure 4* below.

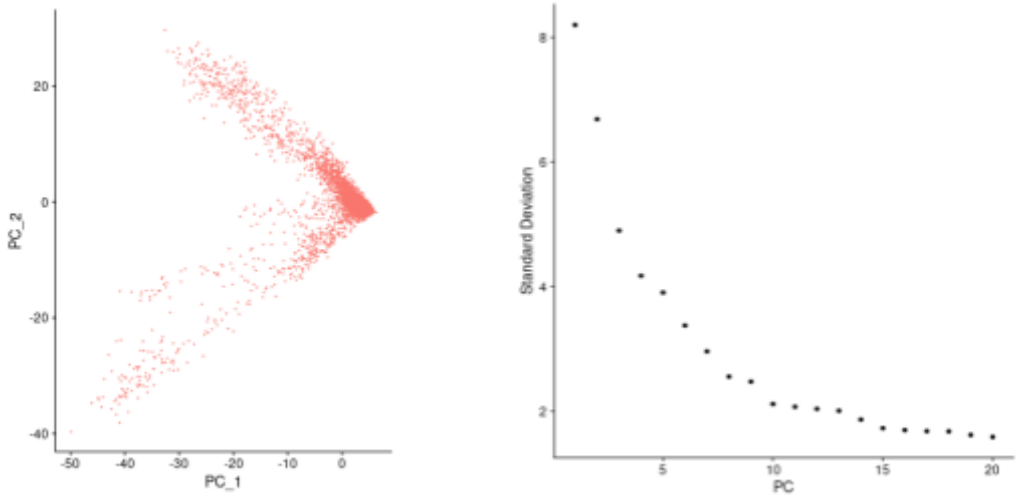


Figure 4 - PCA and Elbow plot for dimension reduction and PC selection

Cell Subtype Clustering

To cluster the cells, I followed the Seurat tutorial which used the Louvain algorithm, to iteratively group cells together. This step is performed using the FindNeighbors function, and takes as input the previously defined dimensionality of the dataset). The FindClusters function implements this procedure. The twelve clusters found can be seen in the UMAP below *Figure 5* . We are now able to reduce the dimensionality of the data and view the clusters in a 2D space.

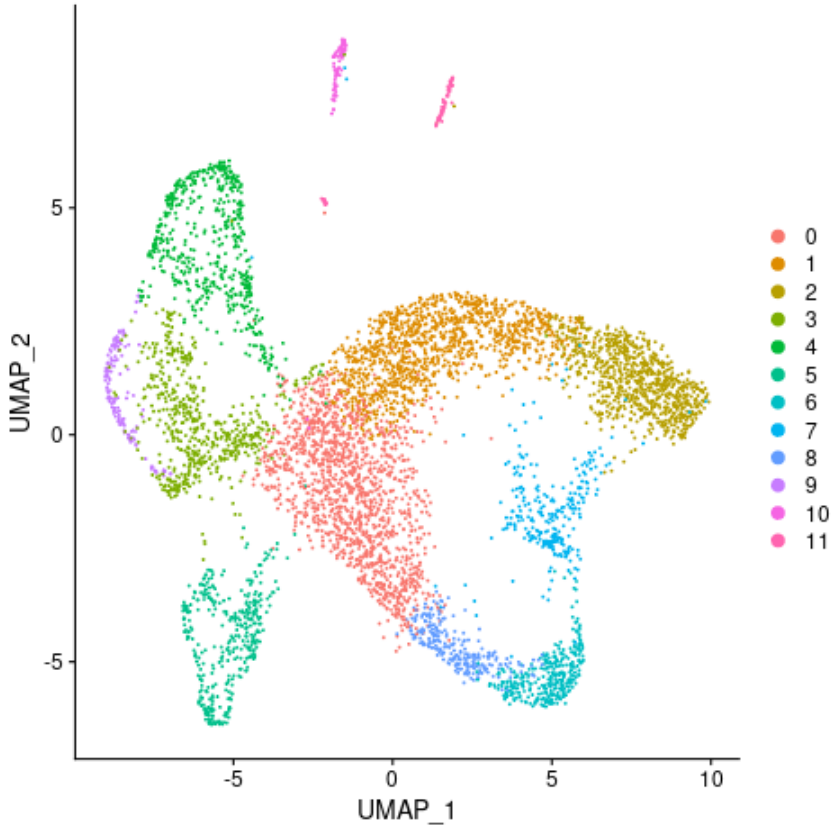


Figure 5 - UMAP showing clustering

The numbers of cells in each cluster and the proportion of cells in each cluster are shown below by *Figure 6*. The values are also listed below.

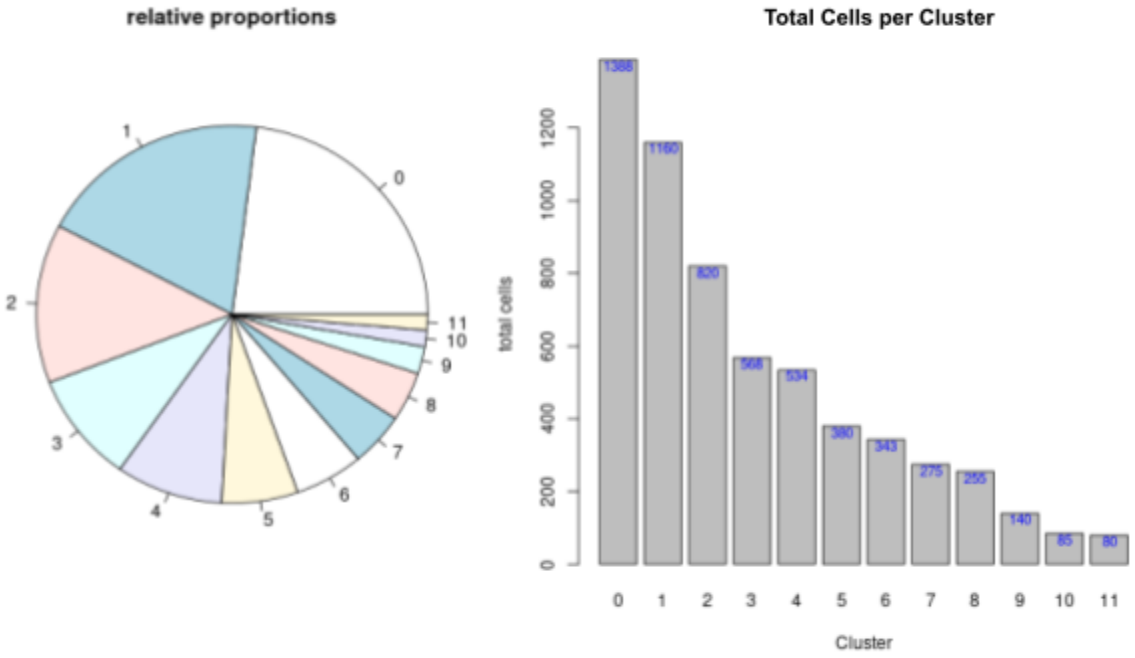


Figure 6 - Cluster Proportions and Counts
Total Clusters: *12*
Counts: *0: 1388, 1: 1150, 2: 820, 3: 568, 4: 534, 5: 380, 6: 343, 7: 275, 8: 255, 9: 140, 10: 85, 11: 80*

Clustering with Biomarkers
Marker genes were identified for each cluster using the FindAllFeatures command included in the Seurat package. This command automatically identifies differentially expressed genes in each cluster, compared to all other clusters, previously identified in the cell subtype clustering performed above. The default differential expression test of Seurat was used: the Wilcoxon Rank Sum Test. This is a nonparametric test used to compare outcomes between two groups, often interpreted as the comparison of the medians between the two populations tested.

A minimum percent threshold value of 0.25 was placed during the generation of the biomarkers to be used for cell type classification. This corresponds to a feature being detected at a minimum of 25% in either of the two groups of cells in order to be considered informative. The top ten marker genes for each of the twelve clusters, by average $\log_2$FoldChange, were then exported for further analysis. This corresponded to the top ten most differentially up- or down-regulated genes per cluster.

Using the above-generated marker genes, clusters were assigned their corresponding cell types. These were visualized via UMAP projection, as detailed in the *Results* section.

Gene Set Enrichment Analysis
Gene set enrichment analysis was performed on both unfiltered and filtered sets of marker genes produced by Seurat for all twelve gene clusters. Filtering involved selecting marker genes that had an adjusted p-value of less than 0.05 and a $\log_2$ fold change of greater than 1. Analysis was done with Metascape (Zhou et al. 2019) using the express analysis option.

Results

Cell Type Identification

Marker genes for each of the twelve clusters were identified using the FindAllFeatures command in Seurat, as detailed in the *Methods* section above. This resulted in a total of 4562 features being selected as "marker" genes. Statistics such as the p-value, adjusted p-value, average $\log_2$FoldChange, and cluster of origin were also automatically generated. The top ten marker genes, by most significant average $\log_2$FoldChange, were exported for each of the clusters, and used to generate *Figure 7*, below.

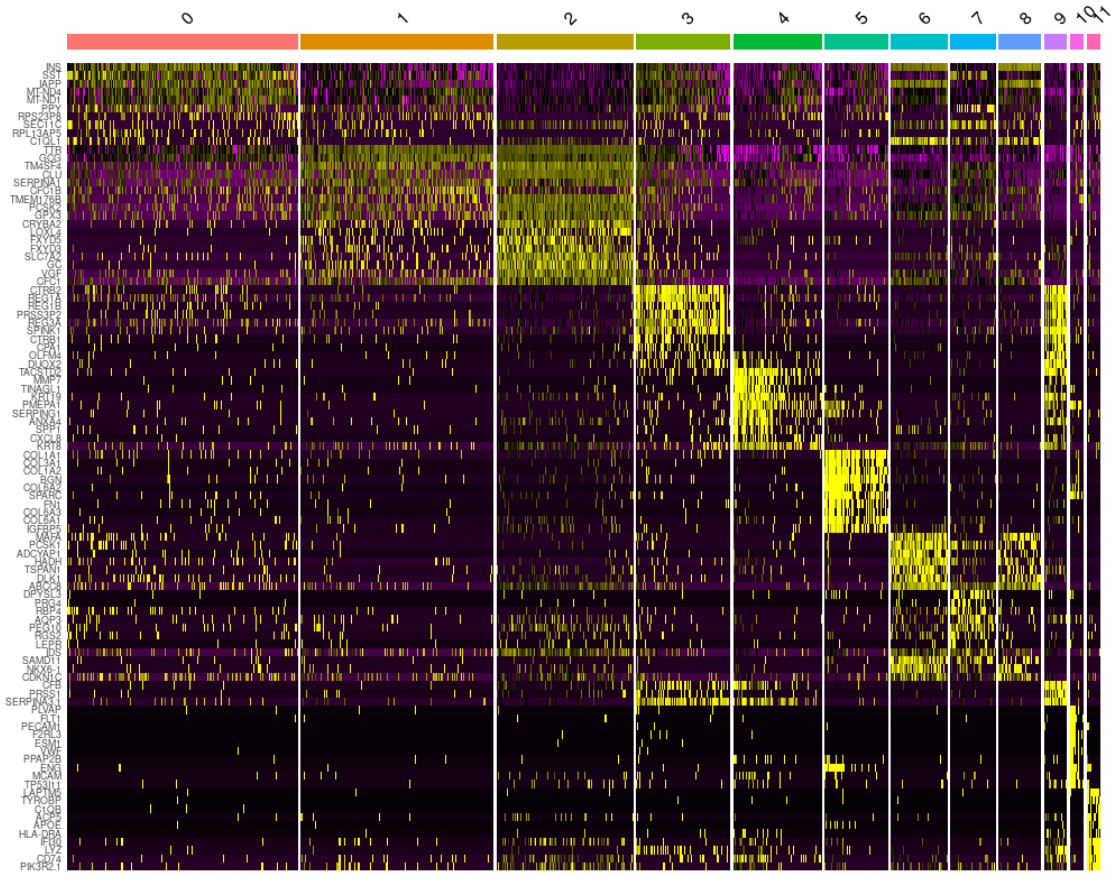


Figure 7 - Clustered heatmap of log normalized UMI counts, showing top ten differentially expressed genes in each of the twelve clusters.

This heat map was generated prior to celltyping, and informed many of the decisions made during the celltyping process. For example, Clusters 1 and 2 appear to share many of the same up- and down-regulated genes, as do Clusters 3 and 9, and Clusters 6 and 8. Therefore, it is reasonable to postulate that despite being sorted into different subpopulations, these pairs of clusters may belong to the same, or very similar, cell types, pending confirmation from further analyses. It was reassuring to note that the cell subpopulations appeared to separate well, with distinct highly expressed "bands" in the heatmap, for each of the subpopulations identified.

In order to best replicate the celltyping performed by Baron et al., the table of marker genes used for cell identification during their analysis was downloaded from their supplementary data (Baron et al., 2016).

For some of the marker genes used by Baron et al., the gene representative of a specific cell type was found to be in the top marker genes of only one cluster, such as the PDGFRB gene. This gene was found differentially expressed only in Cluster 5, which was then assigned the corresponding "stellate" cell type. A full table of the top marker genes for each cluster can be found in the *Supplementary Data*.

Other marker genes from the Baron et al. analysis were less informative, as they were present in the list of marker genes of multiple clusters. For example, as per *Figure 8* below, the GCG gene is the marker gene for alpha cells, while the INS gene is the marker gene for beta cells. However, both appeared to be highly enriched in Cluster 0. At the same time, GCG was also highly enriched in Clusters 1 and 2, while INS was highly enriched in Clusters 6 and 8.

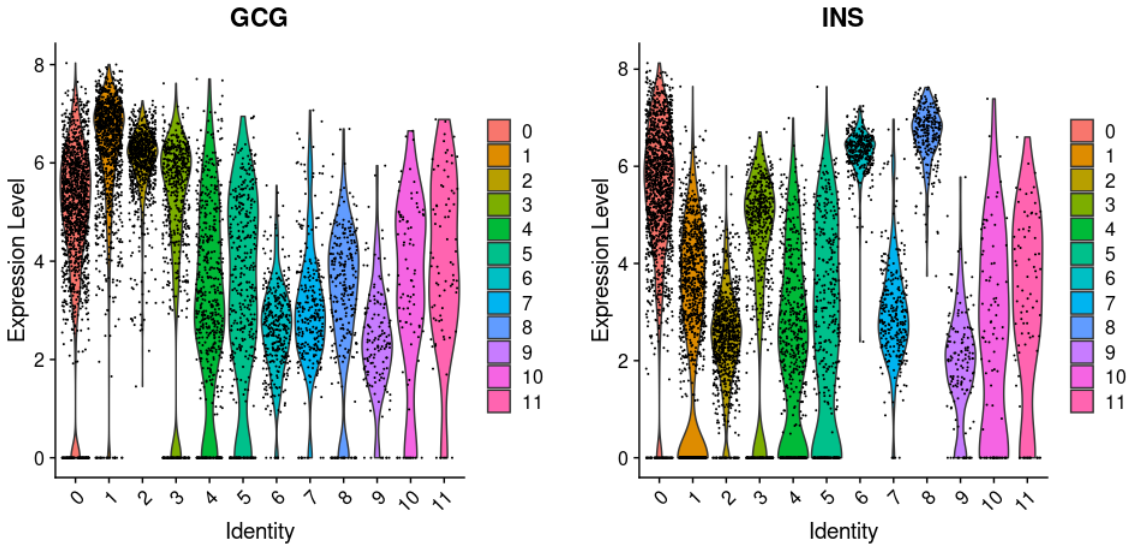


Figure 8 - Violin plots detailing the UMI counts for the GCG and INS genes, grouped by cluster. These two genes both appeared to be highly enriched in Cluster 0, amongst others, making identifying cell type using the marker genes provided by Baron et al. slightly more challenging. Violin plots for each of the genes supplied by Baron et al. can be found in the Supplementary Data.

Because these violin plots were relatively uninformative, FeaturePlots were generated, using the FeaturePlot command in the Seurat package, to align particular genes with their respective clusters. A UMAP projection was first generated of the data, colored by cluster. Each of the genes were then supplied to the FeatureMap, such that only the cells for which the selected gene was enriched were colored. This allowed us to select clusters for which a particular gene was enriched, based on their positions in the UMAP, as shown in *Figure 9,* below.

In *Figure 9B*, it can be seen that the GCG gene is most highly enriched in the areas corresponding to Clusters 1 and 2. This, accompanied by the information provided in *Figure 7* suggesting that Clusters 1 and 2 belong to the same cell type, are highly indicative that Clusters 1 and 2 belong to the cell type best known for their enrichment of the GCG gene: the pancreatic alpha cells. A similar process was used in the identification of the clusters belonging to the pancreatic beta cell type, marked by their enrichment of the INS gene, as seen in *Figure 9C*. This gene appeared to be most localized in the area of the UMAP

corresponding to Clusters 6 and 8. Because we know from *Figure 7* that Clusters 6 and 8 are likely to belong to the same cell type, these clusters were labelled pancreatic beta cells.

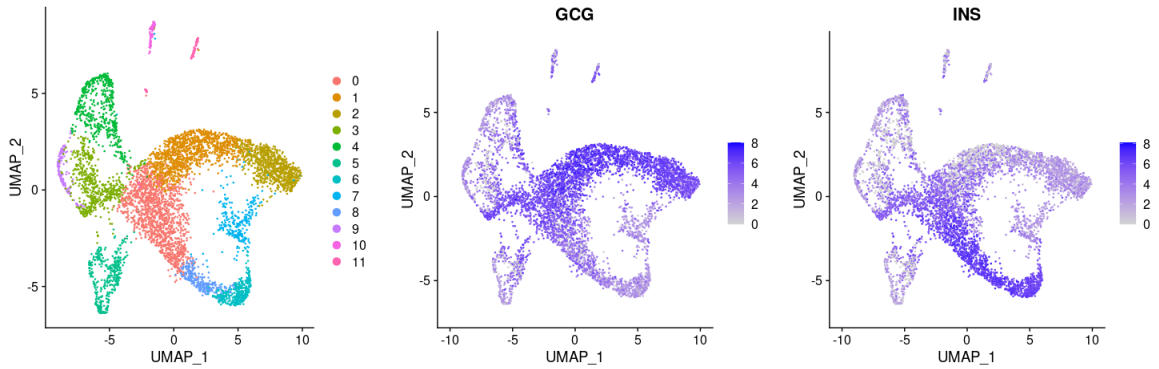


Figure 9 - UMAP projection of clustered cells, colored by (a) cell subpopulations, as identified previously, (b) GCG gene enrichment, and (c) INS gene enrichment. Similar FeaturePlots for each of the genes supplied by Baron et al. can be found in the Supplementary Data.

Other cell types were much more easily identified, using the Baron et al. marker genes. *Figure 10* below shows the violin plots for stellate, vascular and macrophage cell type markers, respectively, grouped by cell subpopulation cluster. These genes were only significantly expressed in one of the twelve clusters, and were able to be assigned without significant difficulty.

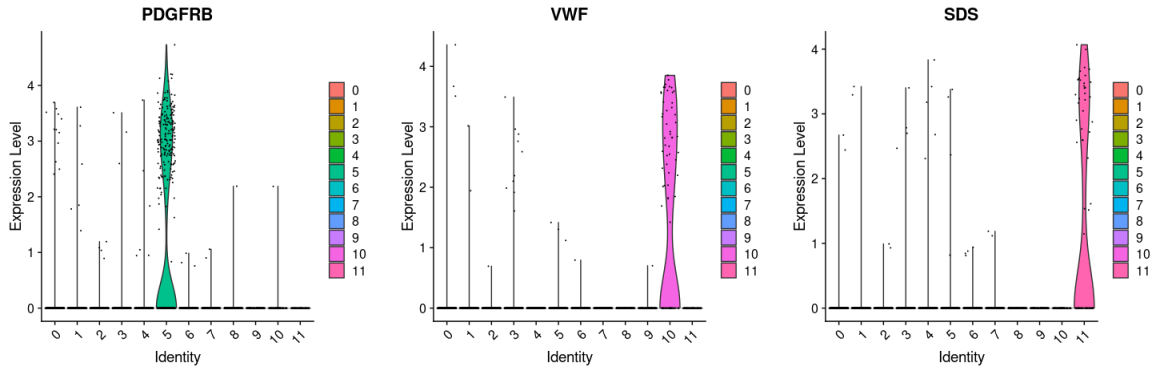


Figure 10 - Violin plots of the expression level of the PDGFRB, VWF and SDS genes. As per Baron et al., the increased expression of these specific marker genes correspond to the presence of stellate, vascular and macrophage cells, respectively. Note that Baron et al. use the terms "vascular" and "epithelial" interchangeably when drawing comparisons.

One group of cells that was particularly difficult to assign to a cell type was Cluster 0, as it seemed to show enrichment in many of the FeaturePlots and violin plots for other cell types that had already been assigned. For example, the pancreatic alpha and beta cells, previously identified by their enrichment in the GCG and INS genes in *Figure 9*, both seem to show enrichment of Cluster 0 when examining their respective marker genes. Therefore, another strategy was used to identify the cell type of Cluster 0.

PanglaoDB is a curated database of both human and mouse single cell RNA-Sequencing samples. The database contains 305 different human samples, across 74 different tissues, including the pancreas (Franzén et al., 2019). The database information is available as a CSV, which can be downloaded and loaded into R as a dataframe. In order to determine the cell type of Cluster 0, the top ten marker genes of the cluster were found within PanglaoDB, whilst filtering the species for *Homo sapiens*, and the organ for "pancreas". Of the top ten marker genes of Cluster 0, two were found to be indicative of pancreatic delta cells. The remaining eight genes were either not found in PanglaoDB, or were indicative of cell types that had already been clearly labelled. Therefore, operating under the hypothesis that Cluster 0 was composed of delta cells, *Figure 11* was generated, using the marker gene of pancreatic delta cells, SST, provided by Baron et al.

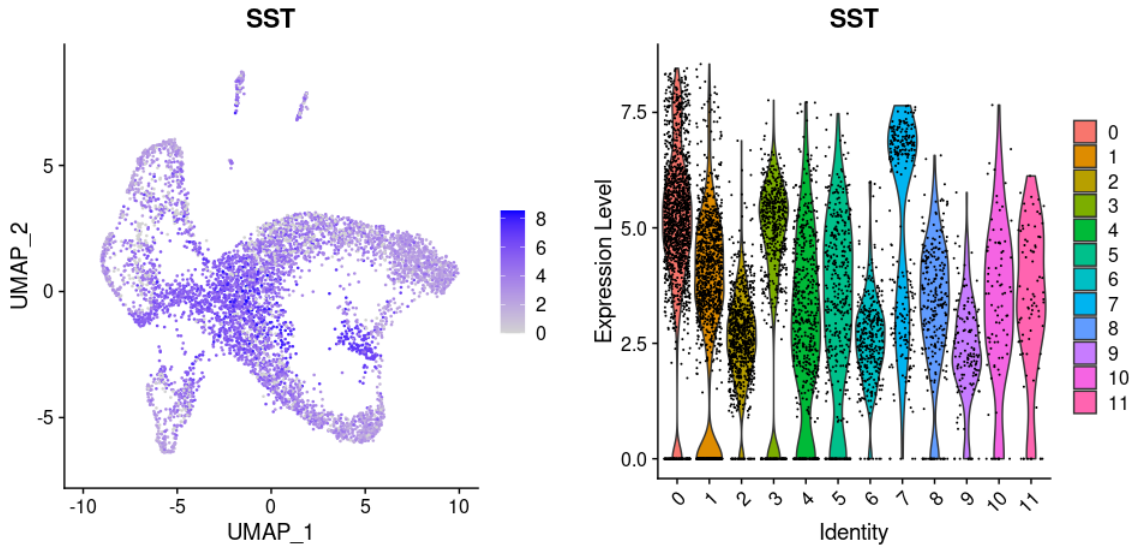


Figure 11 - Violin and Feature Plots of the SST gene, indicated to be a marker gene for the pancreatic delta cell cell type. From the above plots, the SST gene appears to be most highly enriched in Clusters 0 and 7.

As per *Figure 11* above, the cells enriched in the SST gene seem to be localized around Cluster 0, supporting the hypothesis that Cluster 0 is composed of pancreatic delta cells. In addition, the violin plot supports this conclusion, as it shows significant enrichment of the SST gene in Cluster 0. Of note, Cluster 7 was also considered when identifying the cell type of Cluster 0, as the SST gene appears also to be enriched in the area of the UMAP corresponding to Cluster 7, and the SST gene appears to be upregulated in Cluster 7, as well as Cluster 0. However, this hypothesis was ruled out after Cluster 7 was labelled as pancreatic delta cells due to their enrichment of the pancreatic delta cell marker, PPY. The FeaturePlot and Violin Plots for PPY gene enrichment can be found in the Supplementary Data.

After confirming each of the cell type categories using both a violin plot and featureplot, the twelve cell subpopulations were classified into nine different cell types, as visualized in *Figure 12* below. Of note, some of the cell types identified in the Baron et al. analysis were not found in ours. Notably, the pancreatic epsilon cells (identified with marker gene GHRL), the cytotoxic T-cells (identified by marker genes CD3, CD8 and TRAC), as well as the mast cells (identified by marker genes TPSAB1, KIT and CPA3) were absent. The conclusion that these cells were not present in our sample was reached because

none of the marker genes provided by Baron et al. for these cell types were present in the differentially expressed marker genes for each of the clusters. Interestingly, both mast cells and cytotoxic T-cells, two of the three cell types that were not identified in our samples, are commonly associated with the immune response. Further research will be needed to ascertain whether the presence or absence of an immune response during the sample collection process was significant in this finding.

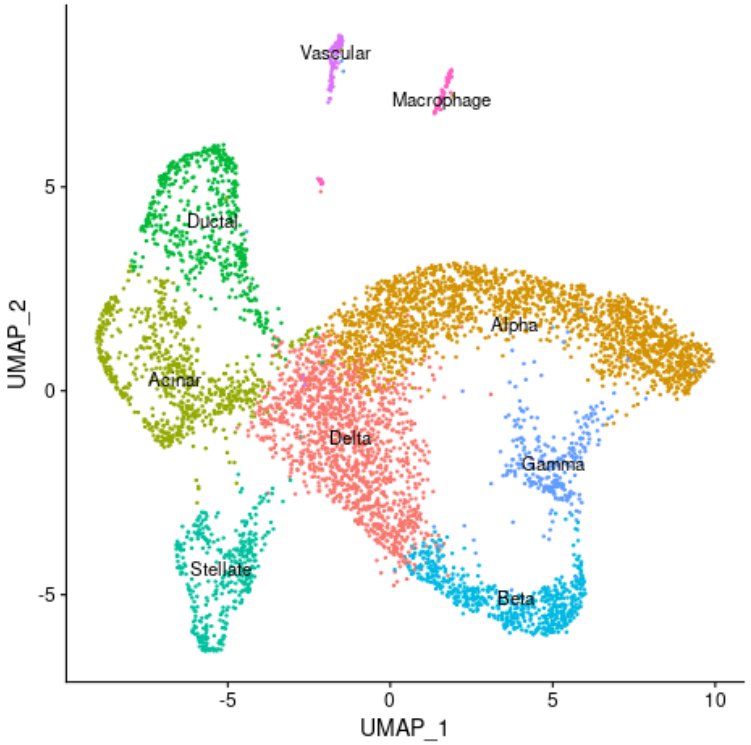


Figure 12 - UMAP of log-normalized count matrix, colored by celltype. From the twelve clusters, nine different cell types were identified. The marker genes provided in the Supplementary Data of the Baron et al. were used to inform the cell type identities.

To assess the efficacy of the cell clustering, *Figure 13* was generated, as can be seen below. After grouping by cell type, clear "bands" of highly enriched genes can be clearly visualized in each of the distinct cell types. The heatmap also provides insight into novel marker genes for each cell type identified. For the purpose of our analysis, "novel" marker genes are those that appear to be differentially expressed in each cell type, yet not identified by Baron et al. to be indicative of a particular cell type or subpopulation. A table of novel marker genes, along with an accompanying heatmap, are available in the Supplementary Data.

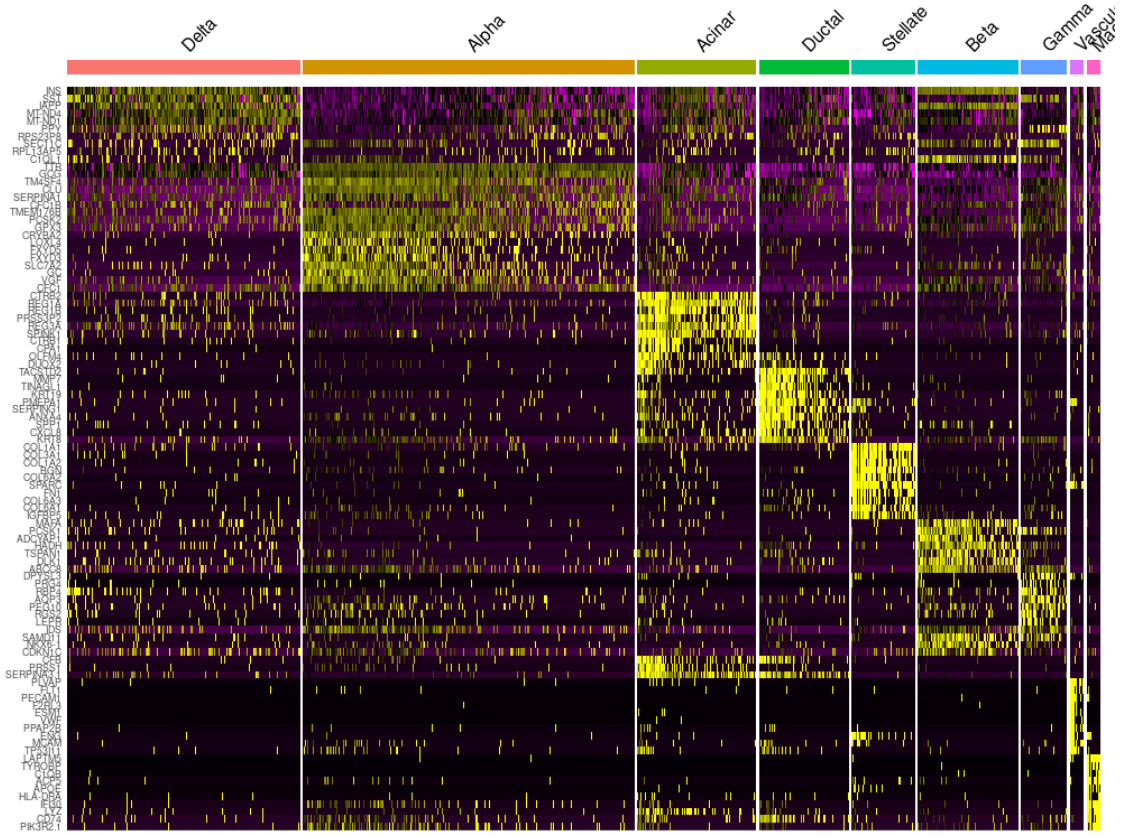


Figure 13 - Heatmap of the top ten most significantly enriched genes per cluster, grouped by assigned cell type. Six of the twelve clusters belonged to three groups, therefore resulting in nine cell types being identified in the above plot.

Gene Set Enrichment Analysis

Table 3 - Top enriched biological process from gene set enrichment analysis for filtered and unfiltered sets of genes from each cluster.

Cluster	Label	Unfiltered		Filtered	
		Number of Marker Genes	Biological Function	Number of Marker Genes	Biological Function
0	Delta	46	Electron transport chain	3	insulinoma
1	Alpha	105	Regulated exocytosis	7	Amyloidosis/Chronic pancreatitis
2	Alpha	695	Signal release/Regulation of vesicle-mediated transport	30	Extracellular matrix organization
3	Acinar	60	Response to	26	Cell wall disruption in another

			wounding/Thermogeneis		organism/Immune response
4	Ductal	406	VEGFA-VEGFR signalling pathway	146	Positive regulation of cell migration/Cell adhesion molecules
5	Stellate	275	Axon guidance/Nervous system development/ Extracellular matrix organization	133	Extracellular matrix organization
6	Beta	638	SRP-dependant cotranslational protein targeting to membrane	43	Protein secretion
7	Gamma	456	Regulation of neuron development	15	Organ growth
8	Beta	195	Eukaryotic translation elongation	27	Regulation of gene expression in beta cells
9	Acinar	1130	Meyloid leukocyte activation	189	Response to wounding
10	Vascular	375	Blood vessel development	213	Blood vessel development
11	Macrophage	183	Regulated exocytosis/Leukocyte activation as immune response	113	Leukocyte activation as immune response

Marker genes derived from Seurat for all 12 clusters were used in Gene set enrichment analysis.
Top GO terms before and after filtering of marker genes for an adjusted P value of less than 0.05 and a $\log_2$ fold change of greater than 1 are summarized in Table 3. For most clusters there was a large drop in the number of genes from filtering which indicated that, of the genes that significantly differed between clusters, the majority were only slightly different. This similarity is reflected in the Circos plots of gene overlap between the clusters for both filtered and unfiltered clusters, which show a large amount of overlap in genes between clusters (fig. 14). Finally, heatmaps of enriched ontology clusters show that for both filtered and unfiltered clusters, cluster 5, 10, 4, 9, and 11 represent one metacluster of similarity and are highly enriched for the GO terms found in the analysis, while clusters 0, 1, 2, 3, 6, 7, and 8 are far less enriched and appear to be more distinct from one another (s.fig. 4).

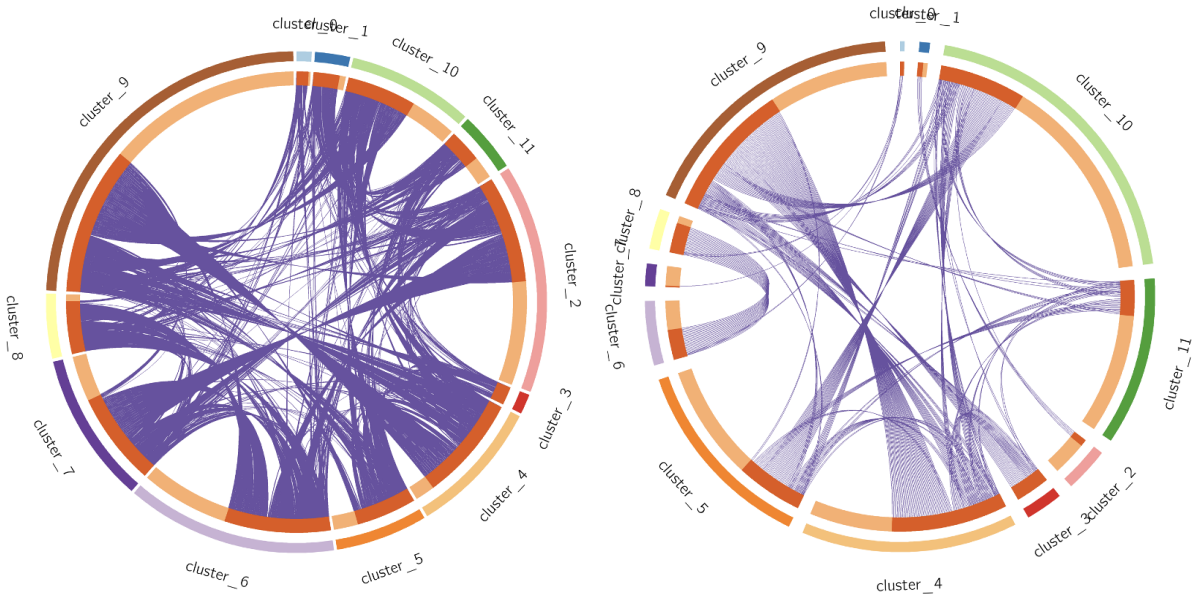


Figure 14 - Circos plots for unfiltered (left) and filtered (right) gene sets for the 12 clusters, showing the amount of overlap of genes between clusters. The same genes are linked with purple lines, and the amount of dark orange in the inner ring represents the proportion of genes that are not unique for each cluster.

Discussion

Heatmap Comparison to Baron et al.

During our analysis, we attempted to recreate Baron et al.'s Figure 1B using the same genes present in their heatmap. However, we quickly noticed that many of the genes that they had noticed significant enrichment within, and had thus used as marker genes in the identification of different cell types, were not present in our analysis. *Figure 15* below shows heatmaps with the same genes, in the same order, as those in Baron et al.'s Figure 1B, grouped first by cluster, then by later assigned cell types.

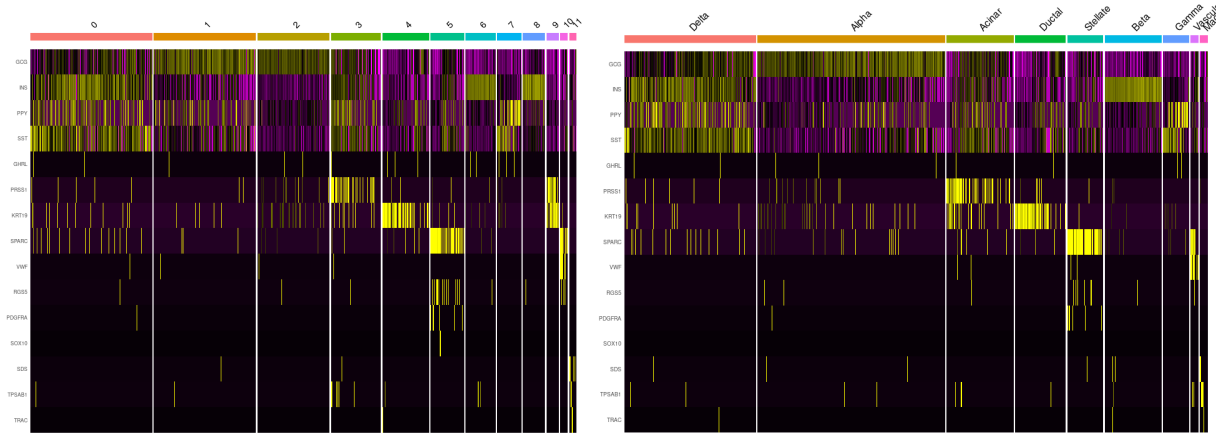


Figure 15 - Heatmap recreation of Baron et al.'s Figure 1B, in which significant marker genes are identified and used to sort cells into different cell types, grouped by (A) cell subpopulation cluster, and (B) later assigned cell types.

It is interesting to note that while some strong bands are present, such as those delimiting the acinar, ductal, and stellate cell populations, other cell groups did not appear to have as strong of a "band" of differential expression. Most of the cells, for example, appear to have relatively high enrichment of the first four genes (GCG, INS, PPY and SST, respectively). This makes the identification of the cell types that rely on the upregulation of these cells, notably the pancreatic alpha, beta, and delta cells, relatively difficult.

Other genes that were found by Baron et al. did not appear at all within our data. The latter six genes in the list, for example (RGS5, PDGFRA, SOX10, SDS, TPSAB1 and TRAC) had very little expression across all cell groups, with no apparent pattern or "banding" within the heatmap to suggest a pattern or identifiable transcription profile associated with them.

Many of these noted differences are likely due to Baron et al. having a much larger cell population to identify trends from. In their Figure 1B, for example, they demarcate the samples from which each of the cells are obtained from, within their heatmap. A large portion of this bar appears to be orange, corresponding from Subject/Donor 3. Because we used data from Subject/Donor 2, represented by the blue/turquoise labels, it is understandable that many of their findings did not translate directly to what we observed. Therefore, further analysis, and perhaps the use of more data from different subjects may be useful in order to confirm Baron et al.'s analysis and replicate their findings.

Dimensionality Reduction Projection Comparison to Baron et al.

For visualization of our data, we chose to project our single cell data onto a Uniform Manifold Approximation and Projection (UMAP). The UMAP is a relatively new dimensional reduction technique, first introduced in 2018 (McInnes et al., 2018). The algorithm underlying UMAP is graph-based and similar to that of tSNE in principle; while Baron et al. chose to project their data onto a tSNE, we chose to use a UMAP for a variety of reasons. Most importantly, tSNE does not preserve the global data structure, instead focusing on preserving local data structures. This means that while the shapes of individual groups may be meaningful and consistent, the distances between clusters can be drastically different, and therefore, meaningless. As we are looking not only at the clustering of the cells themselves, but also at the distances between them in order to assign appropriate cell types, we chose to visualize our data on a UMAP, as we felt this would be more informative.

Baron et al.'s clustering in Figure 1D showed fourteen easily distinguishable, separate clusters (Baron et al., 2016). However, as noted, the distances between clusters in this case are arbitrary. Therefore, the distinctiveness of their fourteen clusters remains unclear, as no UMAP was generated to visualize the global structure of their dimensionality reduction. However, it is interesting to note that Baron et al. were able to identify five additional subpopulations of cells in their analysis than we were. Possible reasonings are explored in the section below.

Cell Identification Comparison to Baron et al.

In our analysis, we were able to identify nine different pancreatic cell types, while Baron et al. were able to identify fourteen. The five cell types that they had identified that were not present in our analysis were quiescent stellate cells, cytotoxic T-cells, mast cells, Schwann cells and pancreatic epsilon cells.

It was unsurprising that we were unable to detect the pancreatic epsilon cells or the cytotoxic T-cells, as Baron et al. note that these each only constituted an approximate 0.1% of a mixture of cell types each (Baron et al, 2016). Therefore, as our analysis consisted of less than 5000 cells, it was statistically improbable that enough of these cells would be present such that they could be identified and subsequently clustered.

Additionally, Baron et al. noted that one specific donor (Donor 1) had a surprisingly high number of epsilon cells, believed to be absent in adults. As our samples came from Donor 2, a 51-year-old female, it was unsurprising that these cells were not identified.

It is unclear as to why the quiescent stellate cells were not present in our analysis. One possible hypothesis is that there was simply not enough information in the form of cell counts in order to fully differentiate the quiescent and activated stellate cells. One can postulate that the transcriptional profiles of these two closely related cell types would require much more information to ascertain a noticeable difference, and therefore, be clustered as two distinct cell subpopulations.

It was also interesting to note that the marker cells used in Baron et al.'s Figure 1G did not correspond directly with those that they had listed as the marker genes of interest in the Supplementary Data. The figure labels identify fifteen different cell populations, while only twelve are listed in their Supplement. For example, their Supplementary Data lists only one marker gene for stellate cells, while their Figure 1G label distinguishes between marker genes for both activated and quiescent stellate cells, amongst other differences. The authors also appear to use the terms "epithelial" and "vascular" interchangeably when examining cells with the marker gene, VWF. Because it was unclear where this discrepancy arose from, and because the authors did not appear to provide a source for these marker genes, we used both sets of marker genes provided in order to best identify our cell populations.

Gene Set Enrichment Analysis

The Circos plots showed a large degree of overlap between the gene sets for the different clusters, which suggests that they may be performing similar biological roles (fig. 14). Clusters 0, 1, 3, and 8 had an especially high degree of overlap, sharing almost all of their genes with other clusters, and were likely defined more by the transcripts they lacked rather than the transcripts they had. The heatmap of enriched GO terms also shows that clusters 4, 5, 9, 10, and 11 were more similar to one another than to the other clusters, which provides support to their identification as alpha, beta, and delta cells (s.fig. 4). These cell types are three major endocrine cells in the pancreas, so it follows that they would all share similar GO terms (Lawler et al. 2017).

When it came to the top GO term for each cluster, many of the clusters were found to have roles that corresponded to their identification, while others produced surprising results. Between filtered and unfiltered gene clusters, there was a slight discordance in GO terms (table 3). Though not reported, many

of the top GO terms for the filtered clusters were returned for the unfiltered clusters, though with lesser probability. Some of the clusters returned top terms that matched their identification exactly, like cluster 10 which was identified as vascular cells and returned a top GO term of “blood vessel development”. Other clusters, such as the ones identified as the alpha, beta, and delta cells returned terms generally associated with endocrine activity such as “protein secretion”, “regulated exocytosis”, and “regulation of vesicle-mediated transport”. Again, these cells serve major secretory roles within the pancreas, so these results provide confidence for the clusters’ identification (Breton et al. 2015, Chen at al. 2017). Despite this, identifying the cell types of these cluster *de novo* would be impossible based on the top GO terms alone, and would likely require a more holistic approach considering all returned GO terms, as well as greater knowledge of pancreatic cells types to be able to distinguish between the alpha, beta, and delta cells. Additionally, clusters 5 and 11 both returned top GO terms that corresponded with their identifications of stellate and macrophage cell type, such as “extracellular matrix organization”, which correlates with stellate cells role of creating connective tissue, and “leukocyte activation as immune response”, which relates to the function of macrophages as immune response cells (Apte et al. 2012, Padoan et al, 2019).

Interestingly, the returned GO terms for both clusters identified as acinar cells were related to immune responses as well, returning top GO terms of “response to wounding”, “cell wall disruption in another organism”, and “immune response”. This may suggest that the sampled pancreas may have been undergoing an immune response at the time of sampling, though these results are more likely the product of the data processing and lack of sample size in the filtered gene set rather than a biological phenomenon. Additionally, immune response behaviour is uncharacteristic of acinar cells, which primarily serve a secretory function (Macdonald et al. 2010). Both clusters 0 and 1 (labelled as delta and alpha cells) returned pancreatic diseases as a top GO terms for the filtered gene sets, though this was again likely due to the lack of sample size rather than an indication of the health of the sample. Finally, cluster 7 (labelled as gamma) returned terms associated with neuron development and organ growth, neither of which were related to gamma cell functioning, which is endocrine like that of alpha, beta, or delta cells (Lawler et al. 2017).

Conclusion

Collective
Collectively, the overall conclusion that we hope readers are able to draw from our analysis is that technologies for the analysis of single cell data has drastically improved, allowing for the discovery and annotation of previously unknown cell types or transcriptional profiles. While our analysis consisted of only a small subsample of Baron et al.'s original data, we were able to not only identify key marker genes to be used in the assignment of different cell types, but also elucidate novel marker genes that can be used in further analyses of pancreatic cell types and transcriptional activity.

Data *Curator*
Working with single cell sequencing data was a new experience all together. It was fun, and a bit challenging but in the end, combining Awk, R and Salmon, extraction of barcode, whitelisting and alignment was successfully executed.

Programmer
As the programmer I found myself making a lot of difficult decisions in regards to filtering of the cells, as I knew all downstream analysis would be dependent on my decisions. I had a tough time determining which mapping was the most important. In the end I chose MT-genes vs. having all the umapped genes understanding that this could directly affect our analysis. I also had to learn how to navigate working with sparse matrices and huge amounts of data. I usually am able to filter with data frames but I had to learn some new ways to work around not being able to see my actual data which was another challenge.

Analyst
As the analyst for this project, one of the challenges that I encountered was deciding on the marker genes that should be used in the identification of cell types. While Baron et al. provided a table of marker genes that they used during their analysis, this table seemed to be inconsistent with the ones they ultimately used to label their figures. Therefore, our identified cell types were corroborated with use of an external database containing cell type information and corresponding marker genes. However, this database was published in 2019, three years after Baron et al.'s study was published, which may contain information that was not available at the time of the initial experiments.

Another challenge that I faced during my analysis was that many of the marker genes mentioned by Baron et al. were not identified in our datasets, which ultimately led us to determine that these cell types were simply not present in our data. Further analysis with a larger dataset, containing cells from more donors, as was the case with Baron et al., may be useful in the identification of all the cell types that were noted in the original paper.

Biologist
While computationally the role of the biologist for this project was not demanding in the least, the greatest challenge for me was interpreting the results. Initially I had used EnrichR for the gene set enrichment analysis, but I had difficulties making sense of the returned GO terms. I then tried using metascape, which outputs far more analyses than EnrichR, and allowed me to dig deeper into the data. After a sufficient amount of reading about pancreatic cell types, the outputs from Metascape began to

make more sense in respect to the labeling of the clusters, and I was able to draw conclusions from the data.

References

Apte MV, Pirola RC, Wilson JS. Pancreatic stellate cells: a starring role in normal and diseased pancreas. *Front Physiol*. 2012;3:344. Published 2012 Aug 28. doi:10.3389/fphys.2012.00344

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

Brereton MF, Vergari E, Zhang Q, Clark A. Alpha-, Delta- and PP-cells: Are They the Architectural Cornerstones of Islet Structure and Coordination. *J Histochem Cytochem*. 2015;63(8):575-591. doi:10.1369/0022155415583535

Chen C, Cohrs CM, Stertmann J, Bozsak R, Speier S. Human beta cell mass and function in diabetes: Recent advances in knowledge and technologies to understand disease pathogenesis. *Mol Metab*. 2017;6(9):943-957. Published 2017 Jul 8. doi:10.1016/j.molmet.2017.06.019

Franzén O, Gan LM, et al. PanglaoDB: A Web Server for Exploration of Mouse and Human Single-Cell RNA Sequencing Data, *Database,* 2019 (2019). doi:10.1093/database/baz046

Hao Y., Hao S., Andersen-Nissen E., et al. Integrated analysis of multimodal single-cell data. *bioRxiv* 2020. doi:10.1101/2020.10.12.335331

Klein AM, Mazutis L, Akartuna I, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*. 2015;161(5):1187-1201. doi:10.1016/j.cell.2015.04.044

Lawlor N, George J, Bolisetty M, et al. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res*. 2017;27(2):208-222. doi:10.1101/gr.212720.116

Li J, Klughammer J, Farlik M, et al. Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. *EMBO Rep*. 2016;17(2):178-187. doi:10.15252/embr.201540946

MacDonald RJ, Swift GH, Real FX. Transcriptional control of acinar development and homeostasis. *Prog Mol Biol Transl Sci*. 2010;97:1-40. doi:10.1016/B978-0-12-385233-5.00001-5

Marini F, Linke J, & Binder H, et al. An R/Bioconductor Package for Interactive Differential Expression Analysis. *bioRxiv.* 2020. doi:10.1101/2020.01.10.901652

McInnes L, Healy J et al. UMAP: Uniform Approximation and Projection for Dimension Reduction. *Journal of Open Source Software.* 2018;3(29):861. doi: 10.21105/joss.00861

Padoan A, Plebani M, Basso D. Inflammation and Pancreatic Cancer: Focus on Metabolism, Cytokines, and Immunity. *Int J Mol Sci*. 2019;20(3):676. Published 2019 Feb 5. doi:10.3390/ijms20030676

Soneson C, Love MI, Robinson MD, et al. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4. 2015 doi: 10.12688/f1000research.7563.1.

Stuart T, Butler A, Hoffman P, et. al. Comprehensive integration of single cell data. *Cell* 2019; 1888-1902 doi: 10.1016/j.cell.2019.05.031

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

Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun.* 2019;10(1):1523. Published 2019 Apr 3. doi:10.1038/s41467-019-09234-6

Data Availability

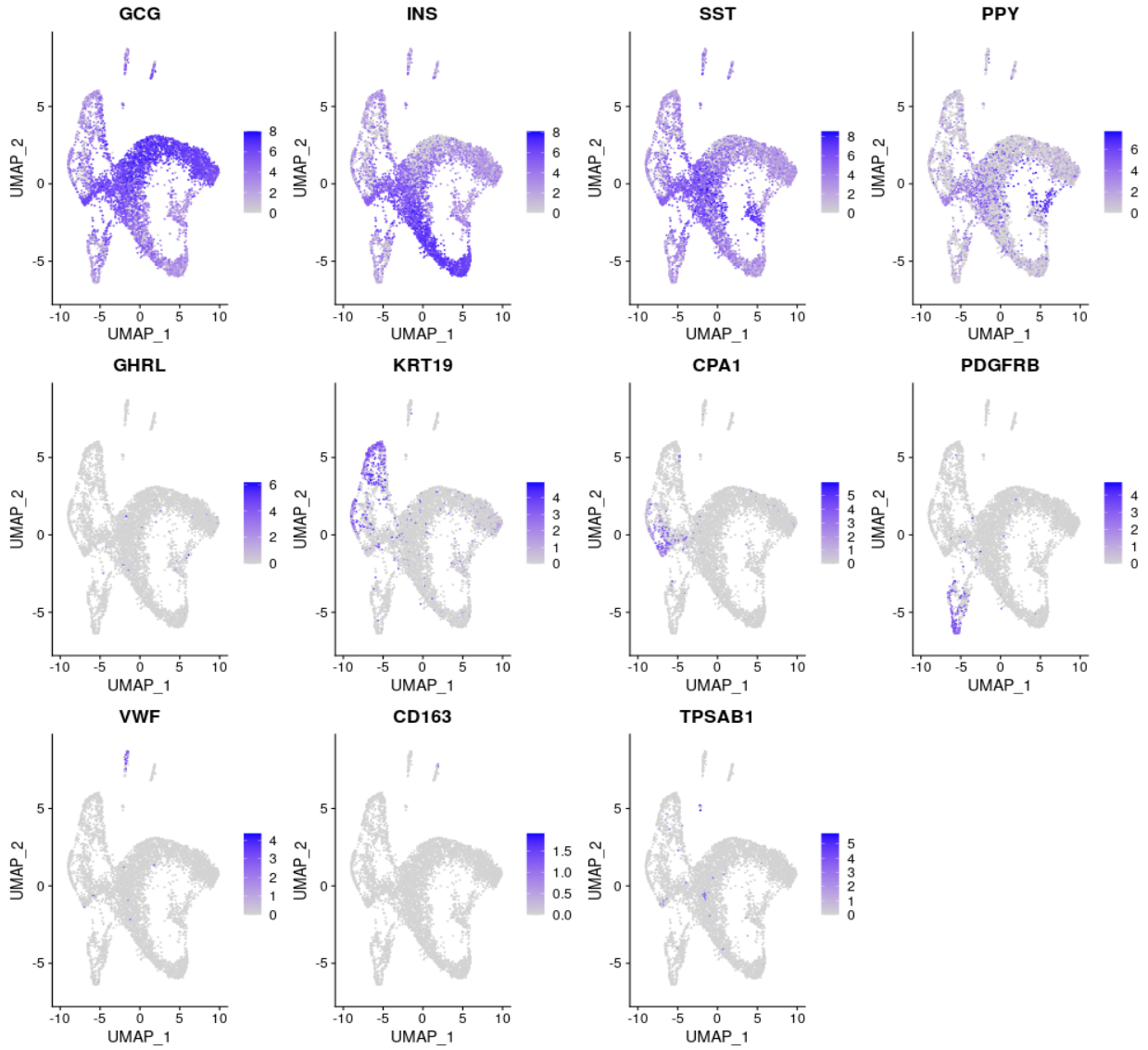
All code used to generate the above analysis can be found at github.com/BF528/project-4-lava-lamp
The data used in this project is accessible via NCBI Gene Expression Omnibus, Accession GSE84133

Supplementary Data

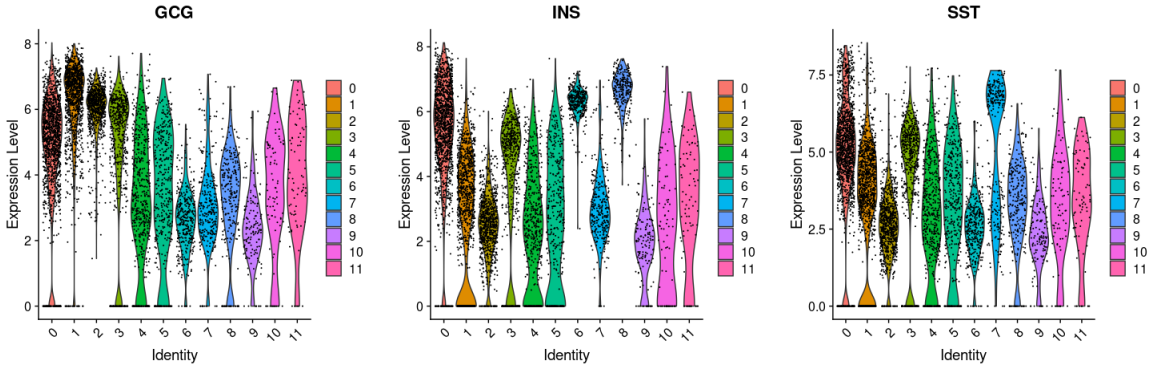
Supplementary Table 1 - Top three differentially expressed genes in each cluster, when compared to all other clusters. A full list of top differentially expressed genes, with the parameters detailed in the *Methods* section, can be found on the Shared Computing Cluster.

Gene Name	Cluster Number	P-Value (Nominal)	Average Log2 Fold Change
INS	0	1.36203833880569e-250	1.7535558892547900
SST	0	2.5451611410896e-204	1.918977868691810
IAPP	0	5.53357434916334e-202	1.63511455718647
TTR	1	3.77096859938277e-295	1.549374004285130
GCG	1	3.29795662842183e-270	1.7694264210700900
CFC1B	1	5.28575224232261e-56	1.386418368225930
LOXL4	2	0	1.6394761458129900
VGF	2	7.20414482269465e-287	1.6012297047601600
TM4SF4	2	7.33473652870561e-282	1.5922977672802200
REG1A	3	2.47549063580888e-297	3.5030101013855100
PRSS3P2	3	1.34007478887635e-250	3.5236862699988000
REG3A	3	2.48899446452007e-242	3.563265027374000
MMP7	4	1.71915862743668e-301	3.106444809156710
KRT19	4	3.83022718915902e-279	3.016037041385690
SPP1	4	9.1299457812822e-169	3.0370084264551100
COL1A1	5	0	5.53422313550408
COL3A1	5	0	5.04722869606767
COL1A2	5	0	4.778074063281130
MAFA	6	0	1.997484053030470
C1QL1	6	2.59467504177587e-239	1.9771795433956500
ABCC8	6	6.43955851002524e-205	2.03216866352383
RBP4	7	1.5793036281622e-132	2.485956472714400
AQP3	7	1.20539018986848e-126	2.238261537149520
PPY	7	1.1888596287157e-19	2.859373059193000
INS	8	5.72472547925224e-116	2.190507837602500
MAFA	8	5.24857452075606e-97	1.974866229872620
IAPP	8	2.38388841314751e-84	2.205835645884890
CFB	9	1.89261995401111e-246	3.6853139608285500
OLFM4	9	6.59239883298584e-186	3.545102160791920
CTRB2	9	1.43076027328194e-152	3.784087003214890
PLVAP	10	0	5.1705489069409600
FLT1	10	0	4.195710840525990
PPAP2B	10	4.61972024231453e-267	4.109267461712850
LAPTM5	11	0	4.590606296673010
ACP5	11	1.24761110043844e-167	5.457484142537950
IFI30	11	1.49108207987366e-84	4.553855701933170

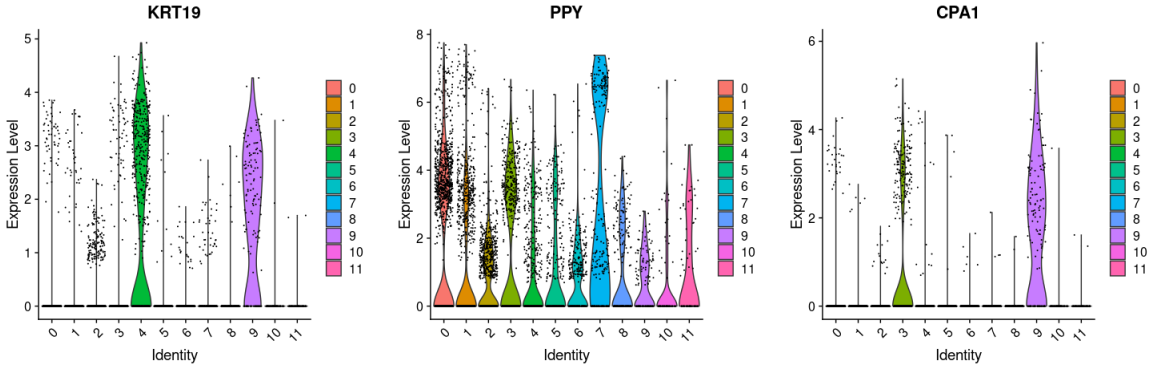
Supplementary Figure 1 - FeaturePlots of all genes used in cell typing. Marker genes are those provided in the Supplementary Data of Baron et al. Some genes used in the original analysis by Baron et al. were excluded, as they were not identified in significant amounts in our samples.



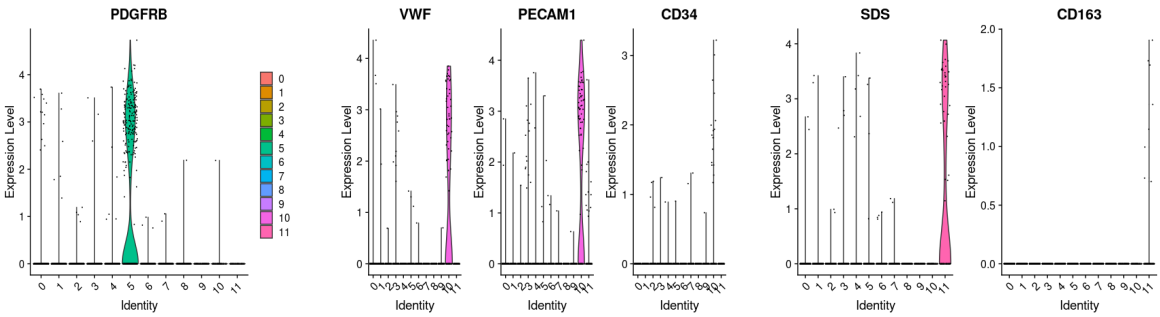
Supplementary Figure 2 - Violin Plots of all genes used in cell typing. Marker genes are those provided in the Supplementary Data of Baron et al. Some genes used in the original analysis by Baron et al. were excluded, as they were not identified in significant amounts in our samples.



Alpha Cells | Beta Cells | Delta Cells

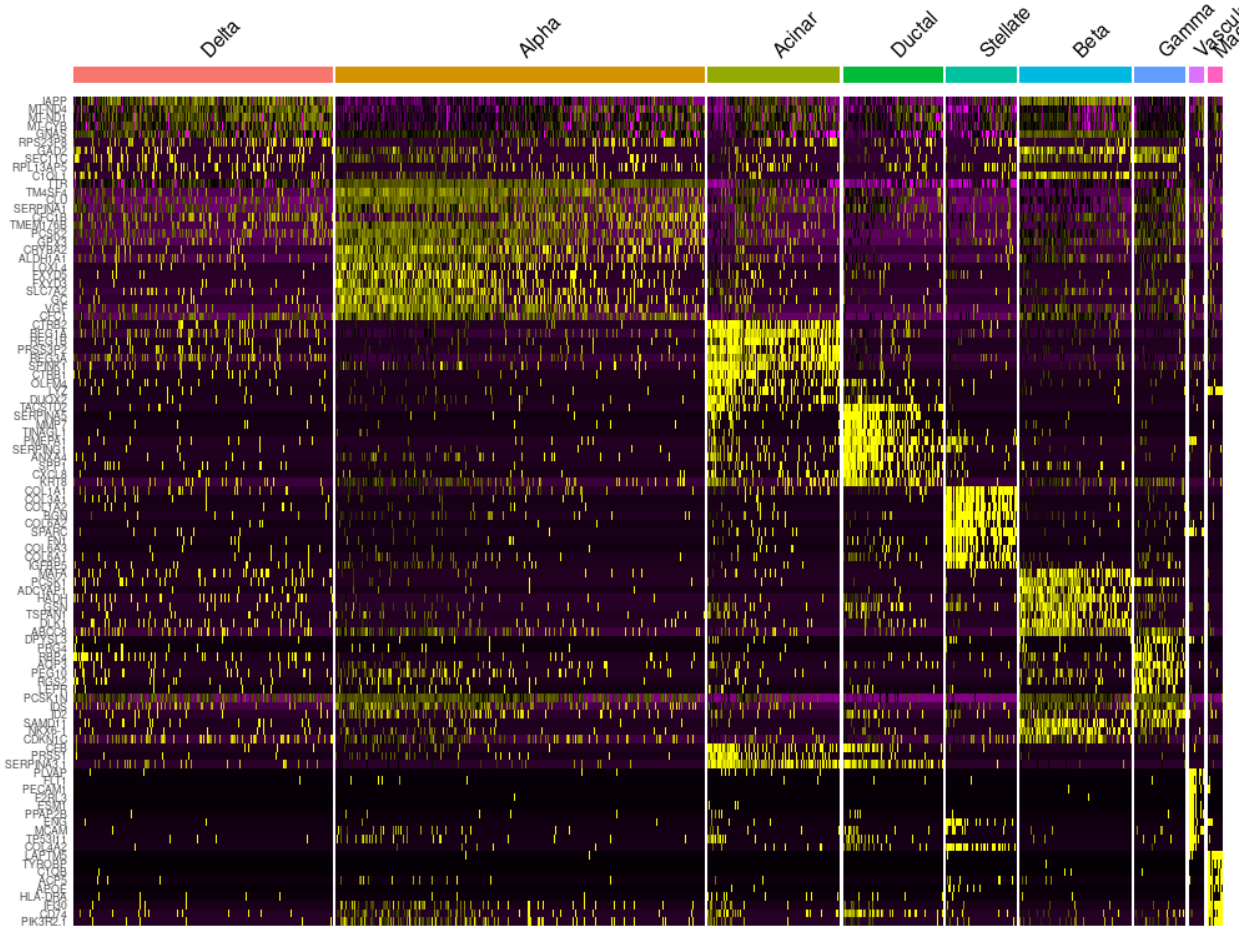


Ductal Cells | Gamma Cells | Acinar Cells



Stellate Cells | Vascular Cells | Macrophage Cells

Supplementary Figure 3 - Heatmap of novel marker genes, clustered by previously identified cell types. A total of 4548 novel marker genes were identified, using the parameters specified in the *Methods* section. The heatmap below was generated using the top ten marker genes for each cell type, by most significant average Log_2 Fold Change.



Supplementary Figure 4 - Heatmaps showing top GO term hits for unfiltered (top) and filtered (bottom) gene sets, with GO terms along the right vertical axis and clusters along the bottom. Dendrograms of cluster relatedness and GO term relatedness are on the top and right axes respectively. Coloration corresponds to the significance of the GO term association with the cluster.