Single Cell RNA-Seq Analysis of Pancreatic Cells

Chris Lin, Yuehting Wang, and Cody Webb
TA: Marzie E. Rasekh

Introduction

The pancreas is an organ that has been gaining attention lately due to the rise of Type 2 Diabetes. Previous studies have classified some different cell types within the pancreas, but they have all struggled to work with the data in a high-throughput manner. In Baron et al[1]., they attempted to classify all the different cell types in the human pancreas via single cell RNA-seq analysis. The authors studied four human pancreases and two mouse pancreases using this technique in order to classify the previously unknown substructures of the human pancreas.

In our replication study, we took the sequencing data from the pancreas of the 51-year old female donor[2] in an attempt to replicate the results of the original study. We parsed through the data to obtain the relevant and well-behaved barcodes and UMIs from the original sequencing and extracted the gene expression levels for each of the cells. We then used K-nearest neighbor (KNN) to cluster our results from the samples. Using previously identified marker genes, we then labelled our clusters as a cell type and visualized them using UMAP. Finally, any novel marker genes which were discriminative of cell type that were not used in Baron et al were identified.

Data

In the original study, the authors used the inDrop platform to create their single cell data and then PCR to subsequently enrich it. We obtained the sequencing data for the 51-year old human woman. First, the number of reads were determined per distinct barcode. This resulted in an overwhelming majority of barcodes that were not usable (Fig 1A). In order to get the unique barcodes that were represented enough to replicate the study, all barcodes that were represented less than 80,000 times were removed from our sample. 80,000 was chosen as this threshold, as our data's average value was approximately 100,000, which was the average value for barcode count in the original study. It also resulted in a distribution of barcodes that was somewhat close to a normal distribution when plotted on a log scale (Fig 1B).

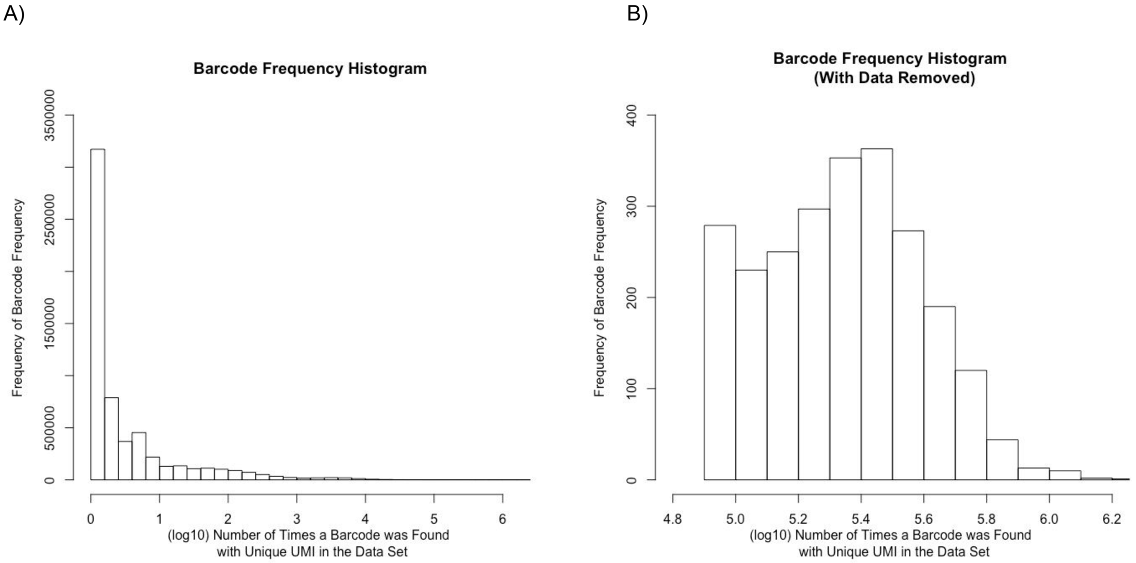


Figure 1. Histograms of the number of times a unique barcode and UMI were found in our samples on the x-axis, and the frequency with which we found these frequencies for the barcodes on the y-axis. Figure 1A represents all of the reads; Figure 1B represents all values with unique barcode frequency of more than 80,000. 1B represents around 0.04% of the data.

We inserted barcodes that had a frequency of greater than 80,000 into a whitelist that would result in useful information from our sample. This whitelist was then used with salmon alevin[3] along with the Human reference genome version 34[4] the gene expression data for each cell. Quality control metrics for the salmon run can be found in Table 1.

Table 1. Quality control statistics provided by salmon after the gene count matrix was constructed. It is somewhat worrying that the mapping rate is so low, but given that the RNA samples were collected post-mortem, it is to be expected that some RNA degradation would occur.

Quality Control Statistics							
Corrected Reads	Mapped Reads	Deduplicated Reads	Mapping Rate	Deduplication Rate	Mean By Max	# of Genes Expressed	# of Genes Expressed Over Mean
274911±174590	151952 ± 98463	4874 ± 2498	0.551 ±0.106	0.958 ±0.060	0.019±0.048	2285 ± 972	427 ± 221

Methods

Single-cell RNA-seq data quality control filtering

We converted the Ensembl gene identifiers to Gene Symbols and removed the duplicated genes in the UMI matrix. Then, we processed the UMI matrix and initialized the Seurat[5] object with the non-normalized data, keeping all genes expressed in >= 3 cells and keeping all cells with at least 200 detected genes. (Fig 2A) shows the visualization of the number of genes, the number of molecules counts, and the percentage of reads that map to the mitochondrial genome. Based on the distribution of the features (Fig 2A), cells that have unique gene counts (Features) over 4000 or less than 200 are excluded. We found the percentage of mitochondrial genes is high and that might be due to the specific cell type[6] so we chose to exclude cells with >10% of reads that mapped to the mitochondrial genome. After removing unwanted cells from the dataset, we normalized the gene expression measurements for each cell by the total expression, multiplied this by a scale factor 10,000, and log-transformed the result. Finally, we include only top 2000 high variance genes in our further analysis.

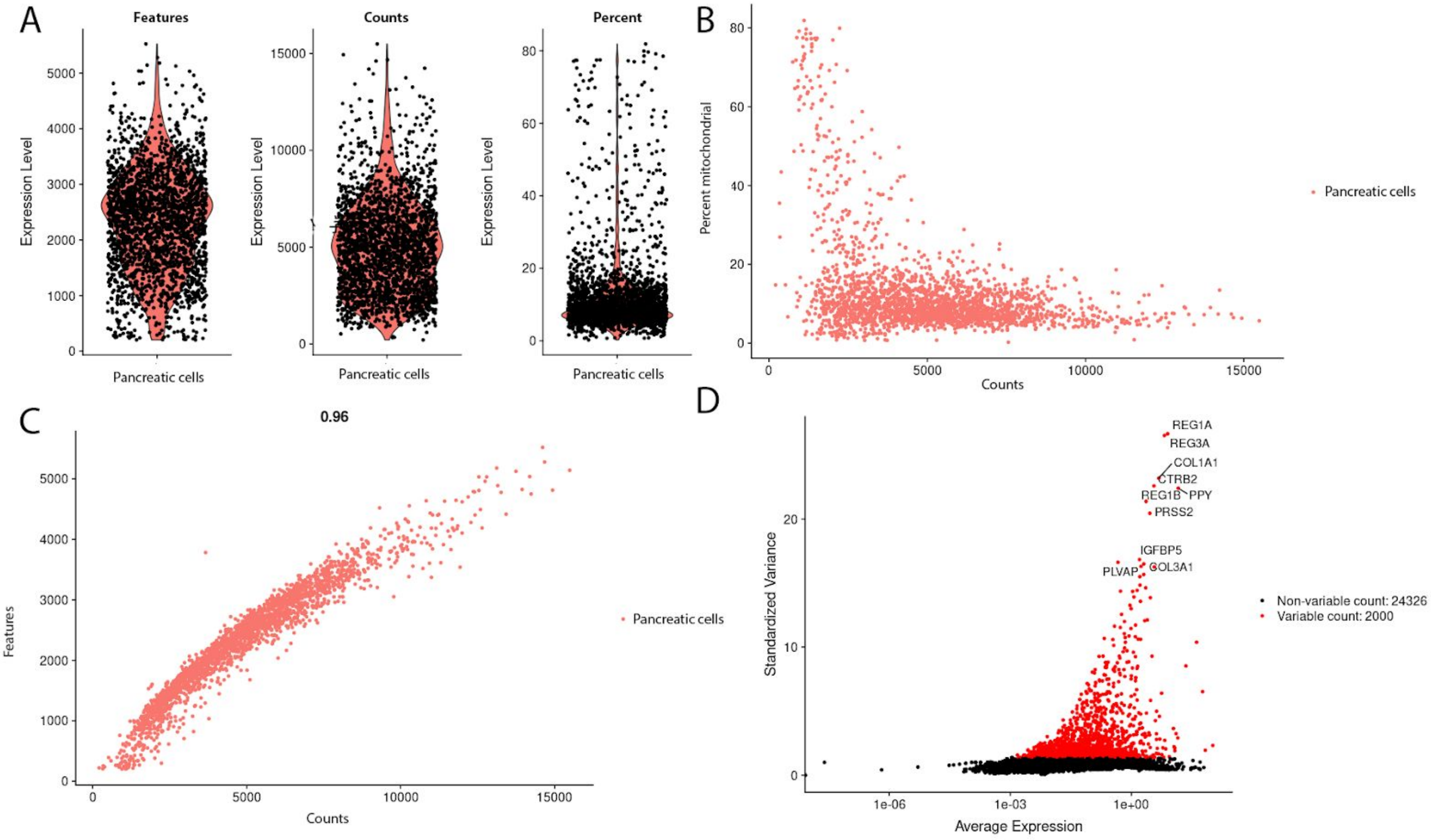


Figure 2. The quality control for the UMI metrix. (A) The violin plot for the number of genes, the number of molecules counts, and the percentage of reads that map to the mitochondrial genome. Most of the cells have the number of genes between 1000 to 4000. Most of the cells have a proportion of the mitochrial genes under 15%. (B) The scatter plot also shows the proportion of the mitochrial among the reads. (C) The scatter plot for gene-gene relationships. (D) The scatter plot shows the 10 most high variable genes.

Dimension reduction and clustering

Next, we scaled the expression of each gene and performed a principal components analysis (PCA) on the scaled data. Then, we used JackStraw procedure and an Elbow plot to determine the number of principal components for the dataset. The K-Nearest Neighbor technique was applied to cluster the cells. The resulting clusters were visualized with Uniform Manifold Approximation and Projection (UMAP) and the relative proportions of cell numbers were visualized with the bar chart.

Marker identification and cluster labeling

Following the processing and clustering of counts, novel marker genes that were discriminative of each cluster were identified. Possible candidates for marker genes were obtained by identifying differentially expressed genes of one cluster compared to all other clusters, imposing a pre-filtering threshold of at least 0.25 log-fold change and an adjusted p-value of < 0.05. This was then repeated for all ten clusters. Results were then sorted by log-fold change to identify strong candidates for marker genes. Using PanglaoDB[7], a database of canonical gene expression markers used to define cell types, in addition to the markers identified by Baron et al, each gene was assigned a cell type. Table 2 shows a subset of our table, highlighting the top three genes for each cluster with the highest average log-fold change.

Table 2. Top differentially expressed genes as determined by average log fold change and their corresponding cluster designations. Specificity is a measure of how frequently this gene is expressed in cells not of this specific cell type. For our clusters, high log-fold changes and low numerical specificity were seen, indicating suitable candidates for marker cells. Rows highlighted in red represent overlap between our discovered markers and cell types with that of Baron et al.

Gene	Cluster	Average log-fold change	Adjusted P-value	Cell type	Specificity	Cluster Designation
LOXL4	0	1.43	7.15E-114	Alpha cells	0	Alpha cells
CRYBA2	0	1.27	1.23E-148	Alpha cells	0.01	
AGT	0	1.22	3.77E-83	Adipocytes	0.02	
IAPP	1	3.37	1.55E-158	Beta cells	0.04	Beta cells
INS	1	3.34	3.45E-170	Beta cells	0.05	
DLK1	1	2.28	8.20E-163	Beta cells	0.04	

GCG	2	1.26	1.31E-80	Alpha cells	0.04	Alpha cells
TTR	2	1.03	6.00E-79	Alpha cells	0.07	
AC092155.1	2	0.77	5.31E-26	Alpha cells	0.01	
SPP1	3	2.51	9.20E-102	Cholangiocytes	0.05	Cholangiocytes
KRT19	3	2.27	2.98E-126	Cholangiocytes	0.15	
TACSTD2	3	2.23	5.27E-184	Cholangiocytes	0.09	
SST	4	3.4	1.12E-76	Delta cells	0.05	Delta cells
RBP4	4	2.72	3.79E-152	Delta cells	0.04	
RGS2	4	1.47	1.02E-43	Monocytes	NA	
PPY	5	3.91	1.70E-46	Gamma (PP) cells	0.04	Gamma cells
AQP3	5	1.96	3.77E-34	Gamma (PP) cells	0.07	
PEG10	5	1.05	3.73E-09	Gamma (PP) cells	0.07	
REG1A	6	4.49	9.98E-85	Acinar cells	0.03	Acinar cells
REG3A	6	4.33	9.33E-68	Acinar cells	0.01	
CTRB2	6	4.15	2.00E-139	Acinar cells	0	
COL1A1	7	4.59	8.13E-107	Fibroblasts	0.08	Hepatic stellate cells
COL3A1	7	4.07	4.18E-234	Hepatic stellate cells	0.1	
BGN	7	3.95	3.17E-107	Hepatic stellate cells	0.06	
TPSB2	8	3.9	3.94E-23	Mast cells	0	Mast cells
TPSAB1	8	3.86	8.93E-37	Mast cells	0	
ACP5	8	3.65	1.23E-24	Monocytes	0.02	
PLVAP	9	4.24	4.56E-234	Endothelial cells	0	Endothelial cells
ENG	9	3.09	4.99E-57	Endothelial cells	0.04	
PECAM1	9	3	1.25E-131	Embryonic stem cells	0.04	

It is important to note that many of the marker genes identified by Baron et al. did not appear in any of our clusters, and that multiple clusters sometimes shared a singular cell type. In these cases, we considered both the singular top differentially expressed gene for that specific cluster and its cell type label in PanglaoDB, as well as how frequently each cell type label appeared in the cluster to assign an overall cluster label. This resulted in the assignment of nine distinct cell types for ten initial clusters. Following cluster identification, a UMAP was generated colored by cell type, in addition to identifying any novel markers that we may have found. Lastly, we performed a stringDB enrichment analysis, identifying any interactions that existed within our cell type clusters.

Results

Single-cell RNA-seq data quality control filtering

The UMI matrix contains 2,357 cells and 60,240 genes with Ensembl gene identifiers. After we converted the Ensembl gene identifiers to Gene Symbols and removed the duplicated genes, there were 60,207 genes left. Then, we filtered out low quality cells based on the criteria (see Method). 1,815 cells remained after filtering out low quality cells. After removing unwanted cells from the dataset, we normalized the gene expression measurements for each cell by the total expression and filtered the counts matrix to include only the top 2,000 variable genes in our further analysis.

Dimension reduction and clustering

In order to cluster the cells, we performed principal components analysis (PCA) and visualized (Fig 3) the distribution of p-values for each Principal Component to determine its dimensionality. It appears that the p-values are significant in the first 15 PCs in the JackStrawPlot (Fig 3A). For the Elbow plot (Fig 3B), we observed an ‘elbow’ around PC 9-11, suggesting that the majority of the true signal is captured in the first 10 PCs. We decided to use the dimensionality of 10 for the clustering, resulting in 10 total clusters. Then, we performed a non-linear dimensional reduction (UMAP) to visualize the result (Fig 3C).

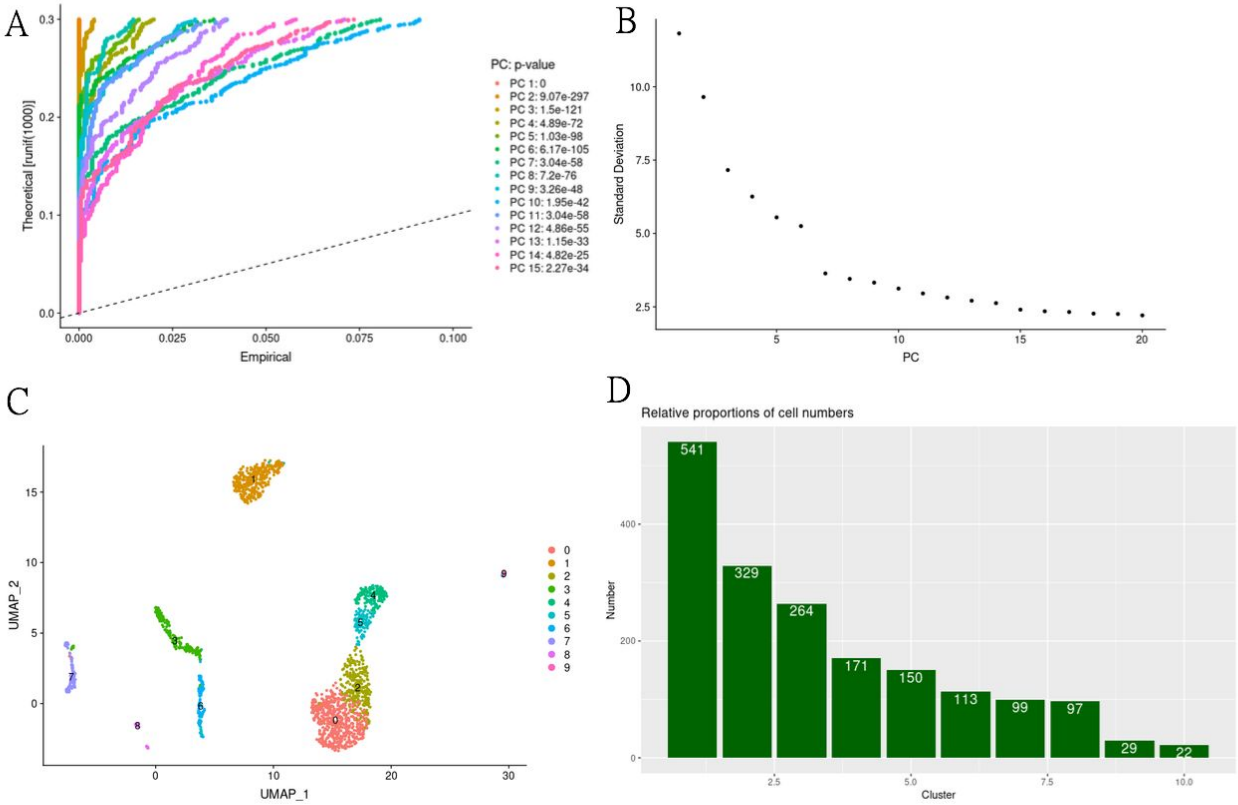


Figure 3. The dimension reduction clustering results. (A) The JackStraw plot of 15 PC with p-values. (B) The Elbow plot for 20 PC. It appears that the sharp drop-off stops around PC10. (C) The UMAP plot. There are 10 clusters for 1815 cells. (D)Bar plot of the relative proportions of cell numbers in each cluster. x-axis indicates the number of cells, y-axis indicates each cluster. Compared with other groups, cluster 1 has a larger proportion of cell numbers.

Marker identification and cluster labeling

To confirm that our differential expression-based method of finding marker genes was valid, a violin plot of the expression level (log TPM) of the top differentially expressed gene for each cell type cluster was created (Fig. 4) As seen from the figure, the top marker gene belonging to each of our assigned categories (indicated by an asterisk) shows a greater amount of overall expression than in unrelated clusters. Additionally, for some cell types such as mast, endothelial, alpha, and hepatic stellate cells, little to no expression is seen in the other clusters, indicating that these genes serve as good markers for their corresponding cell type, suggesting that our methods for identifying marker genes does indeed discriminate between various cell types.

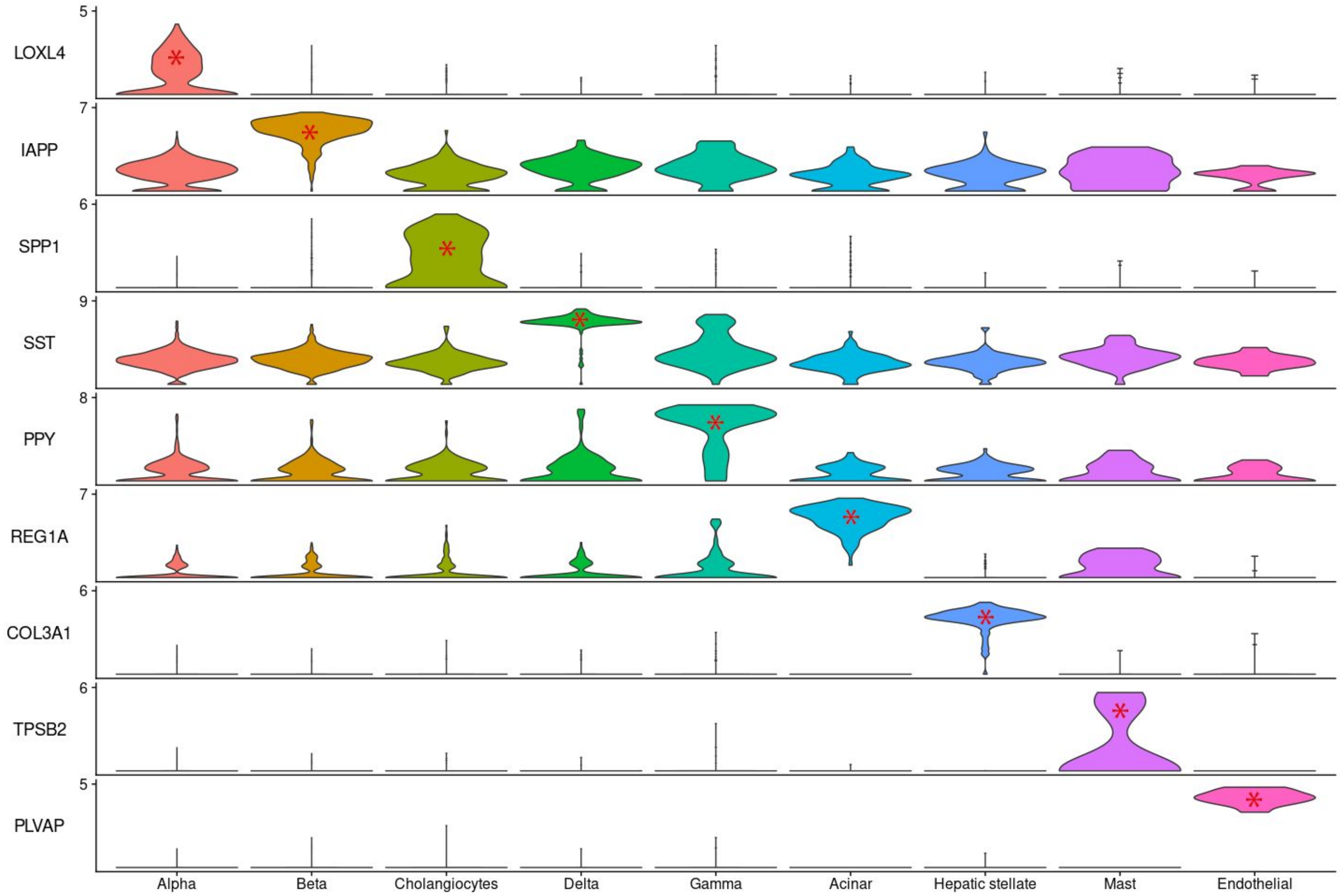


Figure 4. Violin plots of the log expression levels of marker genes for each cell type designation. Asterisks designate the corresponding marker gene with the highest log-fold change for that cell type. For each of the cell types, the top marker gene has both a higher median expression level in addition to a higher probability of having a high level of expression.

As the pancreas is a very heterogeneous organ composed of multiple cell types, we chose to use a UMAP[8]-based dimensionality reduction method to create a projection that preserves the overall structure and distances of the clusters (Fig. 5), unlike the t-SNE plot used by the original authors. Like the projection from Baron et al., we also found that our cell types formed clear clusters, indicating cell-to-cell differences at a molecular level. Namely, the beta, hepatic stellate, cholangiocytes/ductal, acinar, endothelial, and mast cells all formed distinct, isolated clusters that were very similar to that of Baron et al. Of the nine major clusters that were identified in the authors' projection, we were able to reproduce eight in our own UMAP. However, our delta, gamma, and alpha cells showed less separation, although this is difficult to say with certainty as the t-SNE used by the authors does not preserve cluster-to-cluster distances. Additionally, it seems like Baron et al. was able to achieve a higher resolution in their labeling and dimensionality reduction, successfully separating stellate cells into quiescent and activated, which we were not able to accomplish.

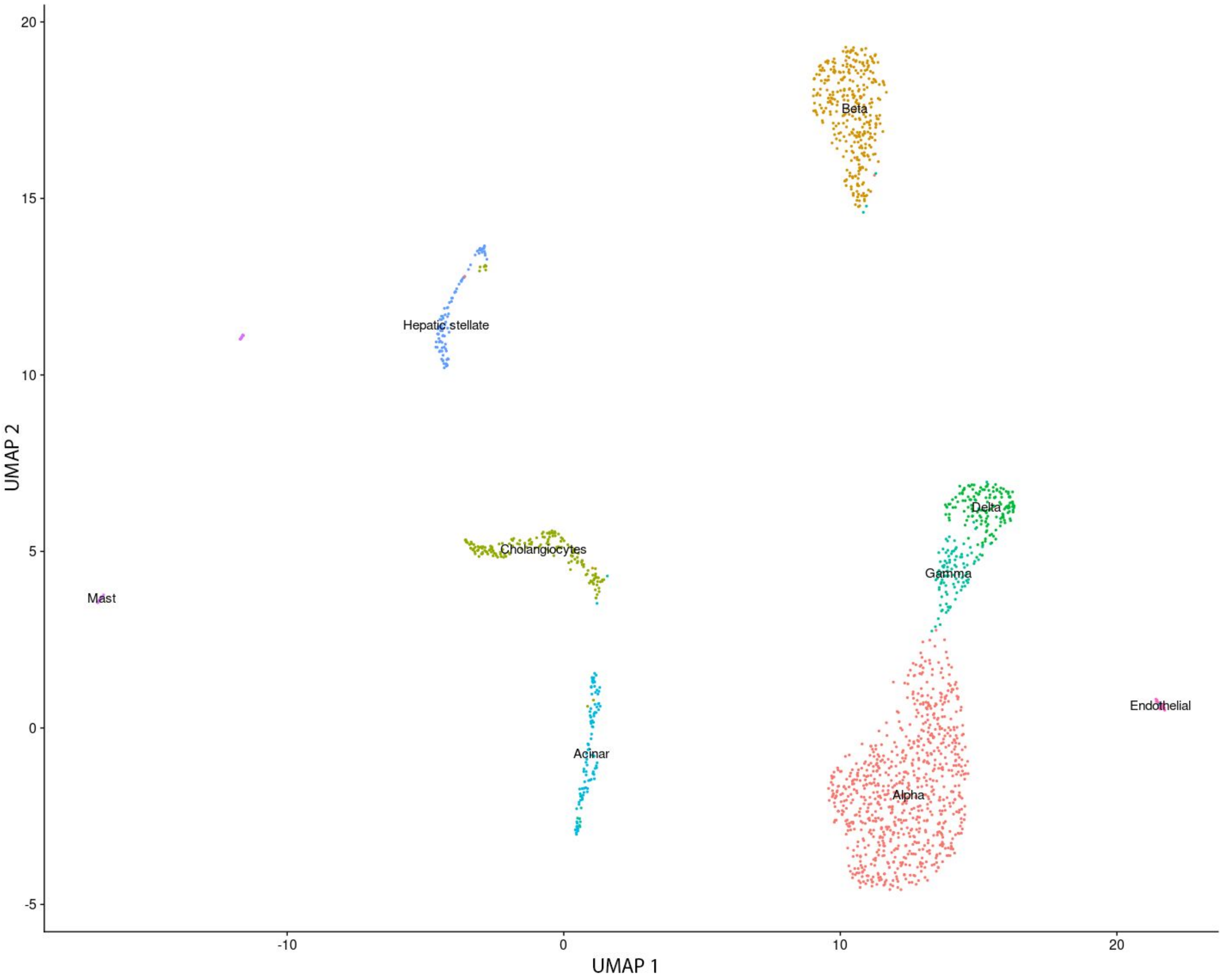


Figure 5. Scatterplot of clustered cells colored by cell type using non-linear dimensional reduction (UMAP). Clear separation of cell types can be seen as a result of dimensionality reduction and labeling.

Following labeling and projection, we further investigated the marker genes that we had identified for each cluster. The top five marker genes with the greatest average log-fold change for each cell type identified was visualized using a clustered heatmap of log normalized UMI counts across all cells (Fig 6). Compared to Baron et al., our heatmap shows similar trends as their results, with marker genes being highly expressed in the cell cluster type that they belong to. One finding of note is that alpha cells were originally separated into two separate clusters, as indicated by the high expression values in two separate sections of marker genes, which was not seen in Baron et al. Additionally, as we included five marker genes for each cell type, some overlap of high expression can be seen between cell types, especially in clusters lacking large numbers of cells, such as the mast and endothelial clusters, which is also observed in the original paper.

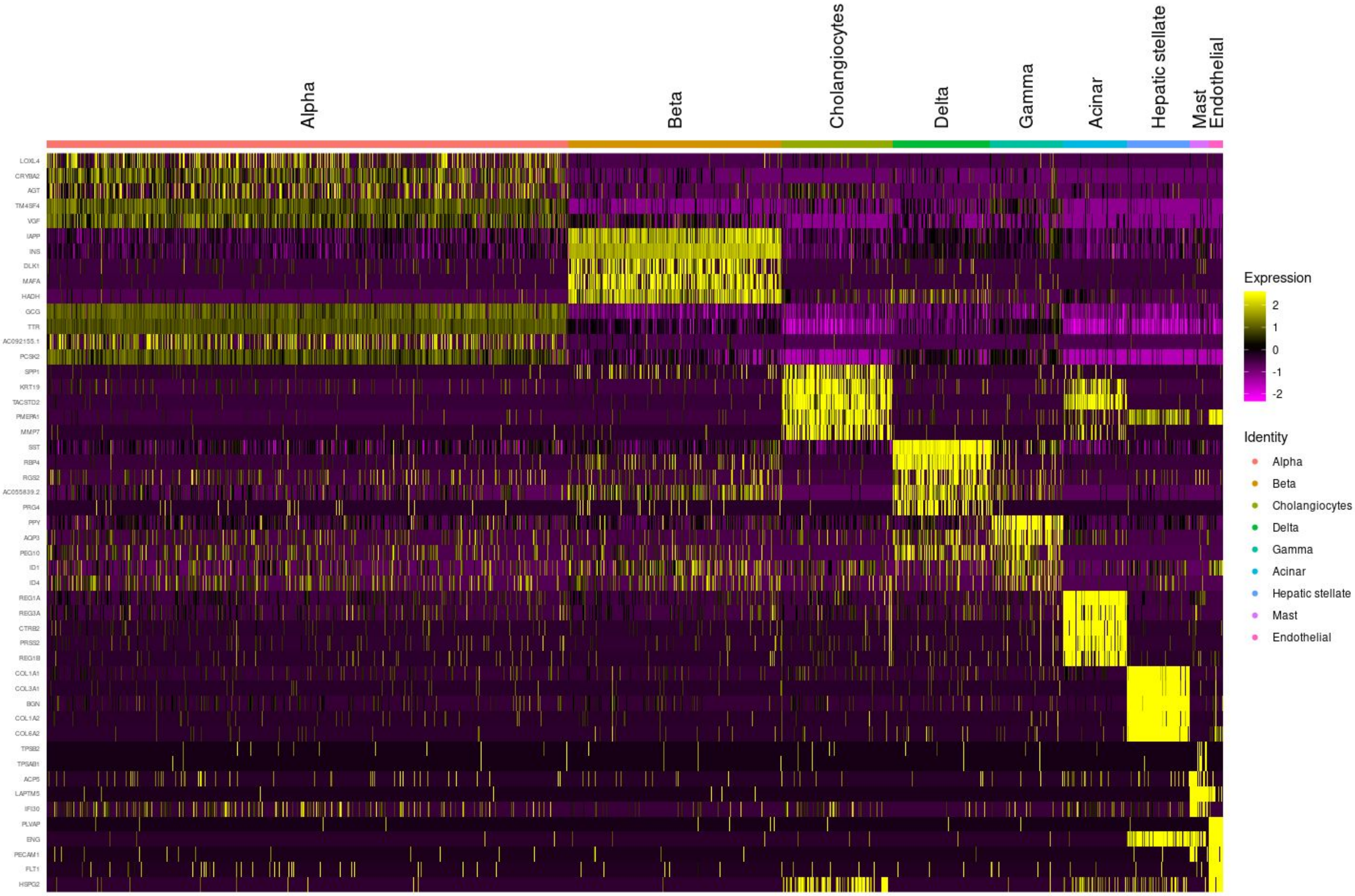


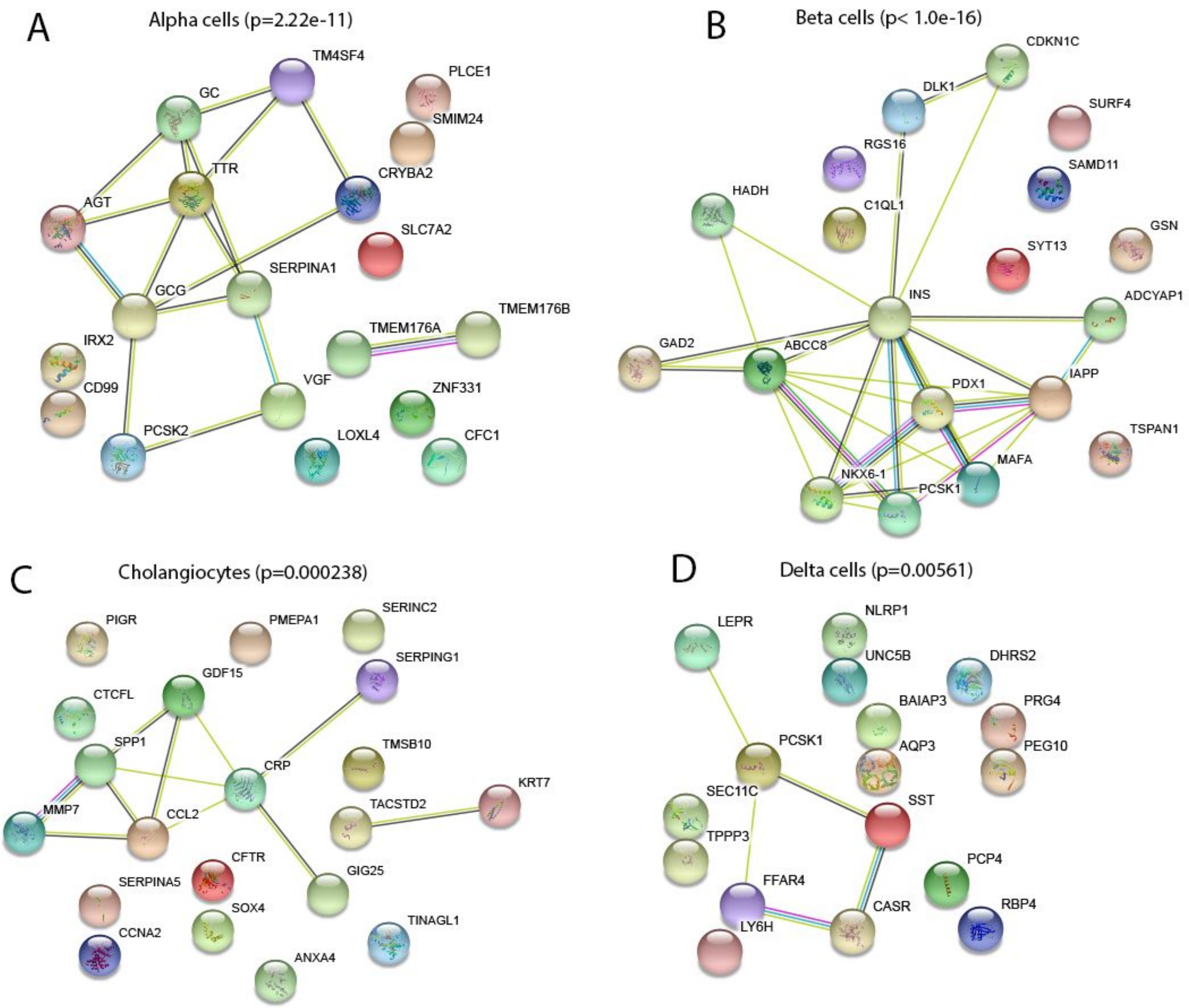
Figure 6. Heatmap of log normalized UMI counts across the top five marker genes for each cluster across all cells. High expression is indicated by yellow and low expression is indicated by magenta. Note the increased expression of marker genes associated with the cell type cluster.

These findings suggest that other discriminative markers of cell type that were not identified by Baron et al. may also exist. We have provided a small selection of novel cell markers in Table 3 that exhibit a log-fold change of greater than 0.75, a p-value less than 0.05, as well as a specificity (how frequently this gene is expressed in cells not of this specific cell type) of less than 10%. This specificity value reflects the entries from PanglaoDB, and acts to ensure that we are only selecting highly expressed genes that correspond to a single cell type.

Table 3. Novel cell markers for the nine distinct cell types that we identified. Each marker gene was selected in order to meet the criteria of log-fold change > 0.75, p-value < 0.05, and specificity < 0.1.

Cell type	Novel marker genes
Alpha cells	LOXL4, CRYBA2, AGT, TM4SF4, TTR
Beta cells	IAPP, DLK1, MAFA, HADH, ADCYAP1
Cholangiocytes	SPP1, TACSTD2, PMEPA1, MMP7, SERPING1
Delta cells	RBP4, AC055839.2, SEC11C, PEG10, NLRP1
Gamma (PP) cells	AQP3, PEG10
Acinar cells	REG1A, REG3A, CTRB2, PRSS2, REG1B
Hepatic stellate cells	COL1A1, COL3A1, BGN, CYGB
Mast cells	TPSB2, ACP5, CPA3
Endothelial cells	PLVAP, ENG, PECAM1, FLT1, HSPG2

To further confirm that our novel marker genes were indeed of a related cell cluster, we performed analysis using stringDB to identify if any functional networks existed within a list of our novel genes for each cell type (Fig 7). For all cell types that we identified in our previous analysis, each cluster had a protein-to-protein enrichment p-value of less than 0.01, with the exception of gamma cells. However, this may be due to our gamma cell genes only containing three entries that met our criteria for definition of a novel marker. Surprisingly, all other categories of cell type showed significant enrichment of the number of edges connecting each protein. This is especially noticeable in the acinar and hepatic stellate cell types, where the number of edge interactions is greater than double the number of nodes. Overall, this suggests that many more markers may exist for the cell types found in both our analysis as well as that of Baron et al.



A Alpha cells (p=2.22e-11)

B Beta cells (p< 1.0e-16)

C Cholangiocytes (p=0.000238)

D Delta cells (p=0.00561)

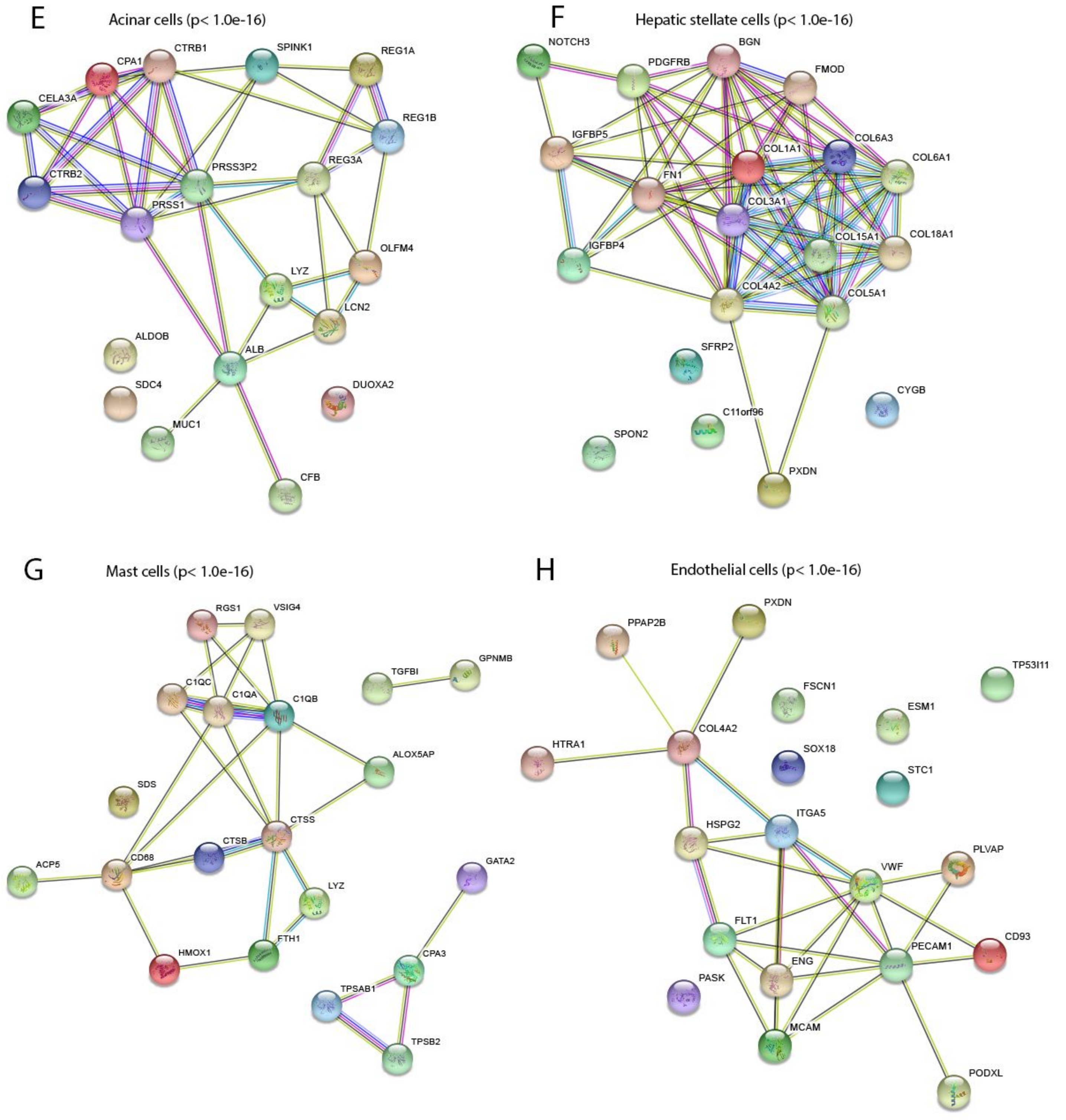


Figure 7. STRING protein-protein interaction analysis. (A-H) STRING PPI network connectivity of cell clusters identified in Table 3. All networks show (8 of 9) have an enrichment p-value of < 0.05, indicating significantly more interactions in all of our cell types. Confidence score threshold was set at 0.4 (medium) for all analyses.

Discussion

Using sample data from one of the four human donors from Baron et al., we were able to implement a similar single cell sequencing analysis pipeline, successfully identifying distinct clusters of cells, and subsequently assign individual cell type labels to those clusters. Overall, both the cluster projections and the cluster cell type identification were similar to those of the original authors. Compared to the 14 clusters identified by Baron et al., we were able to identify 10 distinct clusters of our own. From these 14 original clusters, nine distinct cell types were included in the original authors' t-SNE plots. In our own analysis, we were able to identify and assign distinct cell types to nine of 10 of our clusters. The cell types that we identified as present showed high overlap with that of Baron et al. The only differences that existed were the division of stellate cells in Baron et al into activated and quiescent cells, while we were only able to identify stellate cells in general, as well as the presence of mast cells in our UMAP.

One reason for this discrepancy in cluster number may be the fact that Baron et al. had access to four human samples as well as two mouse samples, while we only had access to one human donor. Any individual differences in gene expression from donor to donor may have an effect on our results, compounded by the fact that we only had one sample to work with. Furthermore, the cutoff we selected to filter out any low quality cells was only based on the distribution in Fig 2. Having different cutoff points will likely result in marginally different results, as this filter ultimately decides which cells end up in the analyses.

Additionally, the authors developed their own iterative hierarchical clustering method to identify clusters, and were able to conduct a thorough literature review to identify marker cells. As we lacked the specialty knowledge pertaining to the pancreas, we chose to reference a database containing canonical cell types when our clusters did not contain markers corresponding to those of the original authors. As a result, it is possible that some genes were mislabeled as the wrong marker cell type. Since the authors only provided one marker gene per cell type, it was exceedingly difficult to confirm whether our marker gene identification for ambiguous clusters were correct. Regardless of these difficulties, our end results did still show strong agreement with Baron et al. in regards to our final scatter plot projection, as well as the cell types identified, supporting

Conclusion

Given the lack of a complete dataset, combined with the lack of homogeneity in the handling and processing of single cell RNA-seq data due to a lack of established protocols, we can not say with complete certainty that we were able to reproduce the findings of Baron et al. Many of the steps taken in both the original Baron et al. study as well as our own were likely suboptimal due to how new the field is. Through collaboration and cooperation, we were able to use the tools given to us to produce very high quality results that did indeed seem to match most of the findings of the original authors, corroborating the fact that both well known and novel cells exist in the pancreas. We hope that the techniques that we used can help ourselves and others in the field as single-cell RNA-sequencing becomes more standardized in the coming years.

References

[1] Baron, Maayan, Adrian Veres, Samuel L. Wolock, Aubrey L. Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, et al. 2016. “A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-Cell Population Structure.” Cell Systems 3 (4): 346–60.e4.
[2] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2230758
[3] Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods.
[4] Harrow, Jennifer et al. “GENCODE: the reference human genome annotation for The ENCODE Project.” *Genome research* vol. 22,9 (2012): 1760-74. doi:10.1101/gr.135350.111
[5] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, III WMM, Hao Y, Stoeckius M, Smibert P, Satija R (2019). “Comprehensive Integration of Single-Cell Data.” Cell, 177, 1888-1902.
[6] Lori L Bonnycastle, Derek E Gildea, Tingfen Yan, Narisu Narisu, Amy J Swift, Tyra G Wolfsberg, Michael R Erdos, Francis S Collins, Single-cell transcriptomics from human pancreatic islets: sample preparation matters, *Biology Methods and Protocols*, Volume 4, Issue 1, 2019, bpz019, https://doi.org/10.1093/biomethods/bpz019
[7] Oscar Franzén, Li-Ming Gan, Johan L M Björkegren, PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data, Database, Volume 2019, 2019, baz046, doi:10.1093/database/baz046
[8] McInnes, L, Healy, J, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426, 2018.