Single Cell RNA-seq Analysis of Pancreatic Cells

Data Curator: Poorva Juneja

Programmer: Alice Zheng

Analyst: Kyler Anderson

Biologist: Priyanka Chary

Introduction

Mammalian pancreas contains a variety of cell types, each of which hosts a wide range of interactions that are essential for the body's function. In addition to regulating blood sugar levels through hormone production, the human pancreas also aids digestion. Abnormalities in these functions can have serious consequences. For example, in diabetes mellitus where the inability of the pancreas to produce sufficient insulin prevents glucose from being removed from the blood, Beta cells are associated with this process and if an abnormality occurs this glucose build-up in the blood can damage the kidneys and heart. Hence, a deeper understanding of the gene expression profiles and transcriptional activity of each of these cell types is important. Several studies on pancreas transcriptomes have already been bulk sequenced and profiled to provide insight into tissues, disease, and development through RNA-seq using the entire population of cells within a sample but there still remains a challenge in obtaining tissues from donors and developing a system that captures a sufficient number of cells.

In this 2016 study, Baron et al[1] implemented a droplet-based single-cell RNA-seq experiment in order to determine the transcriptomes of both human and mouse pancreatic cells of four individuals. Transcriptomics of single cells allows for the classification of cell types by function and expression of marker genes within cell populations. Single-cell RNA-sequencing (scRNA-seq) inDrop deals with the issue stated above by providing a systematic method for capturing thousands of cells without pre-sorting through high-throughput droplet microfluidics that barcode RNA from individual cells[1].

In addition to finding 15 distinct clusters to match with the previously known cell types they detected a new subpopulation of cells, based on their distinct expression profiles. For our analysis, we will be re-examining the data focusing on samples from one individual in the study using current analytical methodology and software packages.

Data

The data set was sourced from the GEO database[2] with the accession number GSE84133. The set included a total of thirteen sequencing libraries, across four individuals out of which samples from a 51-year-old female donor (SRR3879604, SRR3879605 and SRR3879606) were used as primary sequence libraries for our analysis.

Counting number of reads and whitelisting informative barcodes

To process the compressed FASTQ files a combination of command line-based text processing tools and batch commands were used to find the number of reads per distinct barcode. For our analysis, we used 'awk', which is a programming command that allows for efficient processing of large data files. With awk we extracted and combined the barcodes, which were further processed to generate a frequency count of each barcode.

These frequency counts for each sample were then used to create a cumulative distribution plot to determine cutoff for a whitelist with R. the counts were filtered by mean, generating a whitelist of barcodes with count greater than the mean. This whitelist allowed us to focus on the barcoded data with the highest counts. All three counts were then combined as one whole file for the next step.

Generating UMI Counts Matrix using Salmon Alevin

For generating the UMI matrix we first downloaded the v40 human transcriptome annotation file from the gencode database website[3]. Salmon was used to create an index, it was run as a batch job on SCC terminal for each SRR barcode sample 1 and 2 simultaneously along with the combined whitelist file from the previous step. Custom lengths were used for this step with a UMI length of 6, barcode length of 19 and end length of 5.

Methods

After the UMI matrix is produced, it is converted into a Seurat object to be processed for quality control. In the unfiltered dataset, there are a total of 17,288,239 genes and 131,457 cells observed, which can be found by looking at the number of rows of the Seurat object and calculating the sum of the nFeatures_RNA. First the percentage of mitochondrial genes are separated out for each cell, distinguished by the MT label at the front of the gene. The Seurat object now contains three different pieces of information. nFeature_RNA is the number of expressed genes, nCount_RNA is the number of transcripts, and percent.mt is the percentage of mitochondrial transcripts within a cell. In order to visualize the QC metrics prior to filtering, each of these three values are plotted using the VinPlot function on a violin plot in Figure 2. From the violin plots, we can see that for the majority of samples, there are very low nFeature and nCount values. This means that there are low gene counts and total transcripts in the cells. Looking at Figure 2 c, we can also see that for all the cells, there are very low amounts of mitochondrial transcripts, and the percentages are all close or equal to zero.

Next, we used the FeatureScatter function to create scatter plots as an alternative method of quality control and visualization of our data. The scatter plots show the relationship of the number of transcripts against the mitochondrial percents, and the number of transcripts in each cell against the number of expressed genes. The first plot allows us to see if there are any outlier cells which have a low number of transcripts but a high mitochondrial percentage. In Figure 3 a, we can see that most of the samples have both a low number of expressed genes and low number of mitochondrial transcripts. Additionally, the second plot checks the data quality to see if there is a positive correlation between the number of transcripts and the number of genes expressed. We can see that there is indeed a positive relationship of 0.98, which means that the quality of the data is good.

Bad cells were filtered by removing cells whose unique feature counts over 2,500, less than 200, or more than 5% mitochondrial counts. Cells with feature counts less than 200 indicate that there is a low gene count, and therefore no cell in the droplet. Cells with feature counts over 2,500 on the other hand have too high of a gene count, indicating that there are multiple cells in the droplet. Cells with mitochondrial transcript percentages over 5% indicated broken or dead cells

within the droplet, and were also removed. After filtering the dataset, there are 13,356 cells and 6,193,079 genes observed.

After filtering, the data was normalized using a global-scaling normalization method called "LogNormalize". This method works by first dividing the genes UMI count in a cell by the total number of UMIs in the cell. Then the feature expression ratios are multiplied by a scale factor of 10,000, and transformed using the natural log. This normalization step is important due to the differences in sequencing depths between cells.

The next step was to identify the highly variable features, which were genes having high expression variation from cell-to-cell. Because in scRNA-seq data, low expressing genes often have high variance, in order to identify the high variance genes, the variance must first be stabilized. Here, we used vst from the FindVariableFeatures package, or variance stabilizing transformation method, which first fits a curve to predict the variance of each gene using the mean expression. This prediction allows for the computation of a standardized count and standardized variance as a result. After obtaining the standardized variance values, the genes are then ranked, and the top 2000 are taken for use in PCA and clustering.

These results are then plotted using the VariableFeaturePlot function, which labels the top 2000 variable genes in red and the rest in black. In Figure 4, we can see that the variance for the majority of the genes falls around 1, while the variance for the red, high variance genes falls around 2.5. Genes with the highest amounts of variance also had higher levels of average expression. The 5 genes which had the highest variance were TPSB2, .COL1A1, TPSAB1, ALB, and CTRB1. In order to perform PCA on the top 2000 variable genes, first the data needs to be scaled in order to remove unwanted sources of variation. Scaling is done using the ScaleData function, which shifts and scales the expression of each gene so that the mean expression is 0 and the variance is 1. This scaling is used so that the highly variant genes do not dominate the results in the following analyses.

After scaling the data, PCA is performed, using the RunPCA function which takes around a minute. After running PCA, the principal components can be plotted. In Figure 5, we can see the PCA plot of the first two components, and there is slight clustering into three streaks, which could be representative of different gene groups. Using an elbow plot from the ElbowPlot function, it is possible to find out which principal components are significant and contribute the

most to use for clustering. In Figure 6 the elbow plot shows the standard deviation of the principal components, meaning the amount of variation explained by each principal component. Based on the elbow plot, 10 features are selected, since around PC 10 there is a plateau in the variance standard deviations. It is also recommended to have more principal components as opposed to less.

After identifying the number of dimensions we want to use, the FindNeighbors and Find Clusters functions are used to perform clustering. After obtaining the clustering data and identifying the number of clusters, it is possible to discover the number of cells which belong to each of the 11 clusters. Using this data, a pie chart is constructed of the relative proportion of cells in each cluster. From Figure 7, we can see that cluster 0 has the highest proportion of cells, and as the cluster number increases the proportion decreases.

Results

Sample associated with the 51-year-old female- SRR3879604, SRR3879605 and SRR3879606 indicated 1 to 2 million counts of cell barcodes per sample. After filtering data by mean count to eliminate the reads with infrequent counts we were left with a combined total of 235924 samples. With the filtered reads counts to determine cutoff for a whitelist the cumulative distribution plotted (figure1). Results from the salmon alevin mapping statistics showed a mapping rate of 43.3055% with a target index of 245,261.

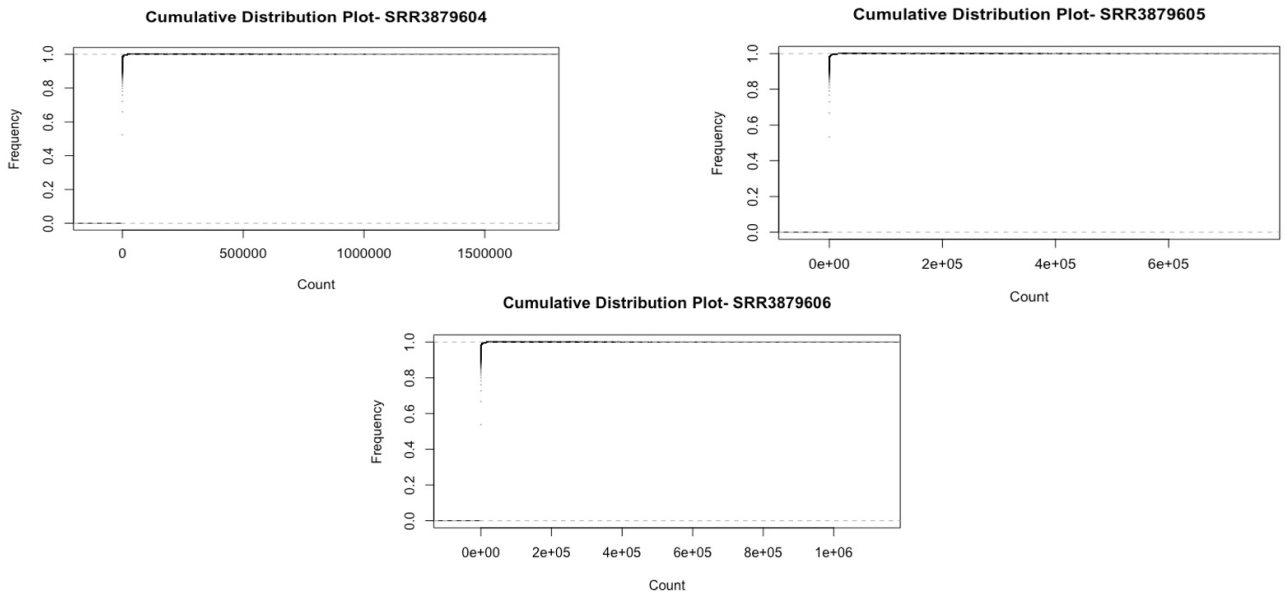

Figure 1: Cumulative distribution plot for single RNA-seq barcode of Samples

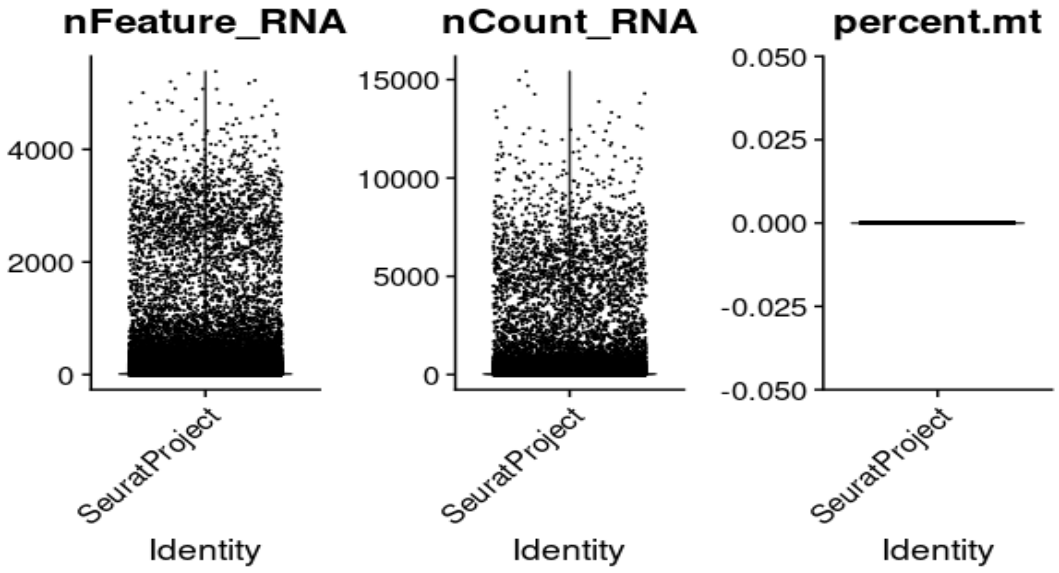

Figure 2 (a, b, and c): Violin plots of nFeatures which is the number of genes in each sample, nCount which is the number of transcripts in each sample, and percent.mt, which is the percentage of mitochondrial transcripts in each sample.

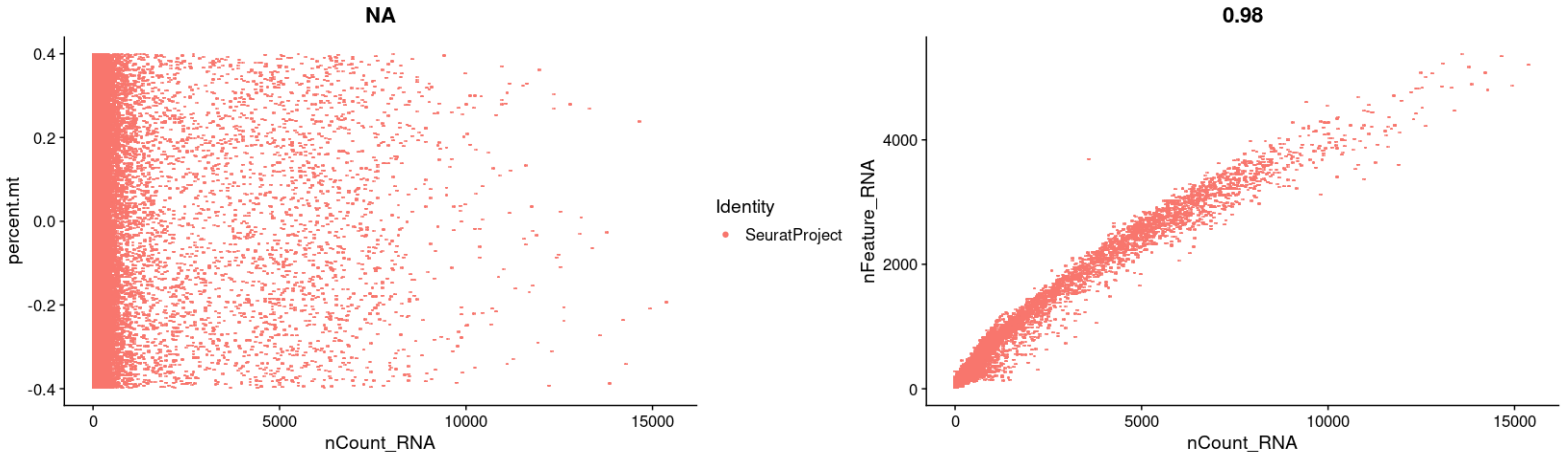

Figure 3 (a and b): Scatter plot of the mitochondrial transcript percents versus the number of total transcripts in each sample, and scatter plot of the number of genes in each sample versus the total number of transcripts in each sample.

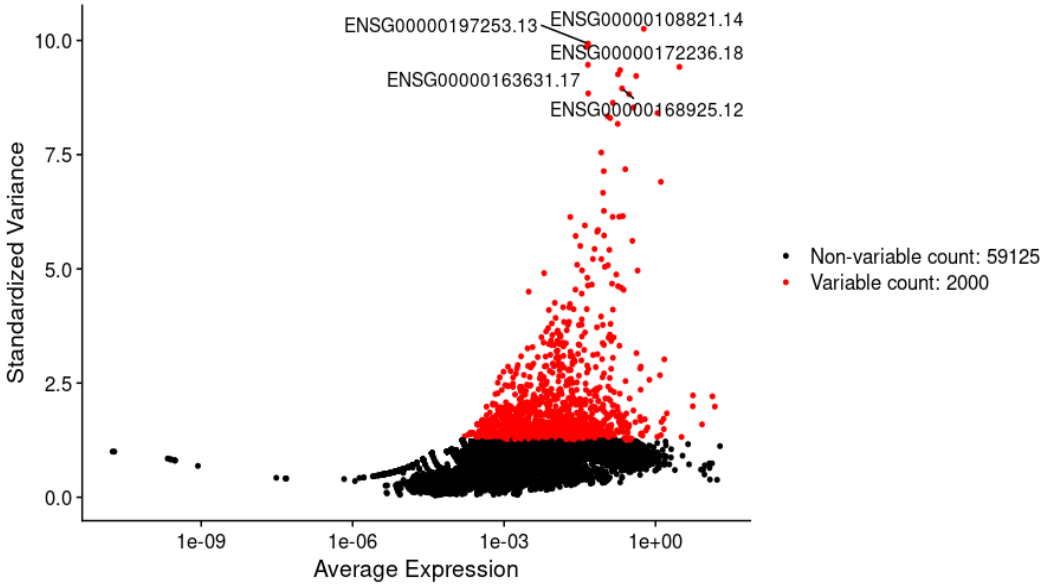

Figure 4: Scatter plot of the standardized variance against the average expression of each gene. The top 2,000 high variance genes are labeled in red, while the rest of the low variance genes are in black. The top 5 genes with highest variance are labeled.

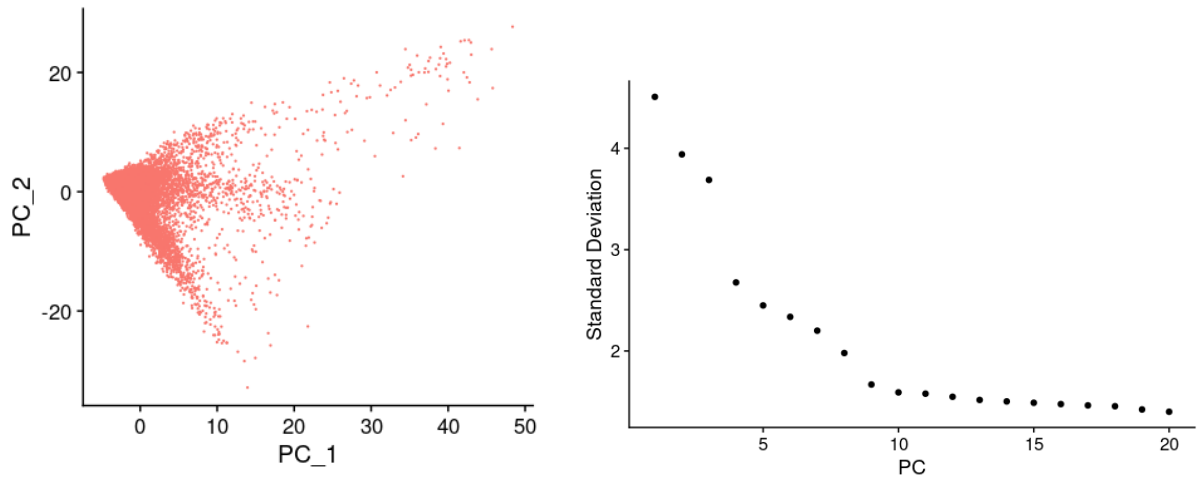

Figure 5: PCA of the cell gene expression data.

Figure 6: Elbow plot of the principal components

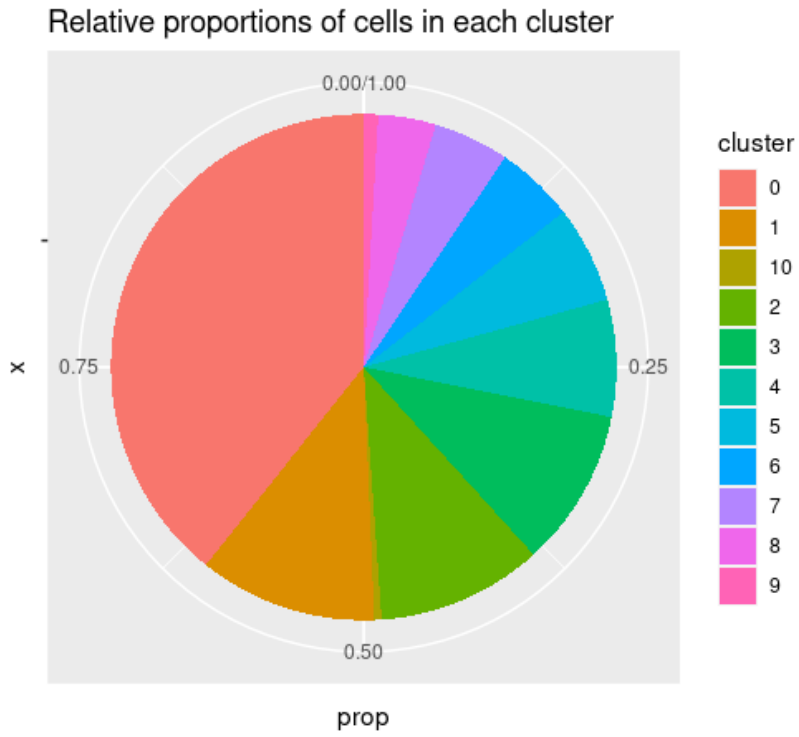

Figure 7: Relative proportions of the number of cells for each of the clusters identified.

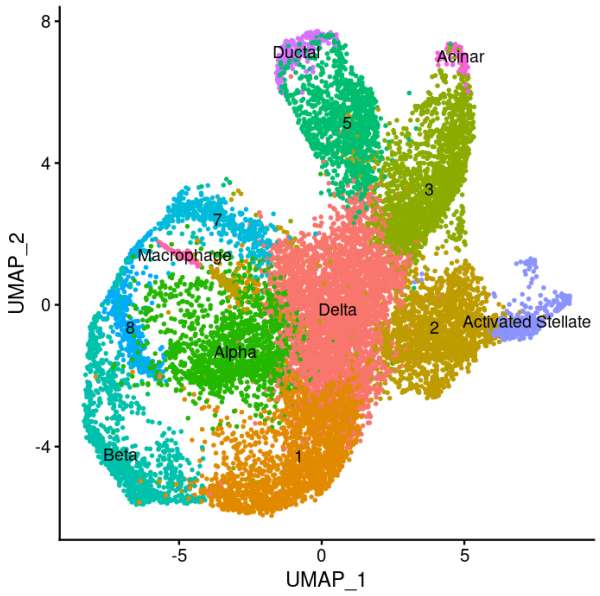

Figure 8: Pre-computed clusters with cell types assigned conservatively. Clusters 1, 2, 3, 5, 7, and 8 proved too ambiguous for assignment; particularly cluster 7 which contained no marker genes used by Baron et. al.

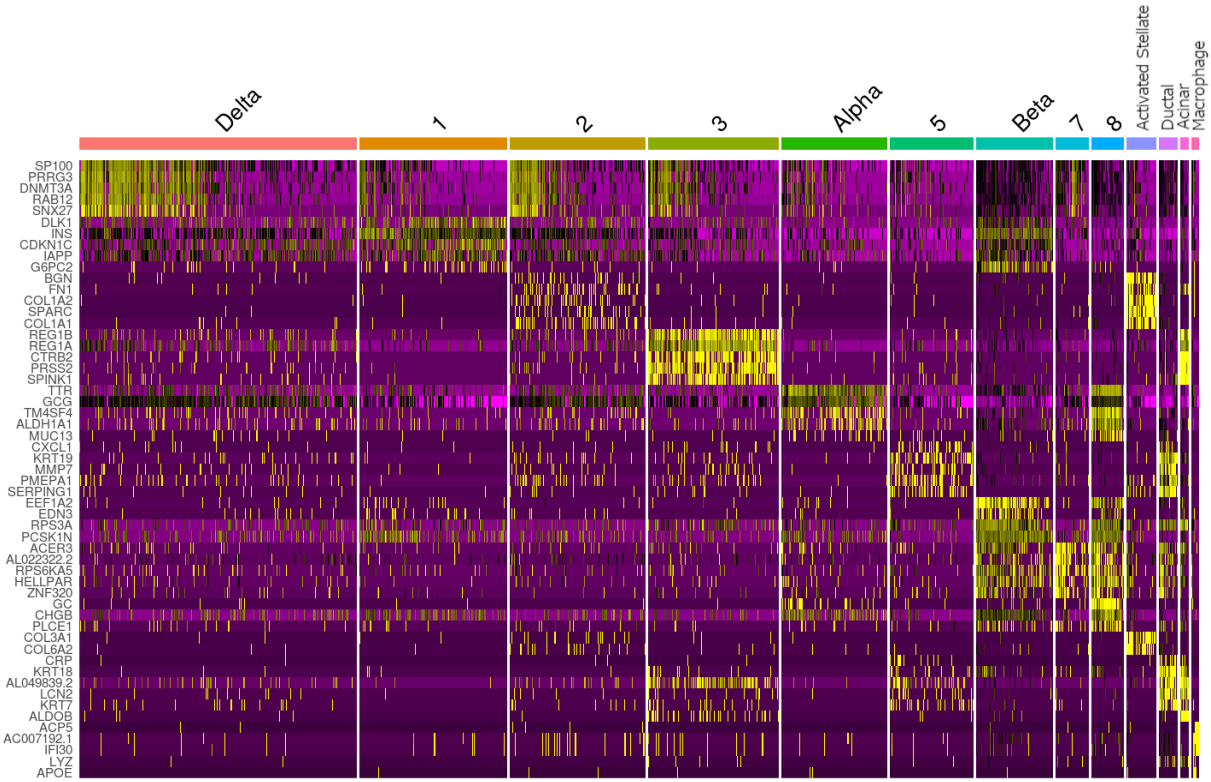

Figure 9: Heat map of the top five marker genes for each cluster as determined by Seurat. Only half the cluster assigned as Delta cells share expression of the top markers, with significant flares in these genes from clusters 2, 3, and 7. The top genes for clusters 1 and 2 only show weakly and flare again for the Beta cell and Activated Stellate clusters. The Alpha and Beta cells occupy most of the markers for cluster 8, and cluster 5's markers show stronger in the Ductal cells.

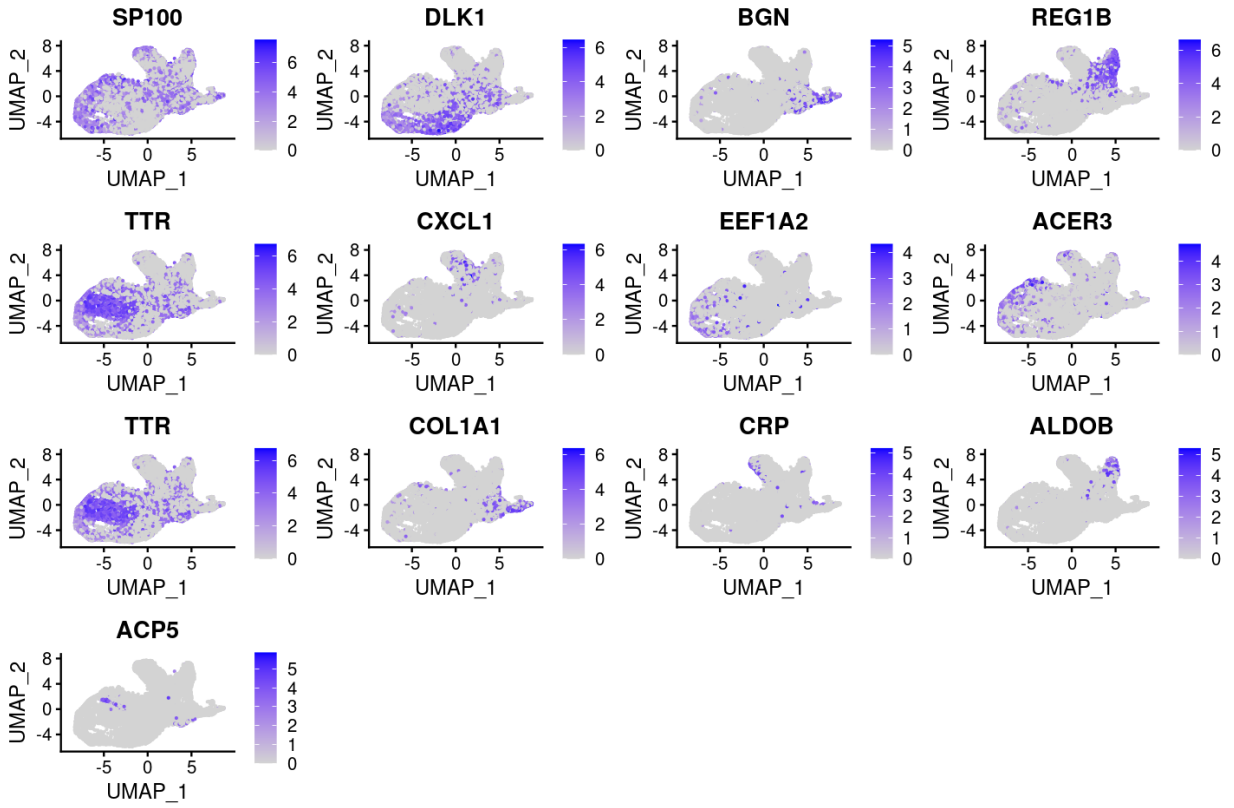

Figure 10: Analysis of the top novel marker gene by log_2FC for each cluster (in order from left to right, top to bottom). In all cases the marker does not localize to one cluster. Two clusters (4 and 8) share their top marker.

Cluster	David Terms	David Terms for top 10 genes
0	Protein binding NADH activity Mitochondrion	NADH dehydrogenase (ubiquinone) activity Mitochondrial respiratory chain complex I Negative regulation of transcription from RNA Polymerase II promoter
1	Plasma membrane Extracellular region	Extracellular space Mitochondrial respiratory chain complex I

	GTP binding	NADH dehydrogenase (ubiquinone) activity
2	Protein binding Endoplasmic reticulum lumen Extracellular region	Endoplasmic reticulum lumen Extracellular space Extracellular region
3	Protein binding Extracellular exosome Structural constituent of ribosome	Extracellular space Proteolysis Serine-type endopeptidase activity
4	Protein binding Receptor binding Hormone activity	Extracellular region Hormone regulation Endoplasmic reticulum lumen
5	Protein binding RNA binding Identical protein binding	Protein binding Extracellular exosome CXCR chemokine receptor binding
6	Integral component of membrane Nucleus RNA binding	Translation Cytoplasmic translation Structural constituent of ribosome
7	Protein kinase activity NADH dehydrogenase activity Cytochrome-c oxidase activity	Axon guidance
8	Protein binding Phosphoprotein Acetylation	Extracellular exosome Glycoprotein Integral component of membrane
9	Protein binding Platelet-derived growth factor Binding Extracellular exome Cytosol	Extracellular region Collagen trimer Endoplasmic reticulum lumen
10	Protein binding Cytosol Wound healing	Extracellular space Cell surface Extracellular exosome
11	Protein binding Extracellular exosome Structural constituent of ribosome	Extracellular space Proteolysis Serine-type endopeptidase activity
12	Protein binding Plasma membrane Extracellular exosome	Integral component of membrane Identical protein binding Integral component of plasma membrane

Table 1: The table above shows each cluster's DAVID gene enrichment results. The middle column shows the enrichment for all genes in the cluster, while the last column shows the enrichment for the top 10 differentially expressed genes in each cluster.

In order to analyze each cluster's biological function, the gene symbols from each group were extracted and uploaded to the DAVID Functional Annotation Tool. Then the most counted annotations were documented to form the table above. For the last column of the table, the rows were ordered in order for ascending p-values. The smallest p-values (most differentially expressed genes) were shown at the top of the list. Only the top ten gene symbols from the chart were extracted and uploaded to DAVID's functional annotation tool to show the gene enrichment results for the most differentially expressed genes.

Discussion

The sequencing libraries were processed into UMI counts matrices using salmon alevin. The mapping statistics were noted and the results on how well the reads mapped were recorded. Next, the UMI counts matrix was filtered based on the quality of cells, and low variance genes. The counts matrix was then normalized and eleven clusters were formed as shown in the Figure 6 pie chart.

The Seurat R package provides tools for finding marker genes for each cluster, particularly the FindAllMarkers method. This method compares each cluster to all others in turn, looking for genes whose expression corresponds to the cluster designation. Pre-calculated clustered cells were analyzed for markers in this way, with a minimum $\log_2$FC of 0.25. Of the 23 markers used by Baron et. al. only 12 were assigned to these clusters, with many overlapping. Notably, cluster 7 did not contain any of the markers used by the original paper. Judging by $\log_2$FC, percent of expression, and the violin plots shown in Supplementary Figure S1, cell types were assigned to 7 of the clusters, shown in Figure 8 on a UMAP projection. The markers GCG, INS, SST, and PPY – for Alpha, Beta, Delta, and Gamma cells respectively – are particularly ambiguous for their presence across all clusters (Figure S1). Based on $\log_2$FC values, GCG and INS were deemed most characteristic for clusters 4 and 6, though other measures, including p_{adj}, were not in consensus. Disagreement was frequent among the typical marker genes; such statistics are provided in Table S1.

To observe the general separability of the clusters by any marker genes, a heat map, shown in Figure 9, of log normalized UMI counts for the top 5 marker genes for each cluster was produced with Seurat. Many of these genes are not typical marker genes, though their novelness is abated by the noise of the heatmap. Indeed the latter, smaller clusters often share expression levels of similar sets of genes to larger clusters. The expression of the markers within the clusters, especially for clusters 1 and 2 which could not be identified, is often highly inconsistent.

To assess the performance of potential novel marker genes, the expression levels of the top marker gene by $\log_2$FC for each cluster not used by Baron et. al. was observed overlaid on the UMAP projection (Figure 10). The expression regions often overlapped heavily, had many outliers, and did not cover the entire set of cells. The top ten marker genes for each cluster were extracted for enrichment analysis.

The results from the cluster analysis lend to several conclusions. First, it is likely that the underlying clustering was poor quality, or performed over an uninformed projection space. This is indicated by the failure of typical and novel marker genes to sufficiently distinguish the clusters. Second, as this analysis was performed on only one sample, it may be prone to biases specific to the sampled individual. Pooling over multiple individuals would lead to less noisy and more robust clustering. Furthermore, it may simply be inadequate to rely on only a small set of marker genes for identifying cell types in such a complex setting. Baron et al. used more marker genes for clusters with fewer cells; in order to make up for the loss of information due to cluster size, they introduced more markers so their assignments were still well supported. In this case, since only one individual was used, the lack of data seemed to affect even the large clusters, and typical markers were absent or ambiguous.

The genes were uploaded to DAVID for functional annotation. This process was repeated with the top ten genes with the smallest p-value, to analyze gene enrichment in the most differentially expressed genes. A table of the results was constructed in Table 1. Overall, many of the gene enrichments showed the most counts for protein binding. While they were included in the table for consistency, they are not very telling of the cell type of the cluster since they applied to all clusters. However, there were other enrichment terms for each cluster. Generally, alpha cells are endocrine cells that secrete the peptide hormone glucagon. In the DAVID chart it can be seen that cluster four would most likely fit this description. This matches with the graph provided in the

previous analysis as well as the analysis by Baron et al. Beta cells, most commonly associated with insulin production in the pancreas could also be related to cluster four. However, in the analysis and according to Baron et al., beta cells were identified in cluster six. Since the DAVID terms associated with cluster six were related to translation, this could possibly indicate the proteins related to insulin production were being created. The gamma cells play a large role in the immune system and T cells. While the paper by Baron et al. suggests that cluster zero may be identified as gamma cells, both our paper's results and the DAVID terms do not strongly suggest this. Ductal cells, commonly associated with the pancreatic lining, have been identified as cluster six both in prior analysis and the paper by Baron et al. The DAVID terms shown in Table 1 mention wound healing and cell surfacing which would be understandable in this context. Stellate cells provide the liver with an ability to respond to injury and heal certain types of damage. [4] In relation to the DAVID terms, this cell type would be cluster nine or ten. Cluster nine mentions collagen, found in connective tissue, skin, tendon, bone, and cartilage, binding, and platelet-derived growth factor. Cluster ten mentions wound healing. According to previous analyses, stellate cells would be strongly associated with cluster nine. However, Baron et al.'s supplementary table identifies cluster zero as stellate cells. Macrophage cells have been identified as cluster twelve in both our analysis as well as Baron et al's work. Macrophages, a type of white blood cell, would be associated with plasma membranes and extracellular exosomes as suggested by the DAVID terms.

Overall, our analysis of the clusters differed from Baron et al's paper. This could have been due to the example datasets or the repeat DAVID enrichment terms which provided a level of ambiguity.

Conclusion

While Baron et al. proved their ability to determine transcriptomes of thousands of pancreatic cells, map them to known cell type identities, and point to their functional roles, our analysis and results are not as conclusive. Between the analysis of the cluster cell type labels and DAVID terminology there were inconsistencies.

The main limitations of our analysis were brought up in the cluster labeling and the enrichment terms. While Baron et al analyzed twelve cell types in their analysis, our replication of their work was only able to identify seven. This was potentially due to the different use of data since the

authors pooled four individuals and used their own unique analysis method, while we used the general purpose Seurat package. Additionally, we were limited due to the repetitive DAVID enrichment terms. Many of the terms were repeated for many clusters making the cell type rather ambiguous. Future direction for this work would be to repeat the analysis using Baron et al.'s approach and using another enrichment method such as EnrichR or Metascape.

References

1. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM, Melton DA, Yanai I. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst. 2016 Oct 26;3(4):346-360.e4. doi: 10.1016/j.cels.2016.08.011. Epub 2016 Sep 22. PMID: 27667365; PMCID: PMC5228327.
2. 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.

3. *Genecode*, EMBL-EBI, www.gencodegenes.org/human/
4. Friedman SL. Hepatic stellate cells: Protean, multifunctional, and enigmatic cells of the liver. Physiol Rev. 2008; 88:125–172.

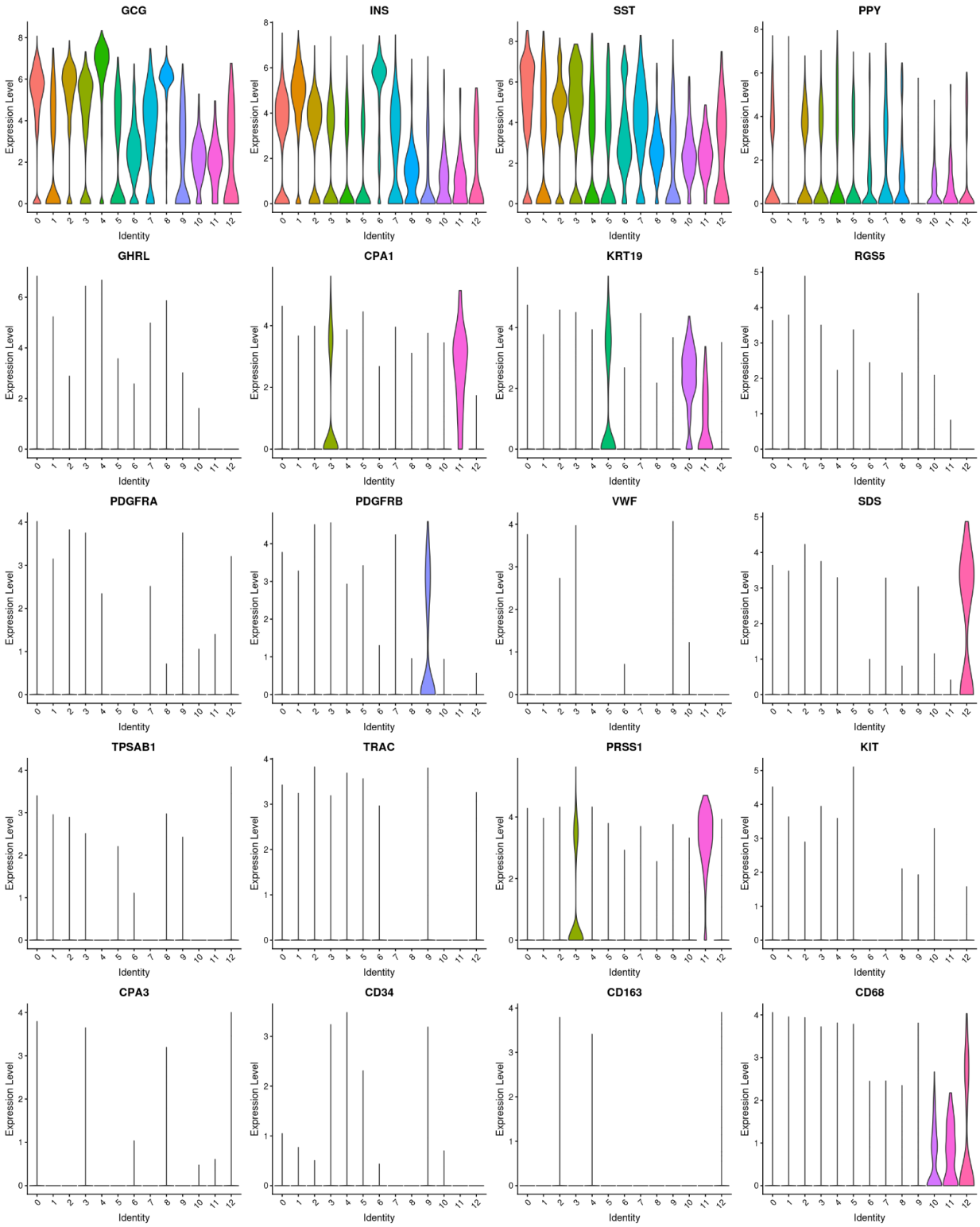

Figure S1: Violin plots indicating expression levels for 20 of the marker genes as used by Baron et. al. Three markers were not present in the filtered counts, and several of the markers have extremely sparse expression levels. The markers GCG, INS, SST, and PPY – for Alpha, Beta,

Delta, and Gamma cells respectively – are particularly ambiguous for their presence across all clusters.

Table S1: Gene marker statistics for marker genes used by Baron et. al. as determined for the single-individual clustering analysis.

Gene	Cluster	log_2FC	pct_1	pct_2	p	p_{adj}
SST	0	1.278	0.875	0.717	8.47e-301	1.85e-296
PPY	0	0.579	0.408	0.359	4.11e-20	8.977140e-16
INS	1	1.758	0.937	0.62	0	0
GCG	2	0.369	0.922	0.779	2.11e-102	4.61e-98
CPA1	3	3.252	0.343	0.033	0	0
PRSS1	3	2.762	0.33	0.04	0	0
SST	3	0.289	0.806	0.751	1.08e-32	2.36e-28
GCG	4	2.246	0.943	0.781	0	0
KRT19	5	2.735	0.44	0.096	1.60e-286	3.51e-282
INS	6	2.183	0.949	0.641	4.35e-294	9.50e-290
GCG	8	0.599	0.993	0.791	4.30e-60	9.39e-56
PDGFRB	9	3.121	0.403	0.011	0	0
PDGFRA	9	1.157	0.144	0.002	0	00
KRT19	10	1.816	0.892	0.109	1.56e-283	3.41e-279
CD68	10	0.833	0.479	0.038	8.06e-245	1.76e-240
PRSS1	11	3.584	0.957	0.068	3.41e-293	7.46e-289
CPA1	11	2.939	0.914	0.063	4.50e-267	9.83e-263
CD68	11	0.974	0.629	0.041	5.59e-197	1.22e-192
SDS	12	4.271	0.587	0.005	0	0
CD163	12	2.025	0.173	0	0	0
CD68	12	2.229	0.288	0.044	3.14e-35	6.86e-31