Single Cell RNA-Seq Analysis of Pancreatic Cells

David Lenci, Nikita Tomar, and Daniel Gealow

Introduction

Single cell RNA-sequencing is a next generation sequencing technique that allows us to better understand the cellular and genomic landscape of cells in similar biological settings. In *Baron et al: A Single Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure,* the researchers sought to determine the transcriptomics of over 12,000 individual pancreatic cells through single cell RNA-seq, and by doing so gain a better understanding of the common and uncommon cell types found in the human and mouse pancreas [1]. Our goal for this project is to replicate the results of the human pancreas analysis done by the researchers in *Baron et al.*

Data

The authors of *Baron et al.* used inDrop single-cell RNA sequencing technology to build transcriptome libraries of pancreatic cells from four deceased human donors and two strains of mice. We restricted our analysis to the three batches from a single human donor, a 51-year-old female with a BMI of 21.1 who did not have type 2 diabetes (GEO accession number GSM2230758). The inDrop protocol attaches a cell-specific barcode and a unique molecular identifier (UMI) to the start of each transcript fragment prior to amplification and Illumina paired-end sequencing. We were given preprocessed read 1 files with the barcode and UMI lengths and positions standardized, and corresponding read 2 files containing the actual transcript sequence. According to the paper, there should have been about 800 cells with an average of 10^5 reads per cell, from each of the three batches. We wrote shell and python scripts to count the number of reads per barcode and generate a plot of the distribution of reads in each sample (Figure 1).

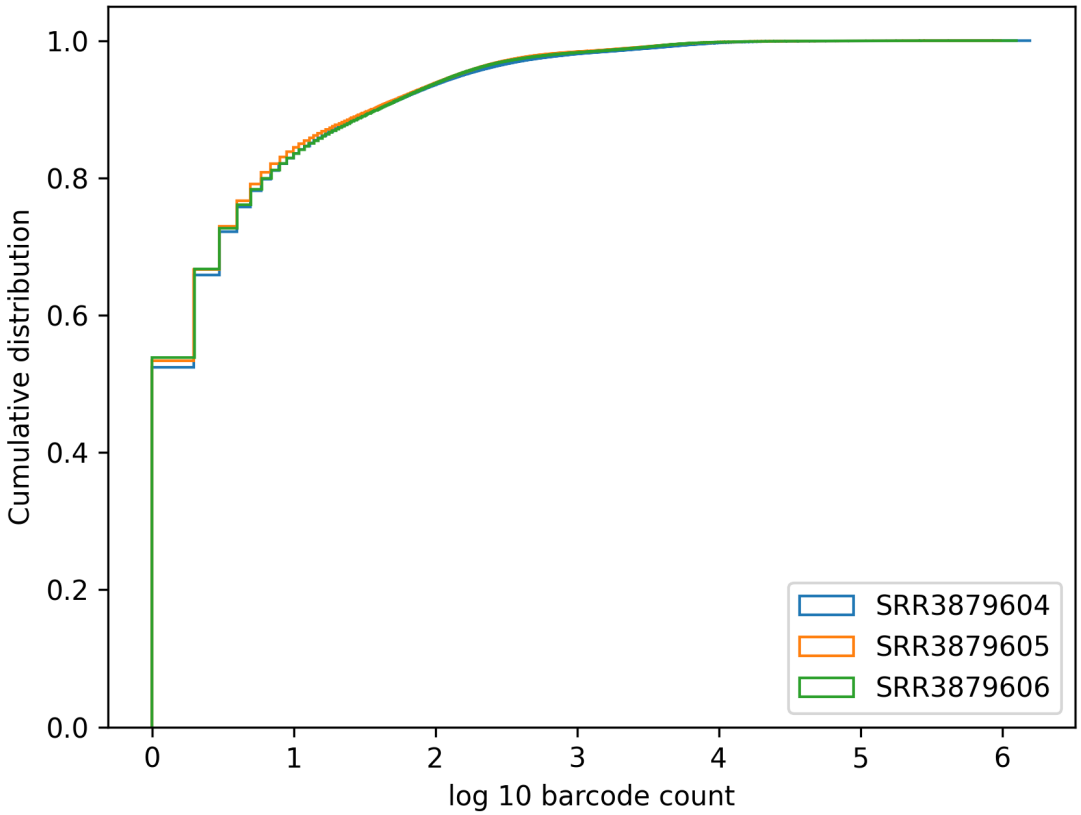

Figure 1. Cumulative distribution of barcode counts for each batch. Even with logarithmic scaling, all three samples have a very skewed distribution, with only a single read associated with about half of the barcodes, ~80% having fewer than 10 reads, and ~90% having fewer than 100 reads.

There was a very large proportion of noisy reads; that is, barcodes with a very small number of counts which were likely the result of sequencing errors and do not correspond to a specific cell. Therefore, we wanted to establish a cutoff of a certain number of counts below which a barcode would be filtered from the data. We plotted the distribution again, but zoomed in on the high end and changed the parameters to produce a plot of how many barcodes would be included for a given filter value (Figure 2).

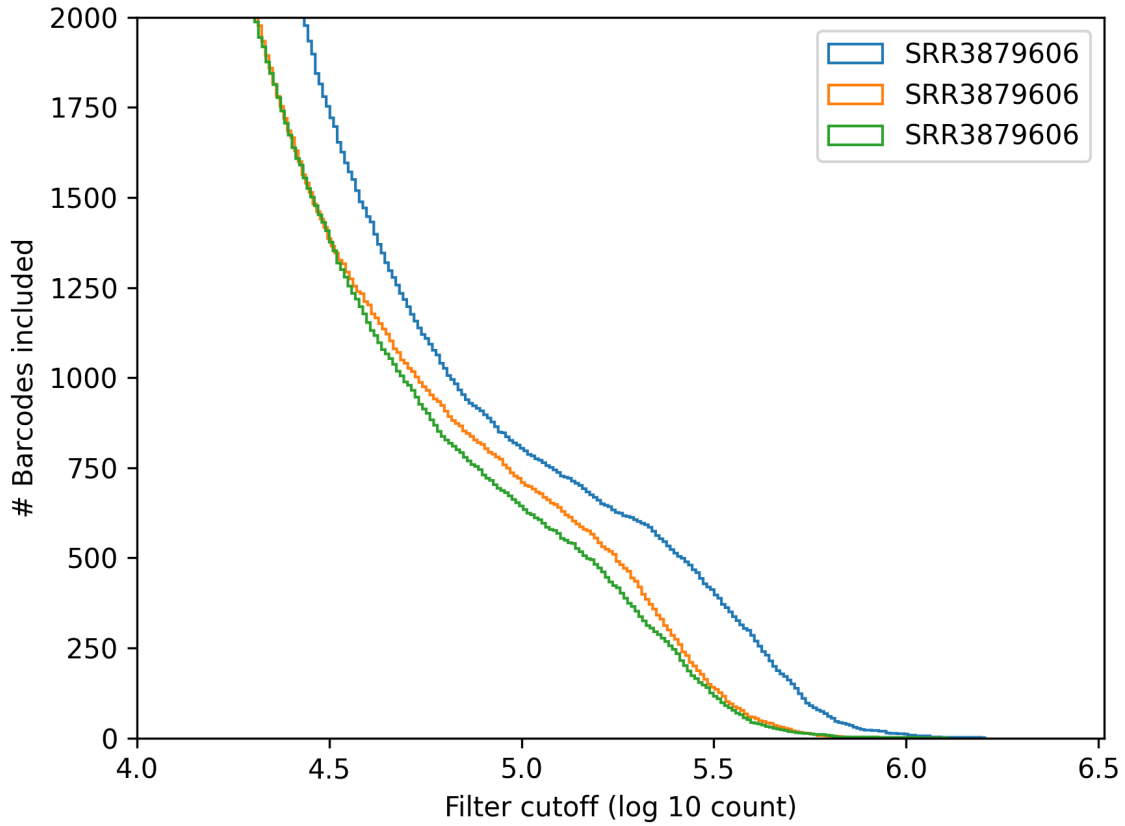

Figure 2. Number of barcodes remaining vs. filter cutoff. The y-axis indicates how many barcodes would be included as valid if a count filter were used with a cutoff of 10 to the power of the x-axis value.

For all the samples, the distribution of barcodes begins to taper off as the filter decreases from 5.5 to 5.0, but then it passes an inflection point and the number of barcodes begins increasing rapidly as the cutoff decreases past 4.5; the “knee” of the plot looks to be around 4.75. A filter cutoff of 10^4.75 also results in slightly more than 800 barcodes being included from each sample, which was the batch size reported in the paper, and leads to an average count on the order of magnitude of what was reported, so we went ahead and used this as our cutoff. We wrote another python script to generate a whitelist file containing all barcodes that had at least 10^4.75 reads in any of the three samples.

Next, we prepared to run alevin by creating a salmon index using the Gencode v40 human reference transcriptome. We used a k of 19 rather than the default 31, since most of the

transcript reads the index would be used for were only about 43 bp long. We also downloaded a transcript ID to gene ID mapping file from Ensembl BioMart's web database interface. Finally, we ran salmon alevin with the provided barcode and read FASTQ files; the index and transcript-to-gene mapping files; and our whitelist. This generated a matrix of the number of UMIs associated with each detected gene in each included barcoded cell in the samples.

Methods

Processing the UMI counts matrix

Having the pancreatic cells dataset generated, we analyzed it using Seurat which is the most popular package for single-cell RNA-seq data analysis. As the first step we used the tximport functionality of the tximport package [5] to read the data and specified the type as "alvin". It summarizes the transcript level information to the gene-level and returns a UMI matrix.

Mapping gene symbols and filtering genes and cells

The rows of the UMI matrix containing the Ensembl gene identifiers (ENSG IDs). We mapped these IDs to their respective gene symbols using EnsDb.Hsapiens.v79 package [2]. A Seurat object was created using the UMI matrix with a filter of data containing the genes expressed in at least three or more cells and cells containing at least 200 detected genes. The initial dataset consisted of 61125 genes. After filtering only 56063 genes remained. After creating the Seurat object we calculated the percentage of counts belonging to a set of mitochondrial genes using the PercentageFeatureSet functionality. We used the pattern "MT-" to identify the mitochondrial genes. Figure 3 shows a violin plot representing the visualization of the counts, features and percent. We filtered out cells with unique feature counts not in the range of 200 to 2500 and having over 5% of mitochondrial counts. Figure 4 shows the feature to count and feature to percent of mitochondrial counts relationships of the filtered dataset. After applying these filters 27006 genes remained.

Processing filtered data

After the filtering was done, the data was then normalized using the NormalizeData function and LogNormalize method. This method uses the feature expression for each cell and normalizes it by the total expression using a scale factor of 10,000 which is the default value. We then subset by filtering out the low variance genes. This was done using the FindVariableGene function in Seurat. We selected 2000 high variable genes using the “vst” method. This method calculates the relationship between mean expression and variance that is inherent to single cell RNA-seq. Figure 5 shows the plot for the standard variance vs average expression and also labels the top 10 genes with high variance. Since the data may contain variation from batch effects, technical or biological noise, scaling the data was necessary to remove the unwanted sources of variation. We used the ScaleData function available in Seurat.

Dimension Reduction

On the scaled data linear dimensional reduction was performed using the RunPCA() and VariableFeature function. Figure 8 shows the top genes associated with the first two principal components and Figure 9 shows the scatter plot of cells in the first two principal components. The JackStraw function was used to identify significant principal components with strong enrichment of low p-value genes. Figure 8 shows the distribution of p-values for each principal component using the JackStrawPlot function. Alternatively, the ElbowPlot function was used to plot the data as well in Figure 11.

Cell Clustering

Graph based clustering method was used which is the default in Seurat for identifying the cell clusters. The KNN graphs (k-nearest neighbors) were calculated based on the euclidean distance in PCA space and the edge weights between any two cells were refined based on the shared overlap in their local neighborhoods. FindNeighbors function was used to conduct this step. For clustering the cells the default Louvain algorithm was applied to group the cells. A resolution of 0.5 was set during the clustering using the FindClusters function.

Using the RunUMAP function of Seurat, non-linear dimensionality was explored and visualized. This algorithm helped to learn the underlying manifold of data so that the similar cells can be placed together in low-dimensional space. The clusters are shown in figure 12.

Identifying Markers and Labeling Clusters:

With the processed and clustered counts we then sought to identify marker genes for our clusters. To do this we utilized Seurat's built in differential expression analysis tools, which allowed us to identify genes that were significantly differentially expressed in each cluster compared to all other clusters. We then looked at the top 10 differentially expressed genes based on average log2 fold change, in order to label our clusters according to the expected cell types. Using these potential marker genes alongside *The Human Protein Atlas* and PanglaoDB, a database that has collected results from single cell expression analyses in order to provide known markers for different cell types, we were able to label our clusters according to cell type [3,4]. Initially, Seurat identified 13 clusters and through labeling and combination of clusters we were able to reduce that number down to nine, which is the number of clusters the researchers identified.

Results

There were originally 61125 genes in total while importing the alevin data using tximport. After mapping the ENSEMBL IDs to gene symbols 56063 genes remained. Creating the Seurat object using the filter for removing genes with less than 3 cells and less than 200 detected genes we get 27006 distinct genes at the end.

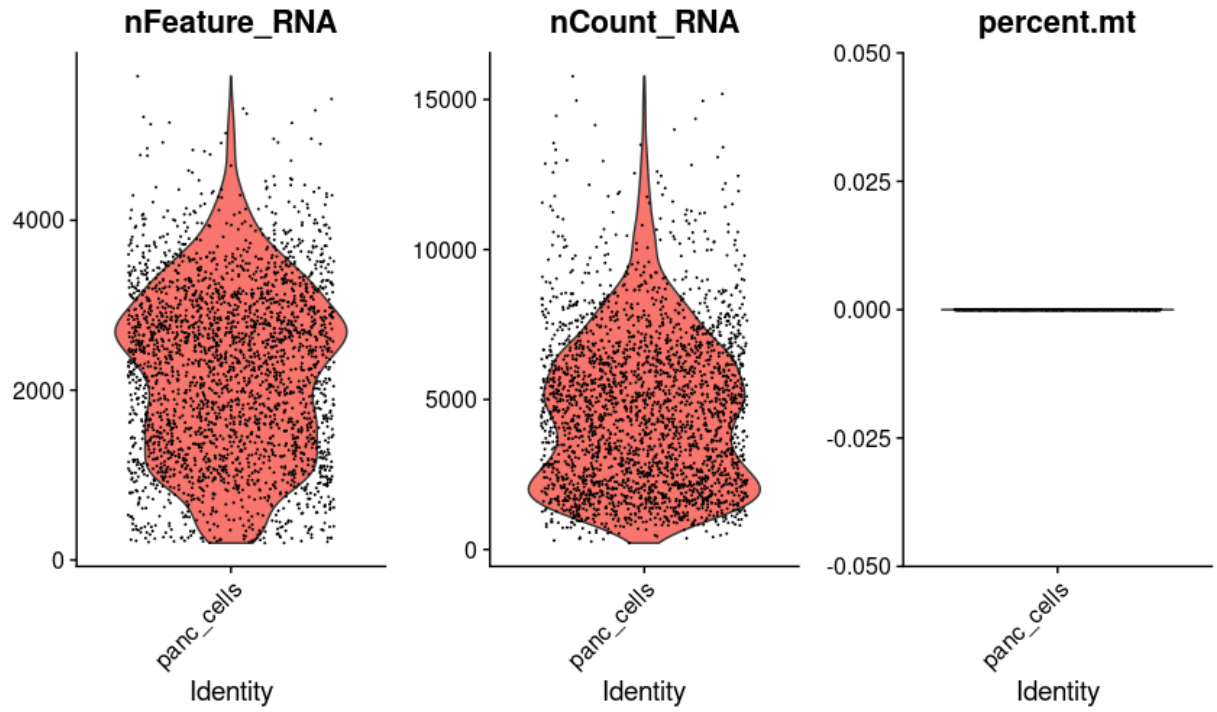

Figure 3: Violin plots for the number of reads, expressed genes and percentage of mitochondria genes for each cell.

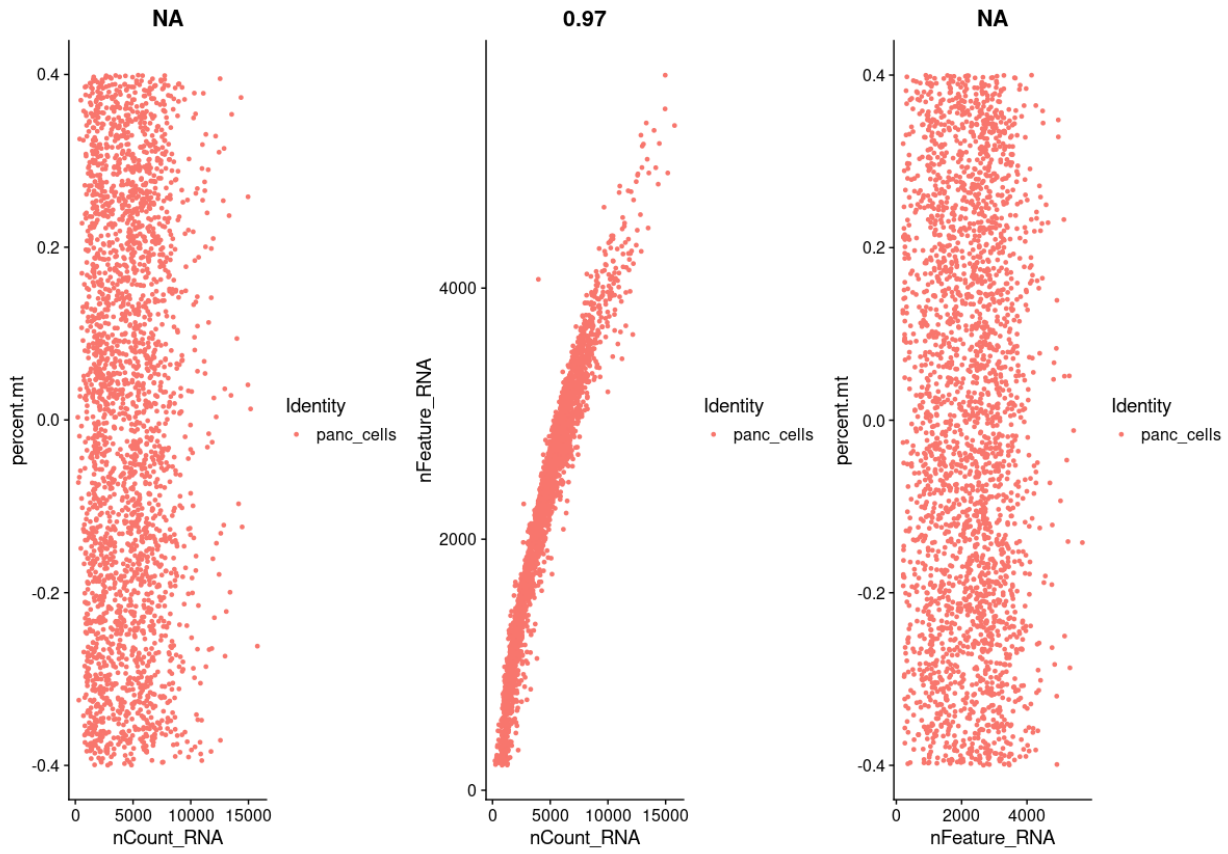

Figure 4: Scatter plots for the reads-expressed genes and reads-mitochondrial genes relationships.

Our original data did not show a high percentage of mitochondrial genes. Nevertheless we filtered by removing cells showing greater than 5% of mitochondrial genes. Low quality cells were also the ones that had fewer than 200 genes or more than 2500 genes, which were also filtered out. All of the 27006 genes remained.

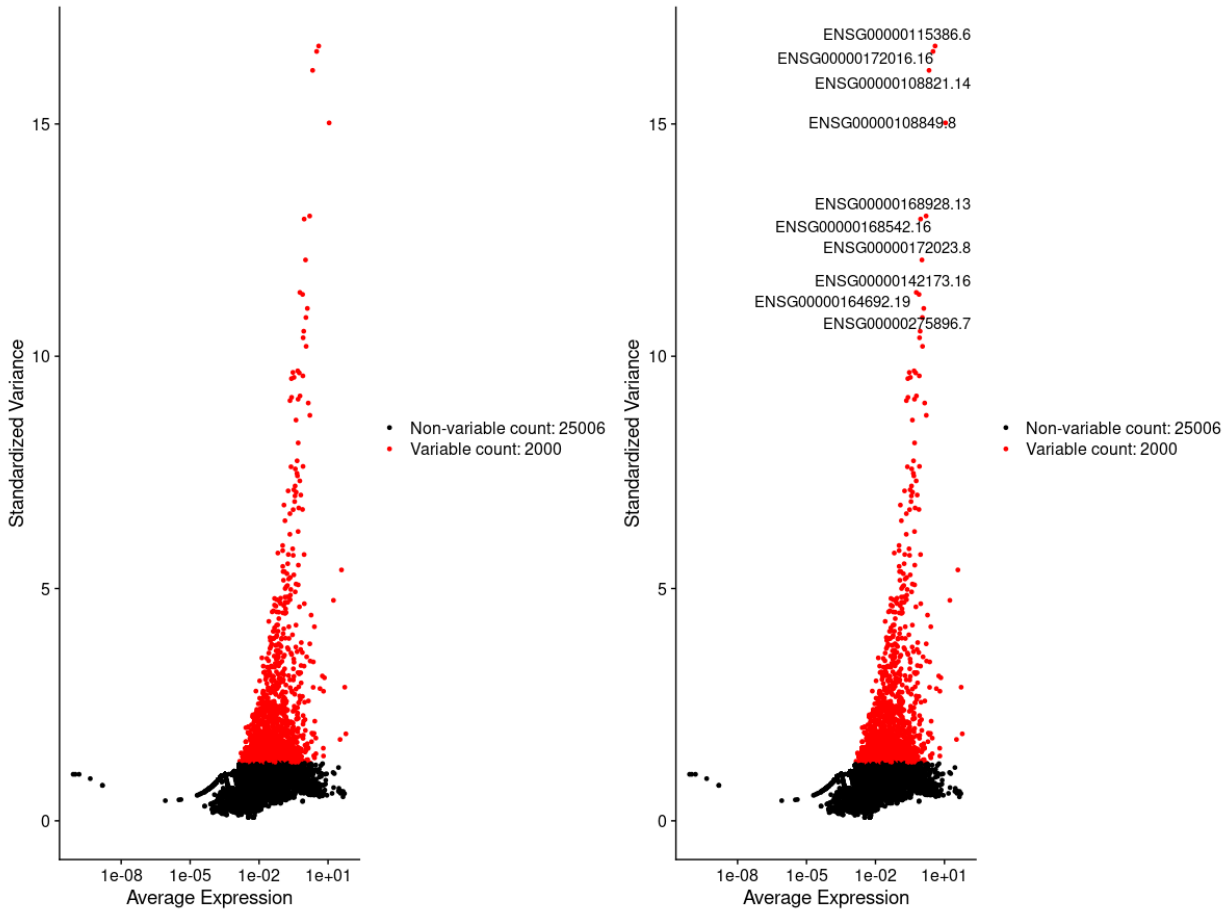

Figure 5: The scatter plot representing the relationship between the Average Expression and the Standardized Variance. The red colored points are the points of interest for the downstream analysis of variable count of 2,000. (Left) This scatter plot represents those without the top 10 most variable genes and (Right) shows the top 10 highly variable genes.

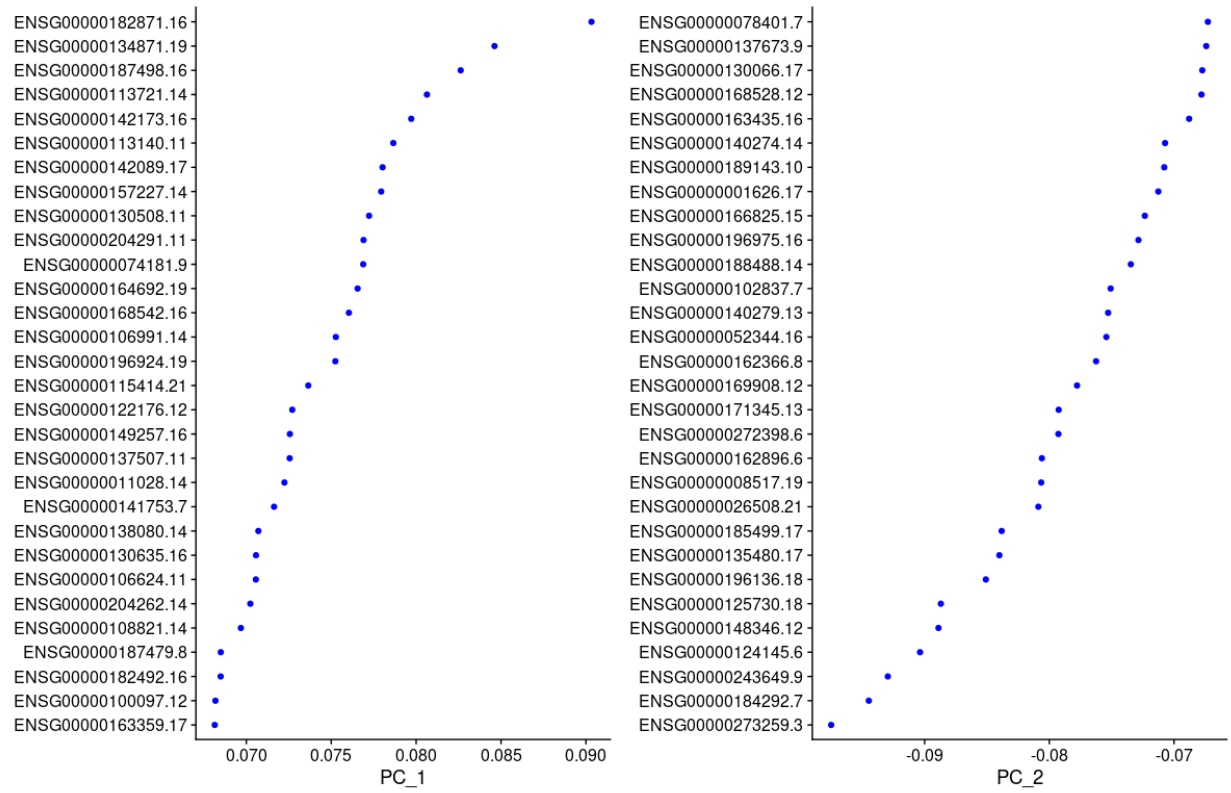

Figure 6: Scatter plot for gene subsets in PC1 and PC2. PC1 has positive gene groups whereas PC2 has negative gene groups.

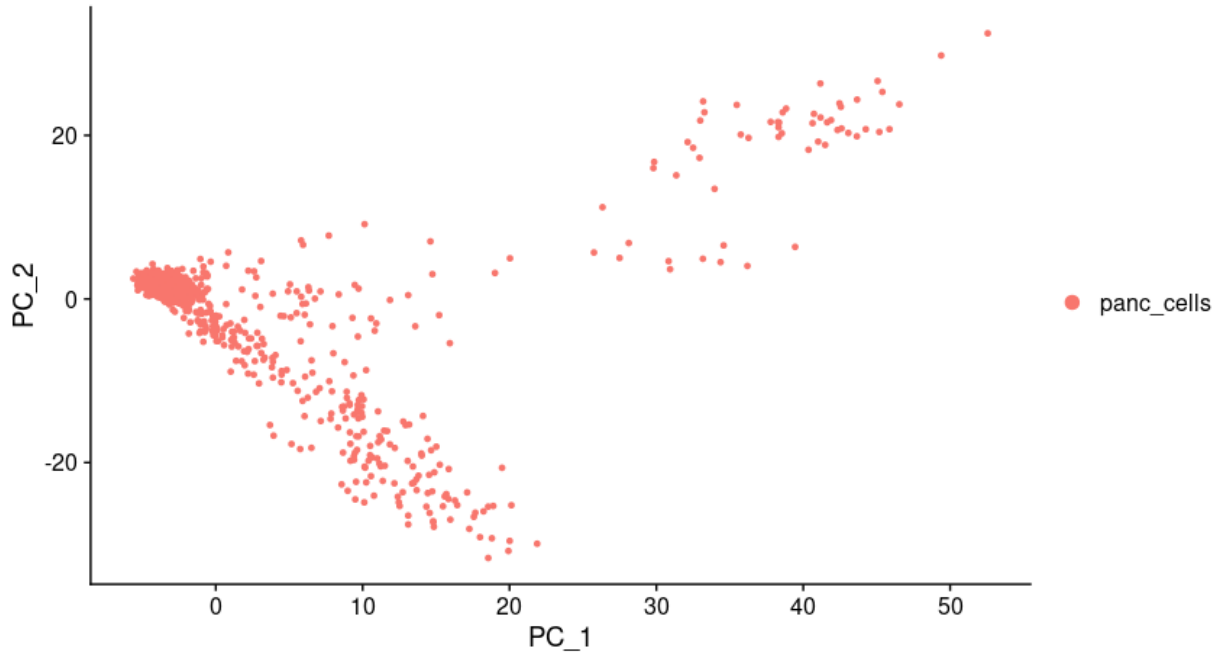

Figure 7: Scatter plot of PC1 and PC2

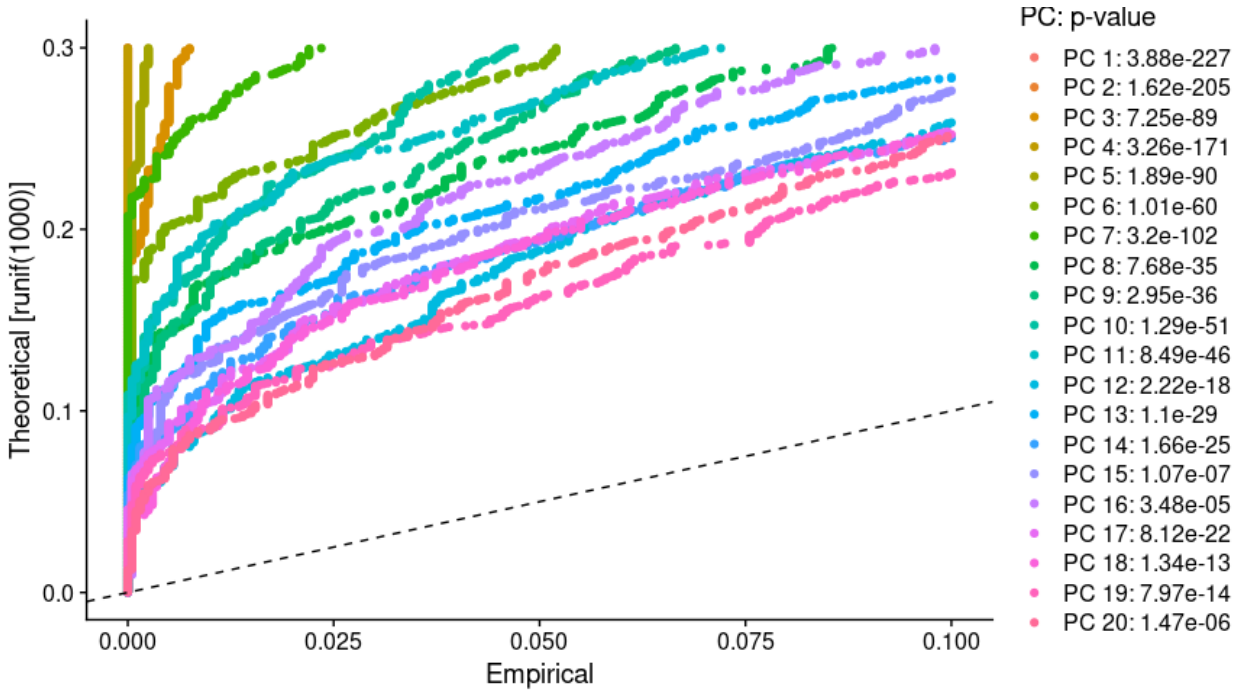

Figure 8: The Jackstraw plot for each PC. The plot displays the p value of each PC formed.

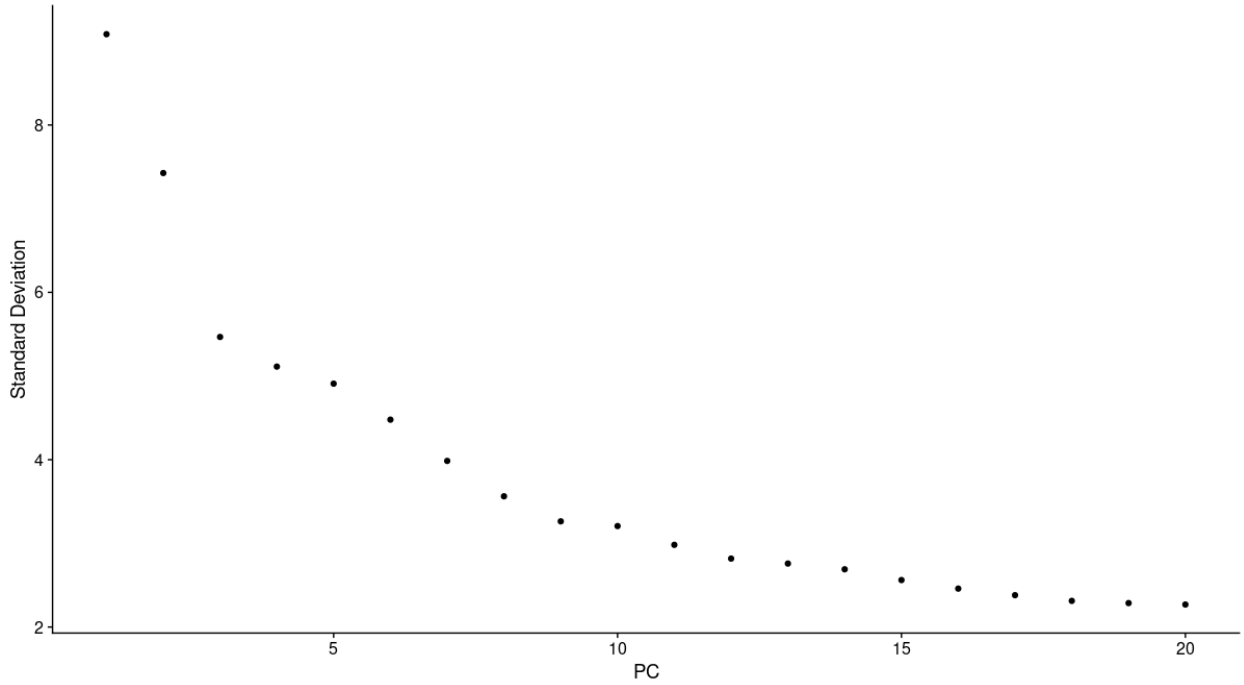

Figure 9: Elbow plot displays standard deviation(y axis) for each PC.

We noticed a considerable decrease in the standard deviation from principal components PC2 to PC3 in Figure 9. A steady decrease was observed hereafter for the rest of the PCs. Using RunUMAP, 9 clusters were identified as shown in Figure 10. A pie chart representing the relative proportions of cell number in each clusters is shown in FIgure 11.

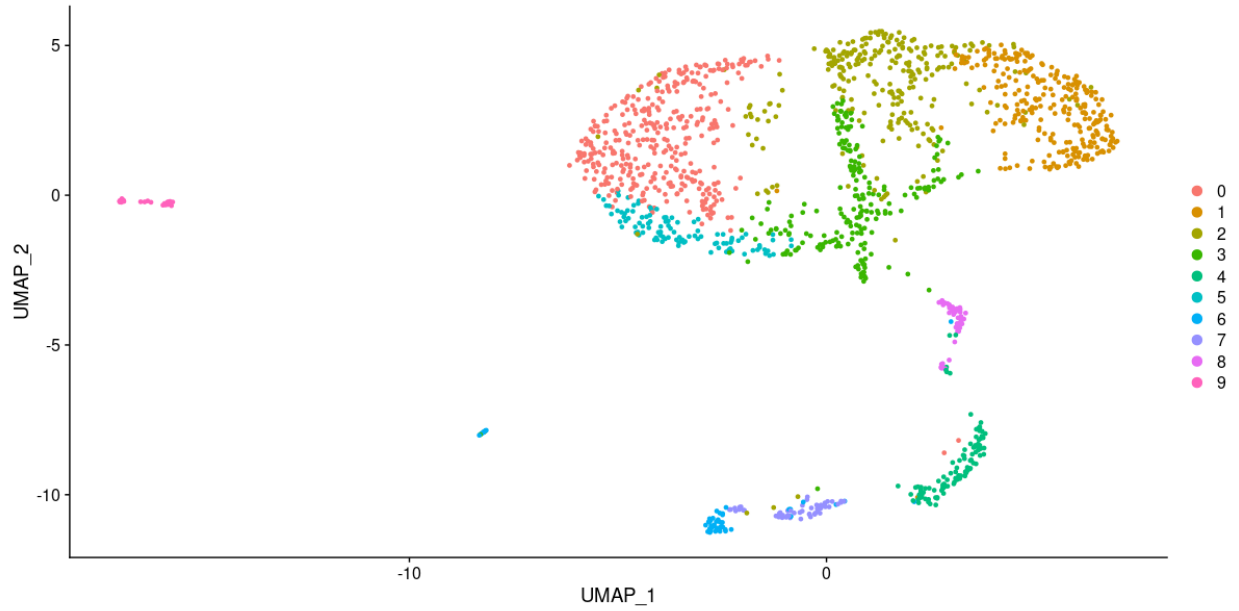

Figure 10: Clustering done using RunUMAP and plotted using DimPlot. 9 Clusters found.

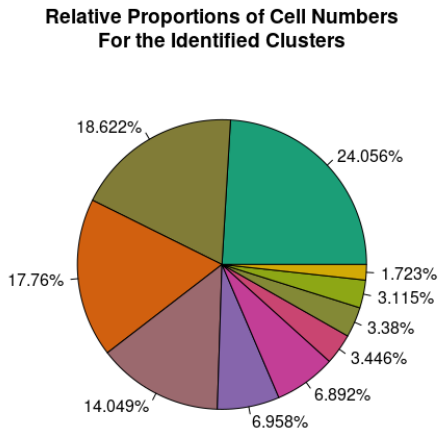

Figure 11: Pie chart showing the relative proportions of cell number in each clusters.

Through identification of marker genes and labeling of clusters we were able to confidently label cells to one of the following cell types: Alpha, Beta, Ductal, Delta, Acinar, Stellate,

Oligodendrocyte, Hepatocyte, and Macrophage. We then performed dimensionality reduction using UMAP in order to visualize our clusters, which can be seen in figure 12.

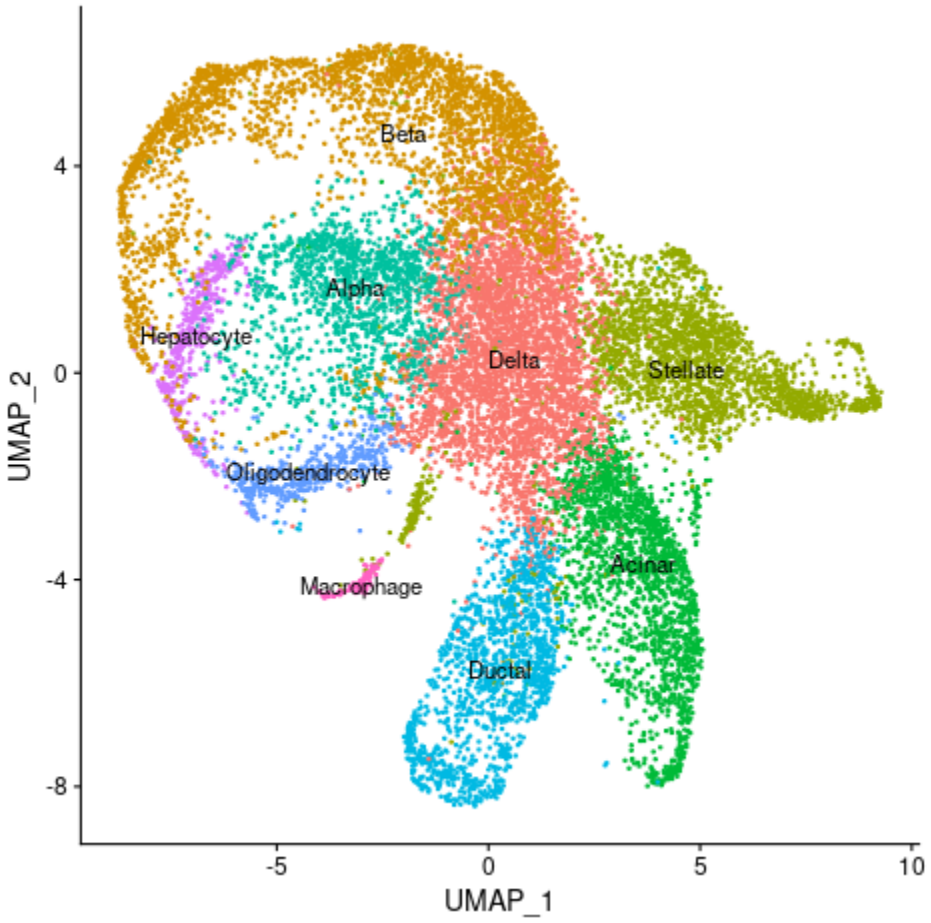

Figure 12 Umap of Clustered Data: This figure shows the output of UMAP on our clustered data with the labels according to cell type determined by marker genes. As opposed to Figure 1D in *Baron et al* the cells are not tightly clustered.

The marker genes used to identify these cell types included SST for Delta, INS for Beta, COL1A1, COL1A2, and FN1 for Stellate, Spink1 and CTRB2 for Acinar, GCG for Alpha, KRT19, MMP7, and KRT7 for Ductal, ACER3 and RPS6KA5 for Oligodendrocytes, GC for Hepatocytes, and IFI30 for Macrophage cells. A heatmap visualizing how these marker genes are expressed across each of the clusters can be seen in figure 13.

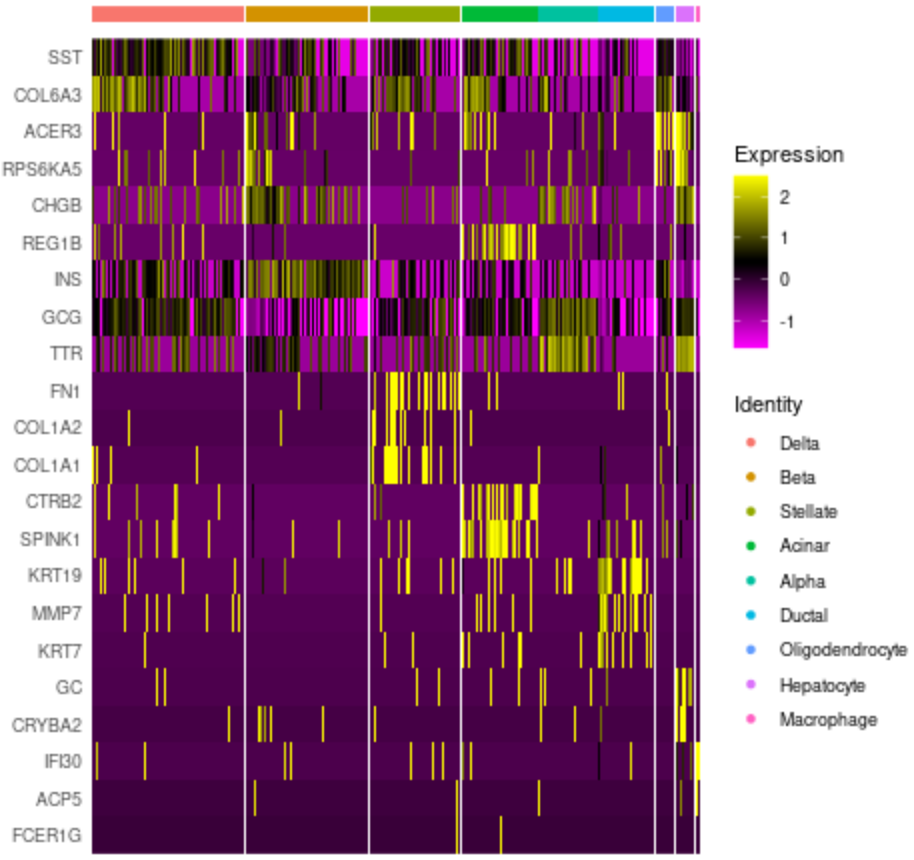

Figure 13: Heatmap of Marker Genes used to label cell clusters. From this we can see some of the weaker assignments like Delta, which used SST as its marker gene. Additionally, we can see the expression of some Novel Markers indicating potential different labels.

Additionally, we identified “novel” markers for some of the clusters that potentially indicate those clusters belong to a different cell type. TTR was found in the Alpha cluster, which is a marker for Hepatocyte cells. CHGB and CRYBA2 were found in the Hepatocyte cells cluster, which are markers for Enteroendocrine cells. COL6A3 was found in Delta cells and is a marker for Stellate cells. ACP5 was found in the Macrophage cells, and is a marker for Hofbauer and Stellate cells. These genes can also be seen in figure 13.

Discussion

Comparing our results to those of the researchers we were able to identify 7 of the 9 cell types that they identified. The two not included in our results were endothelial cells and gamma cells. Endothelial, or more specifically vascular endothelial cells, have many canonical cell markers, and while none of those appeared in the top 10 differentially expressed genes for any of our clusters it is likely that one of our clusters is mislabeled. Specifically, it is likely that the Oligodendrocyte cluster and Hepatocyte clusters are mislabeled, as these clusters did not appear in the results from the researchers. Gamma cells on the other hand, are uncommon in the pancreas, and are more difficult to confidently identify. While some of our clusters did have increased expression of one of the top gamma markers (PPY) this increased expression was insignificant, and is not uncommon in Beta and Alpha cells.

Additionally, we did not use many of the markers the researchers used to identify the common clusters across both results. For Alpha, Beta, Delta, and Ductal we were able to label these cell clusters using the same gene markers used by the researchers. For Stellate, Acinar, and Macrophage we had to identify different marker genes through the Human Protein Atlas and PangloaDB to label these clusters. It is unclear why we did not see the markers that researchers found in their data. One possible explanation is that we did not use all of the data used in Baron et al's original analysis. Lastly, the clustering of our data was not as clear as in the original paper even when using t-SNE instead of UMAP. Again, this could just be a product of the data we used to replicate these results.

Conclusion

From our single cell analysis we were able to confidently identify seven of the nine clusters that the researchers identified. While we did need to use different gene markers, we are still confident that our labels were correct. That being said, the clusters that were generated from dimensionality reduction through Seurat were not as clear as what was demonstrated in the original paper. It is unclear why this was the case, and could be because of the number of dimensions used by the researchers when they generated the plot.

References

1. Baron, Maayan, et al. “A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-Cell Population Structure.” *Cell Systems*, vol. 3, no. 4, Oct. 2016, pp. 346-360.e4. *www.cell.com*, https://doi.org/10.1016/j.cels.2016.08.011.
2. Johannes Rainer. Ensdb.hsapiens.v79: Ensembl based annotation package. 2017. R Package version 2.99.0.
3. Oscar Franzén, Li-Ming Gan, Johan L M Björkegren, *PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data*, Database, Volume 2019, 2019, baz046, doi:10.1093/database/baz046
4. https://www.proteinatlas.org/humanproteome/single+cell+type
5. Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng,Andrew Butler, Maddie Jane Lee, Aaron J Wilk, Charlotte Darby, Michael Zagar, et al.Integrated analysis of multimodal single-cell data.bioRxiv, 2020.