Single Cell RNA-Seq Analysis of Pancreatic Cells

Group: Hedgehog
Teaching Assistant: Joey Orofino

Data Curator: Rojashree Jayakumar
Programmer: Qinrui Wu
Analyst: Merai Dandouch
Biologist: Dylan Beeber

Introduction

Single cell RNA (scRNA) sequencing, unlike traditional sequencing methods, can be utilized to obtain transcriptomic information of low population of cells and reveal information about cell population heterogeneity (Tang et al., 2019). Additionally, single cell RNA sequencing is harnessed to identify differential gene expression, epigenetic alterations and cell specific marker identification (Zhang et al., 2021). Single cell sequencing has several clinical applications in different fields and helps get insights into the disease mechanisms and provides a basis for treatment (Tang et al., 2019).. In the field of oncology single cell sequencing has profound applications. Tumor development is a complex process, where somatic cells accumulate several mutations. Therefore cell heterogeneity is the main driving force in tumor and scRNA sequencing can efficiently detect cell heterogeneity, by measuring the transcription pattern in each cell and distinguishing the cell types in cancer (Zhang et al., 2021).

Baron et al., performed a droplet-based (inDrop), scRNA sequencing method to find the transcriptomes of 12,000 pancreatic cells from two mouse strains and 4 human donors. In the inDrop method, mRNA from lysed cells are encapsulated with droplets with hydrogel beads containing the barcodes. Barcoding of mRNA is done by reverse transcription (Zilionis R. et al., 2016). Baron et al clustered cells into 15 types and detected subpopulations of ductal cells with different expression profiles. Further they were able to detect disease associated differential expression patterns and heterogeneity in the regulation of genes in B-cells in the human pancreas with respect to functional maturation and levels of ER stress.

In our study, we tried to replicate the results by Baron et al., using the sequencing data from a 51 year old female donor. In our study, we processed the barcode reads and generated the UMI counts matrix, performed quality control of the UMI matrix and analyzed the matrix to identify cell clustering patterns and marker genes for each cluster and determine the biological meaning by performing gene set enrichment analysis on the marker genes.

Data

The scRNA sequencing protocol used the inDrops method and paired end sequencing was performed on Illumina Hiseq 2500 machine. The number of PCR cycles for library preparation was 10-12 cycles. Reads were trimmed using Trimmomatic and read 2 was aligned to reference transcriptome using bowtie (Baron et al).

The SRA run selector from accession number GSE84133 for the data was used to identify the three sequencing runs corresponding to the 51 year old non-diabetic female donor (Sample Name: GSM2230758) which contained a total of 131.5G bases. The runs were SRR3879604, SRR3879605 and SRR3879606 and the fastq files containing the mRNA reads corresponding to the runs were used for further analysis. The cells that were sequenced for this sample were the human pancreatic islets. Read 1 was preprocessed so that each sequenced read

will have a barcode with 19 barcode bases and 6 UMIs. The experimental information for this sample can be accessed through the experiment id SRX1935939.

The fastq files from the three runs containing read 1 were combined into a single file to contain all the reads in a single fastq file using the zcat command. The awk command was used to find the barcodes and split them to include only the first 19 unique bases. These bases were sorted using sort command and the frequency of each unique barcode was obtained using uniq -c command. The unique barcodes with frequencies were plotted in R to obtain the inflection point.

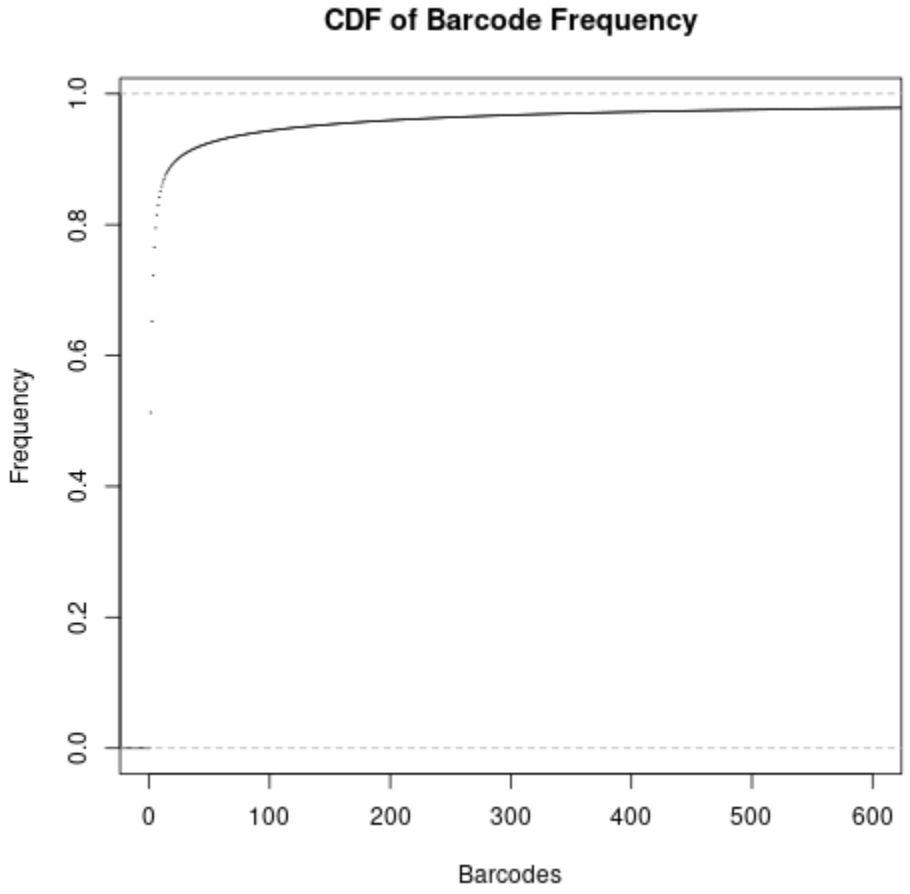


Figure 1: Cumulative distribution frequency of the unique barcodes. The inflection point is not clear from the graph and hence two different thresholds were used to whitelist the barcodes for further analysis.

Since the inflection point was not clearly defined (Figure 1), two different thresholds were used to whitelist the barcodes. One of the thresholds removed the first 100 reads (corresponding to frequency less than 20) were removed. The other threshold was stricter and removed cells with frequency less than 100. Both the resultant barcodes were used to generate the UMI matrix.

Index using the salmon index command was built with both decoys and without decoys. To build a decoy aware transcriptome by using the entire human transcriptome, the fasta files containing the transcript sequences and the genome sequences (GRCh38.p13) corresponding to the latest gencode version (Release 40) were downloaded using wget. The genome targets were extracted using grep command to construct the decoy file. The decoy file and the concatenated transcript and genome files were used to build the index.

To build the UMI matrix, the salmon alevin command was used. The command included the ISR (inward, stranded, reverse strand) library, along with a mapper file containing the mapping of the Ensembl transcript ids to the Ensembl gene ids, the paired end reads fastq files of all three runs, the barcode and UMI fastq file containing the reads and the whitelisted barcode files. Three different matrices were constructed using different indexes to check for differences in downstream analysis and mapping statistics. The first matrix was constructed using a threshold that excluded the first 100 low frequency barcodes and index files constructed without decoy aware transcriptome. The second matrix used index files constructed using decoys aware transcriptome and the same threshold as the first matrix. Since removing the first 100 low frequency barcodes did not change the dimensions of the data as much as expected, a third matrix where barcodes with less than 100 reads were removed and the matrix was constructed using the index built using decoys aware transcriptome. Overall the mapping rate was similar for all the three methods, however when a decoy aware transcriptome was used, many fragments were discarded as they mapped to decoys (Table 1). This is because usually the decoy transcript has regions with high mismappings like pseudogenes and therefore incorrect mapping is avoided when decoy-aware transcriptomes are used. Downstream analysis using the Suerat pipeline yielded the same results from all three matrices.

Mapping Statistics	Matrix without decoy aware transcriptome and barcodes with frequencies greater than 20	Matrix built using decoy aware transcriptome with frequencies greater than 20	Matrix built using decoy aware transcriptome and barcodes with frequency above 100
Number of targets in index	245,261	245,455	245,455
Number of decoys	0	194	194
Rich equivalence classes	245,261	233,703	230,710
Total reads in the equivalence classes	582,176,066	576,608,885	573,480,191
Number of fragments discarded because they are best-mapped to decoys	0	16,196,418	16,070,982
Reads with `N` in the UMI sequence	73657	73657	71281
Mapping rate	**43.9432%**	**43.523%**	**43.2868%**

Table 1: Differences in few mapping statistics of the three UMI matrices

Methods

Quality Control of The UMI Counts Matrix

The tximport and the fishpond R packages were used to import the UMI counts matrix in the form of salmon alevin output. Instead of following the custom analysis methodology developed by Baron et al's paper, we used the Seurat R package to store and manipulate our single-cell data into a Seurat object. In order to select filtering criteria we visualized QC metrics as violin plot (Figure 2) and feature scatter plot after creating Seurat object. As figure 2A showed, a majority of features are located in the range 200-3000. Therefore, we filtered cells that have unique feature counts over 2,500 or less than 200 as well as have less than 5% mitochondrial count to get significant results.

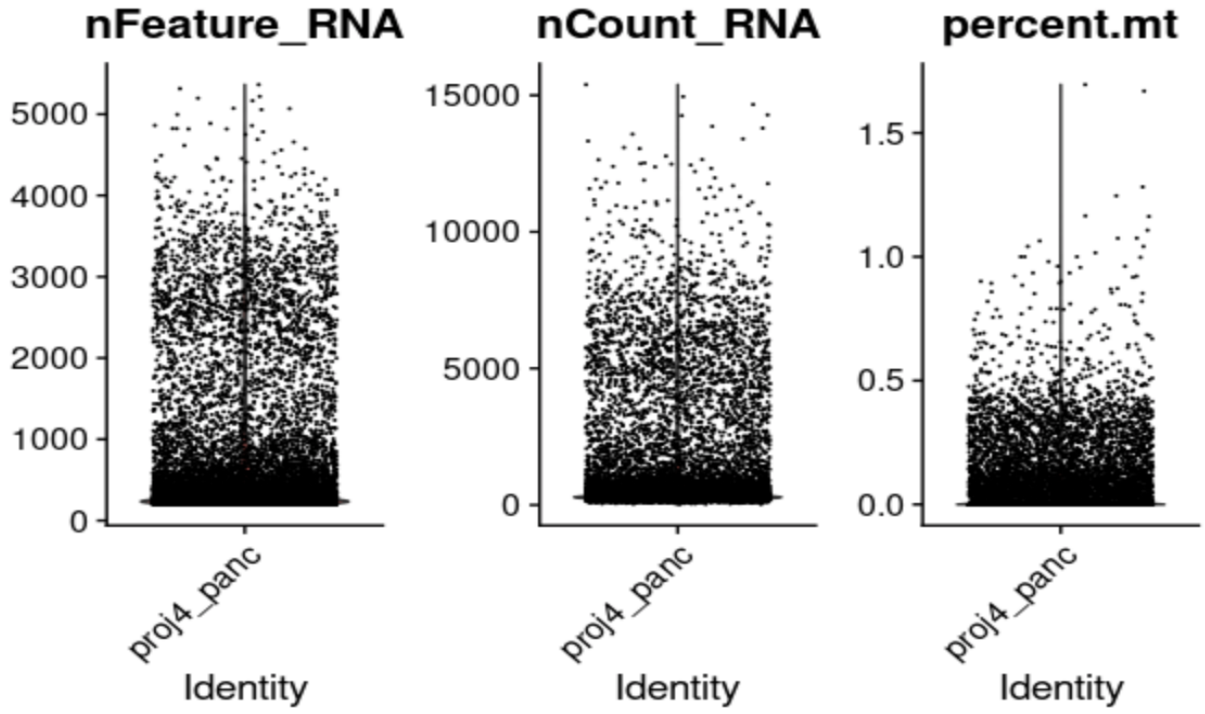


Figure 2: Visualize QC metrics for original UMI matrix. Violin plots of Number of genes detected in each cell (nFeature_RNA), Number of molecules detected within a cell (nCount_RNA), and Mitochondrial percentage (percent.mr)

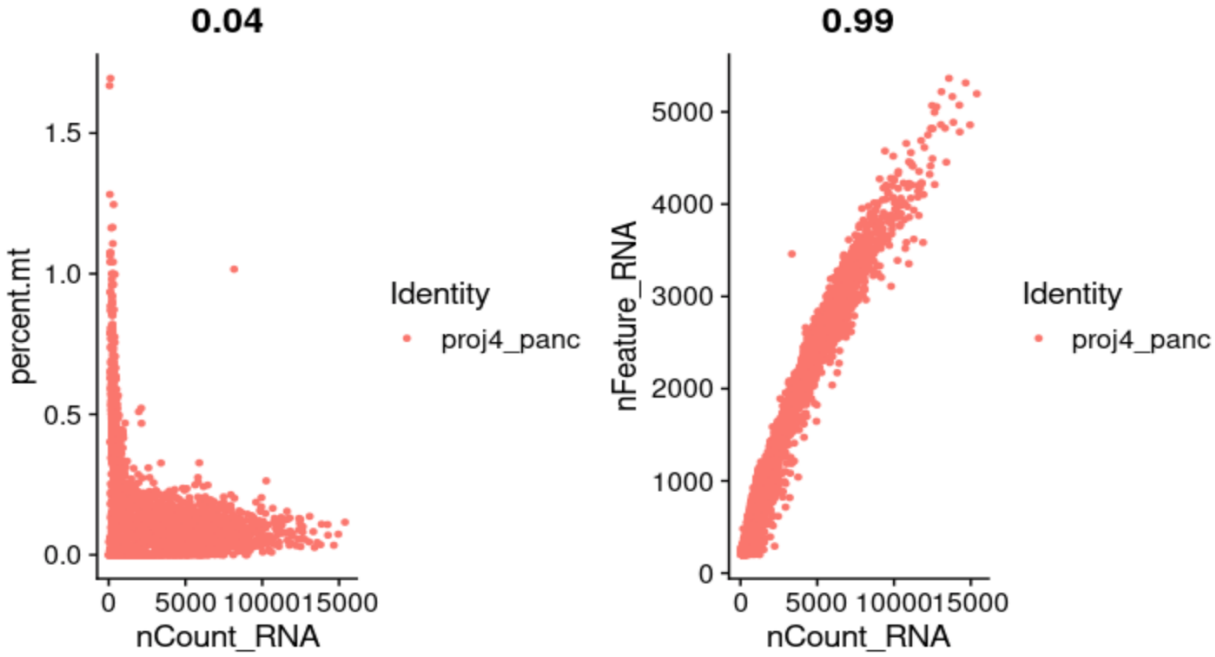


Figure 3: Feature scatter plot of nCount_RNA vs. percent_mt and nCount_RNA vs. nFeature_RNA for original UMI matrix. Based on the value from figure 2.

Data Normalization and the detection of highly variable features

Once having removed unwanted cells from the dataset, the data was normalized by the 'LogNormalize' method to conduct the feature expression measurements for each cell by the total expression with scale factor parameter of 1000. Identification of highly variable features was performed using downstream analysis that helps to highlight biological signals in single-cell datasets (Figure 4). Table 2 indicates the top 10 most highly variable genes.

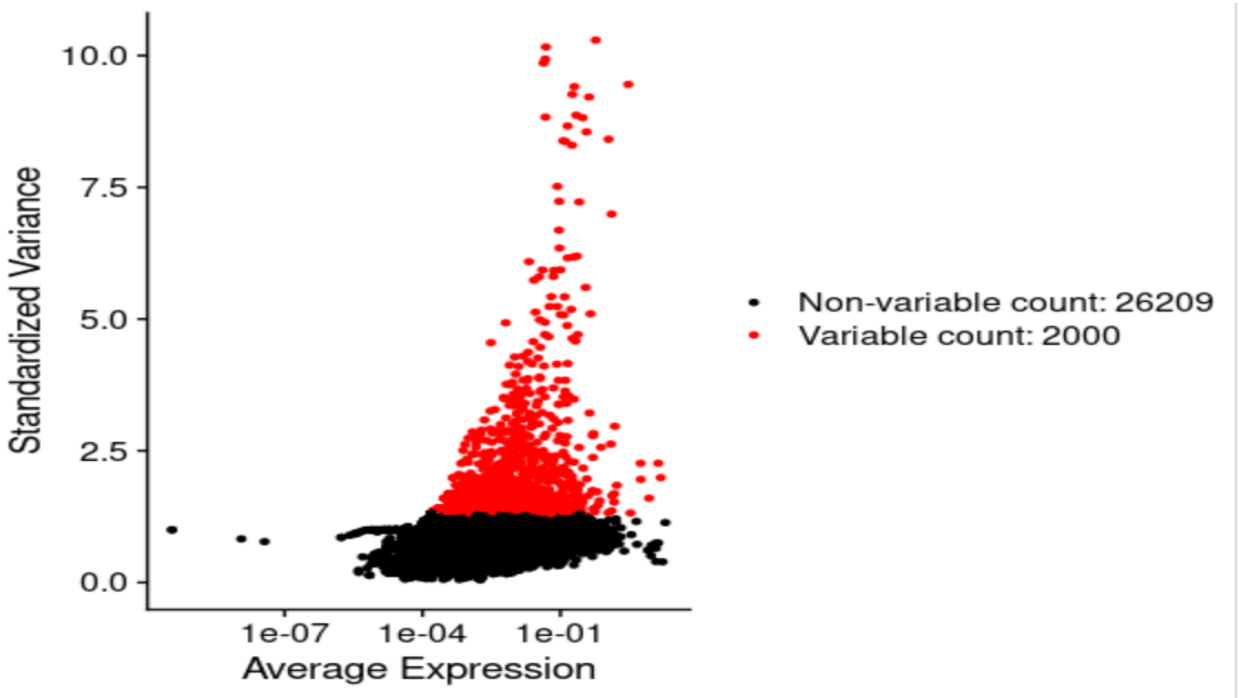


Figure 4: Scatter plot shows the relationship between average expression and standard variance.

No.	1	2	3	4	5	6	7	8	9	10
Genes	COL1A1	PLVAP	TPSB2	TPSAB1	PPY	COL3A1	IGFBP5	CTRB2	CTRB1	ALB

Table 2: The 10 most highly variable genes.

Perform linear dimensional reduction

After two filtrations were applied, an elbow plot was generated to determine the number of principal components. Figure 5 suggests that 'elbow' around PC10-12, that is, the majority of true signal is captured in the first 10 PCs. Based on several non-linear dimensional reduction techniques, UMAP was selected to visualize and explore data as clusters. There are a total 12 clusters represented from figure 6.

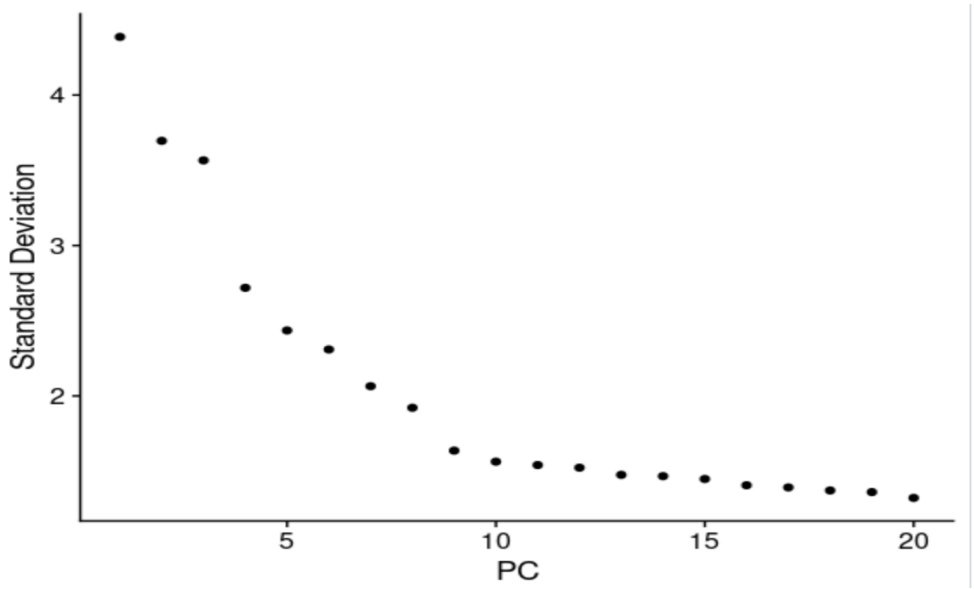


Figure 5: Elbow plot for principal component analysis.

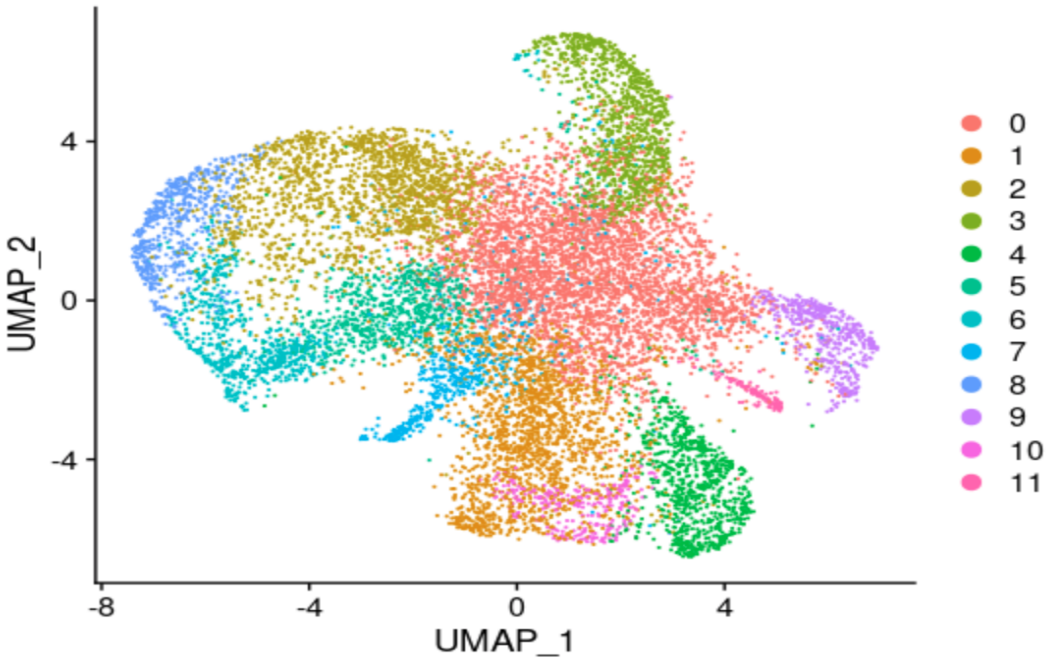


Figure 6: Umap to visualize 12 identified clusters.

In order to investigate gene expression signatures related to pancreas cell types, differentially expressed genes were identified using the Seurat package. Specifically, a wilcoxon test was used to test against all combinations of two UMI gene count pair groupings. Subsequently, the top 10 genes were extracted using average log2FC and filtered using p-values < (10^-5) . These genes were used to differentiate cell cluster types using literature from Baron et al., 2016 and other sources.

Results

Table 3 reports the number of genes and cells in the unfiltered dataset and dataset after applying two filtrations, respectively. The original UMI matrix had 28,209 genes and 14,177 cells. After filtered out low quality cells (nFeature_RNA > 200 & nFeature_RNA < 2500 & percent.mt < 5), there were 26,173 genes and 12,968 cells. Interestingly, the same number of genes and cells remained after removing low variance genes. It may be due to all cells being highly expressed after we filtered out low-quality cells.

	Number of Genes	**Number of Cells**
Original UMI Matrix	28209	14177
Filtering out low-quality cells	26173	12968
Filtering out variance genes	26173	12968

Table 3: Summary of number of genes and cells remaining after two filter steps.

Figure 7 implied relative proportions of cell numbers for 12 clusters. Based on the pie chart, cluster 0 had the highest number of cells with a percentage of 30.59 , about a third of the total number. Cluster 13 had the lowest proportion compared to another cluster, which was 0.93%.

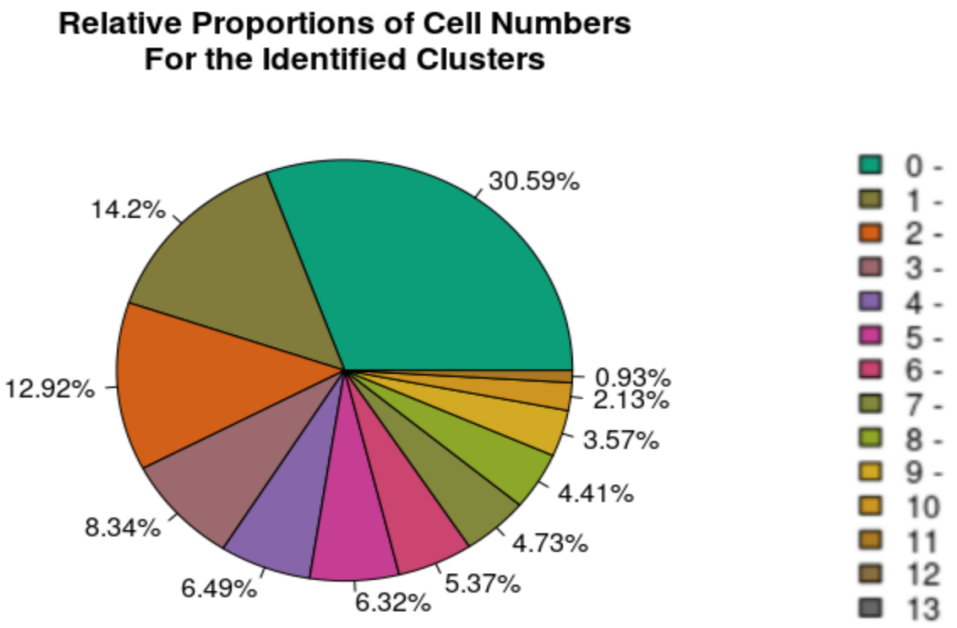


Figure 7: Pie chart shows the relative proportions of cell numbers for each identified cluster.

We used differentially expressed genes to assign gene markers into separate cell cluster types. Using literature from Baron et al., 2016, Segerstolpe et al., 2016 and the Panglao database, the top gene markers were compared and assigned endocrine cell types. For example, cluster 0 was assigned a Delta cell type according to the top 3 enriched gene markers: SST, SP100, PRRG3. A few endocrine cell types such as Beta, Apha, Pancreatic Cell Stellate (PCS), Ductal, and Acinar belonging to separate cluster numbers were merged together into one cell type for sharing enriched gene markers. Table 4 shows the top few enriched gene markers for each cell type. Delta, Beta, Alpha, and Ductal cell cluster assignment remain highly supported here by literature in that the top enriched genes are SST, INS, GCG, KRT19 respectively. There were clusters that could not be labeled using the gene markers from Baron et al. 2016. According to Baron et al., 2016, the top enriched gene for PCS is PDGFRB. However, we found a high correlation between FN1 and COL1A2 genes with PCS cells too (Segerstolpe et al., 2016). This trend can be found for Acinar and Mast cells as well. Moreover, cluster 12 aggregation lacks support and it is unknown whether it belongs to Mast cell or B cell type.

Cell Type	Novel Gene Markers	Novel Gene Markers (Baron et al., 2016)	Novel Gene Markers (Segerstolpe et al., 2016)
Delta	SST,SP100,PRRG3	SST	SST, RBP4, SEC11C, PCP4
Beta	INS,DLK1,CDKN1C,IAPP, EEF1A2, EDN3, RPS6KA5	INS	INS, SCGN, IAPP
PCS	FN1,COL1A2,BGN, COL1A1	PDGFRB	COL3A1, FN1, SFRP2, LUM SPON2, COL1A2
Acinar	REG1B,REG1A,CTRB2, CELA3A, PRSS1, SPINK1, CPA1	CPA1	REG1A, SPINK1, PRSS1, REG3A
Alpha	TTR, GCG,TM4SF4,CHGB, CRYBA2	GCG	GCG, TTR, SSR4, CRYAB2
Ductal	KRT19, KRT18, MMP7,PMEPA1, LCN2, KRT7	KRT19	SPP1, KRT19, MMP7, ANXA4
Mast	ACP5,APOE,HLA-DRA,SDS,LAPTM5,ALOX5AP,IFI30,AC007192.1,LYZ	TPSAB1, KIT, CPA3	TPSB2, TPSD1, TPSAB1, FTL, S100A4, LTC4S, B2M, CPA3

Table 4: A table of the top enriched novel genes for endocrine cell types such as Delta, Beta, PCS, Acinar, Alpha, Ductal, and Mast. Each column pertains to novel genes obtained from literature. The second column shows results from this paper. This table is meant to provide a visual to better compare novel genes against previous studies.

To resolve and support the assigned cell types, a two-dimensional projection of expressed genes was performed and displayed in figure 8. A separation of clusters was obtained for Ductal, Acinar, PCS, and Delta cell types. Additionally, supplementary figure S1 showed distinct violin plot peaks for PCS, Acinar, Ductal. However, Delta, Beta, and Alpha remain ambiguous according to the UMAP figure and violin plots.

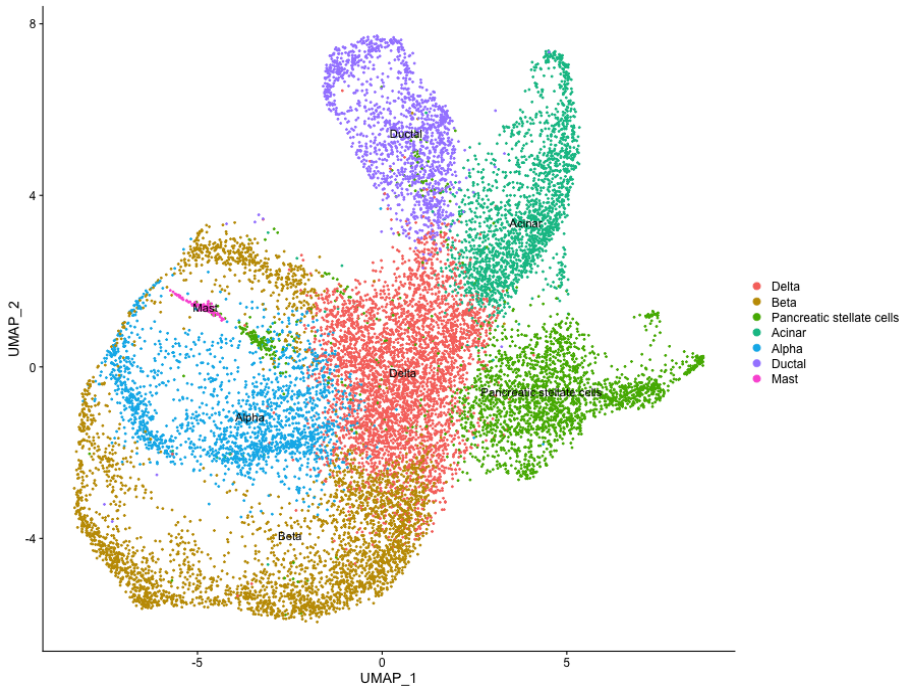


Fig 8: Two-dimensional UMAP projection and scatter plot of expressed genes (n = 19921) based on expression values (avg_log2FC). The obtained clusters were assigned to pancreas cell types. Colors correspond to cell types such as Delta, Beta, (Pancreatic stellate cells), Acinar, Alpha, Ductal, and Mast.

Heatmap with a distribution of the top 10 enriched genes for each cell cluster is shown below in figure 9. A clear distribution of highly expressed genes can be found for PSC, Acinar and Mast cell types. Highly enriched genes associated with Delta cells are also co-expressed in the remaining cell types. Additionally, it is shown that genes belonging to Beta cells are also being co-expressed in alpha and ductal cells.

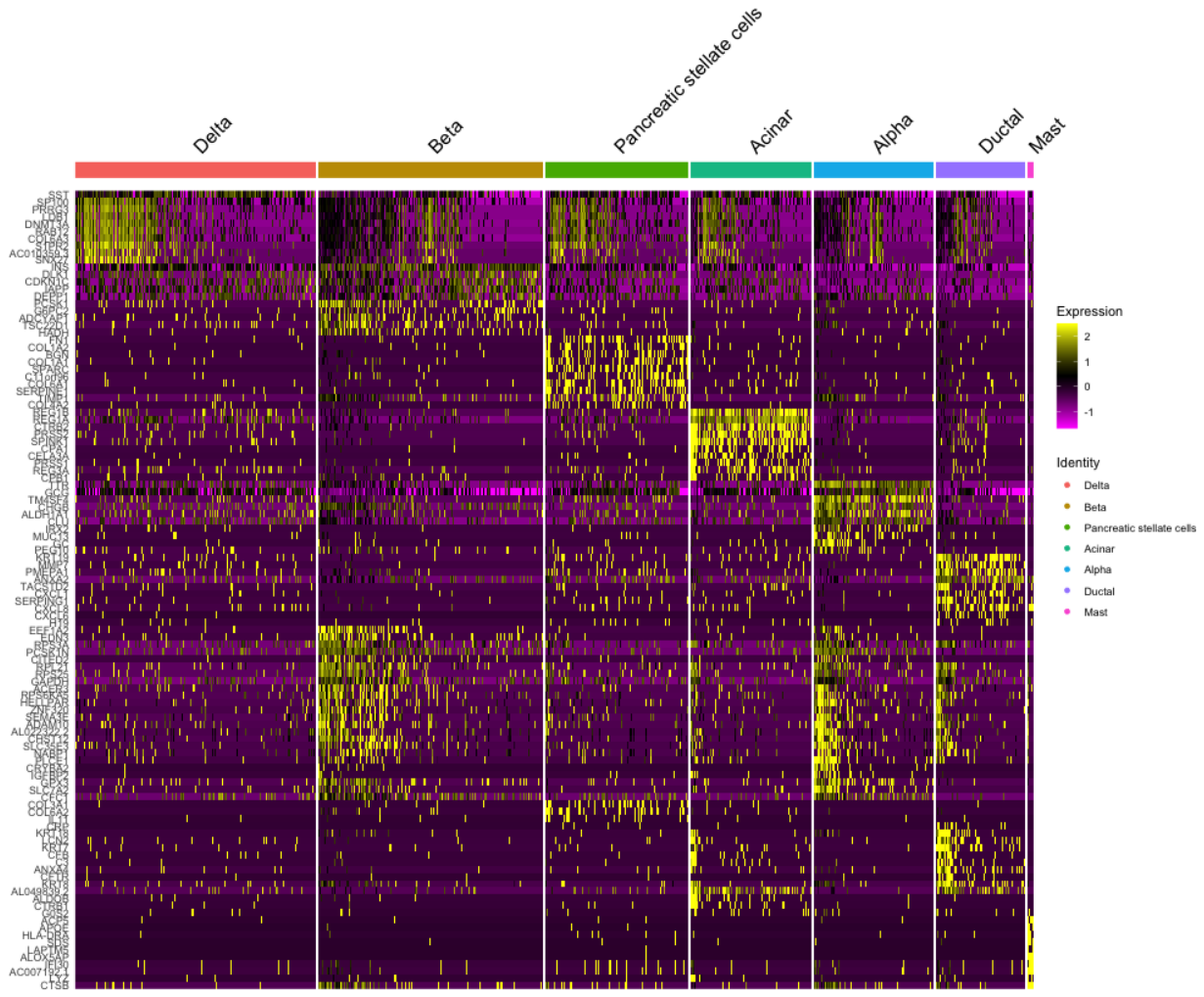


Fig 9: Heatmap with expression distribution for the top 10 enriched genes for each pancreas cell type such as Delta, Beta, Pancreatic stellate cells (PSC), Acinar, Alpha, Ductal, and Mast. The top 10 genes were selected based on average log2FC values.

Gene set enrichment analysis was performed on the top marker genes for each cluster. Functional annotation clustering was performed through DAVID. For each cluster we aimed to only include marker genes with an adjusted p-value of less than 1e-20, however the number of marker genes within this range varied greatly from cluster to cluster. As DAVID can only perform clustering on up to 3000 genes at a time, for cases where the number of marker genes with an appropriate adjusted p-value was over 3000, we manually trimmed the list such that only the 3000 genes with the lowest p-values were included. For the top three functional analyses of each gene cluster, we include a descriptive term that summarizes the findings for each cluster as shown in table 5. All gene clusters returned results through DAVID with the exception of cluster 7, which contained too many gene names unrecognized by DAVID. Further inspection of the unrecognized gene terms indicate that they may be unusual and nonstandard variants of many human genes.

Cluster	Cell Type	Top Three Annotation Clusters
0	Delta	Electron Transport Chain, Protein Localization, Mitochondrion Organization
1	Beta	Transport Vesicle, Extracellular Region, Hormone Secretion
2	Pancreatic Stellate Cells	Extracellular Region, Cell Adhesion, Extracellular Matrix
3	Acinar	Extracellular Region, Cytosolic Ribosome, Response to External Stimulus
4	Alpha	Secretory Vesicle, Hormone Metabolic Process, Hormone Activity
5	Ductal	Extracellular Region, Cell Migration, Secretory Vesicle
6	Beta	Intracellular Transport, RNA Binding, Ribosome
7	Beta	N/A
8	Alpha	Mitochondrion, Intracellular Transport, ATP Metabolic Process
9	Pancreatic Stellate Cells	Extracellular Matrix Organization, Blood Vessel Development, Focal Adhesion
10	Ductal	Focal Adhesion, Extracellular Region, Apoptotic Process
11	Acinar	Extracellular Region, Focal Adhesion, Ribosome
12	Mast	Cytoplasmic Vesicle, Secretory Vesicle, Immune Response

Table 5: Functional Annotation Analysis was performed for each cluster based on the cell marker genes. Cell marker genes were filtered to only include genes with an adjusted p-value of less than 1e-20. Functional annotation was performed through DAVID and descriptive terms were manually selected based on the results of each functional cluster output.

Discussion

Despite the ambiguity involving alpha, mast, delta, and beta cells, a clear separation and distribution for ductal, acinar, and PSC were obtained. Previous studies show that SST, INS, GCG, KRT19 are the top enriched genes for Delta, Beta, Alpha, and Ductal and that remains true in this study as well.

We were able to uncover cluster 7 cell and annotation types. For example, in the heatmap shown in figure 9, cluster 7 belonging to gene cluster (ACER3, RPS6KA5, HELLPAR, ZNF320, SEMA3E, ADAM10, AL022322.2, CHST12, SLC35E3, NABP) is distributed along beta, alpha, and ductal cell types. Additionally, in figure S1, cluster 7 had little to no peaks for all cells containing the top 10 enriched genes. We attempted to replicate the tSNE plot found in Baron et al., 2016, and it can be shown that cluster location and high separation obtained in figure 8 is not similar to the study. This could have been due to the fact that we utilized a UMAP clustering algorithm instead of a tSNE clustering algorithm. Although there were clusters that obtained high violin peaks, some were difficult to identify. For example, cluster 12 was shown to belong to Mast cell types using Segerstolpe et al., 2016 novel gene markers. However, these gene markers were ranked very low in the Mast cell type. Future studies should work to resolve cell marker identification for genes such as ACP5, APOE, HLA-DRA, SDS, LAPTM5, ALOX5AP, IFI30, AC007192.1, LYZ.

Gene set enrichment analysis provides some additional evidence for the cell types assigned to certain clusters of genes. In general, many of the cell clusters were enriched for terms related to the extracellular region and secretion of various products. These findings are expected as pancreatic cells must secrete a variety of enzymes and hormones into the extracellular space to aid in digestion and signaling related to digestion.

The acinar cell type, which plays a primary role in producing digestive enzymes (Baron et al., 2016), was shown to be enriched for genes with terms related to the extracellular region and ribosomes. While fairly general annotation terms, these follow with the expectation that the acinar cells are regularly pumping out digestive enzymes into the extracellular space, and therefore may contain an abundance of ribosomes to aid in enzyme creation. Ductal cells have been shown to secrete bicarbonate and assist in the transport of digestive enzymes (Baron et al., 2016). One of the cell clusters assigned the ductal label was enriched for terms associated with secretory vesicles, possibly related to the transportation of digestive enzymes.

The islet category of pancreatic cells are composed of alpha, beta, delta, gamma, and epsilon cells. These cells secrete hormones that regulate the level of sugar in the body (Baron et al., 2016). We found that two clusters labeled alpha and beta are enriched in terms related to hormone metabolic processes, a result that is to be expected for islet cells. However, we also find that two clusters labeled alpha and delta are enriched for terms related to ATP production and mitochondria. One possible explanation for this is that these cells may be providing energy for other islet cells to produce hormones and carry out other necessary functions. Generally evidence for the islet cell types are more difficult to resolve through this method as they perform similar functions. We also note that the cell cluster in which we were unable to perform functional annotation analysis (cluster 7) was labeled as a beta cell cluster. Our difficulty in performing an analysis may indicate that this cluster is not a true beta cluster and rather contains unusual cells that did not fit well into any other cluster. Examining figure S1 provides further evidence that this may be the case as the evidence for assigning this to the beta cell label is fairly weak.

The final two cell labels - pancreatic stellate cells and mast cells displayed fairly specific evidence in functional annotation analysis for their being assigned to each label. Pancreatic stellate cells have been implicated in angiogenesis with pancreatic tumors (Baron et al., 2016), and functional annotation analysis shows that these clusters were enriched in terms related to blood vessel development and cellular adhesion. Mast cells, a type of immune cell, were also shown to be enriched in immune response.

Conclusion

Overall, there were some uncertainties regarding which cell a gene cluster belongs to such as Alpha, Mast, Delta and Beta. We were unable to fully replicate the results from Baron et al., 2016. When considering the results presented in this paper, it is important to note that two different Seurat models were utilized to obtain parts of these results, so aspects like the number of clusters may be inconsistent in our figures. Gene set enrichment analysis provided some evidence for the clusters labeled as pancreatic stellate cells and mast cells. However, other cell clusters are a bit harder to resolve as their function generally involves secretion of chemicals into the extracellular space. Because the base function of these different types of cells is similar, they are more difficult to differentiate through functional clustering analysis.

References

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

Huang, D.W., Sherman, B.T., Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 37, 1-13 (2009).

Huang, D.W., Sherman, B.T., Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc*. 4, 44-57 (2009).

Segerstolpe, Å., Palasantza, A., Eliasson, P., Andersson, E.M., Andréasson, A.C., Sun, X., Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M.K. and Smith, D.M., 2016. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell metabolism*, *24*(4), pp.593-607.

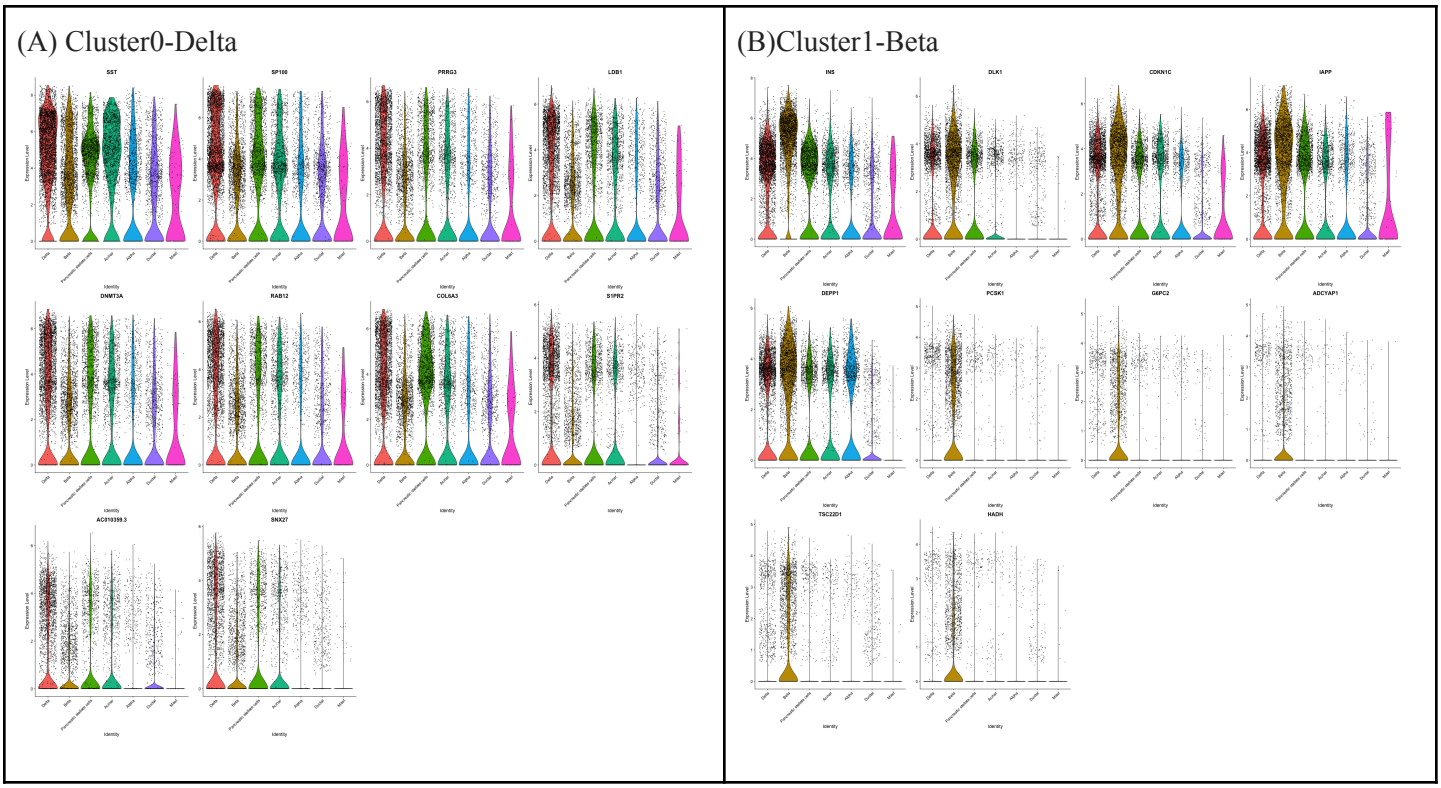
Tang, X., Huang, Y., Lei, J. *et al.* The single-cell sequencing: new developments and medical applications. *Cell Biosci* 9, 53 (2019). https://doi.org/10.1186/s13578-019-0314-y

Zhang, Y., Wang, D., Peng, M. *et al.* Single‐cell RNA sequencing in cancer research. *J Exp Clin Cancer Res* 40, 81 (2021). https://doi.org/10.1186/s13046-021-01874-1

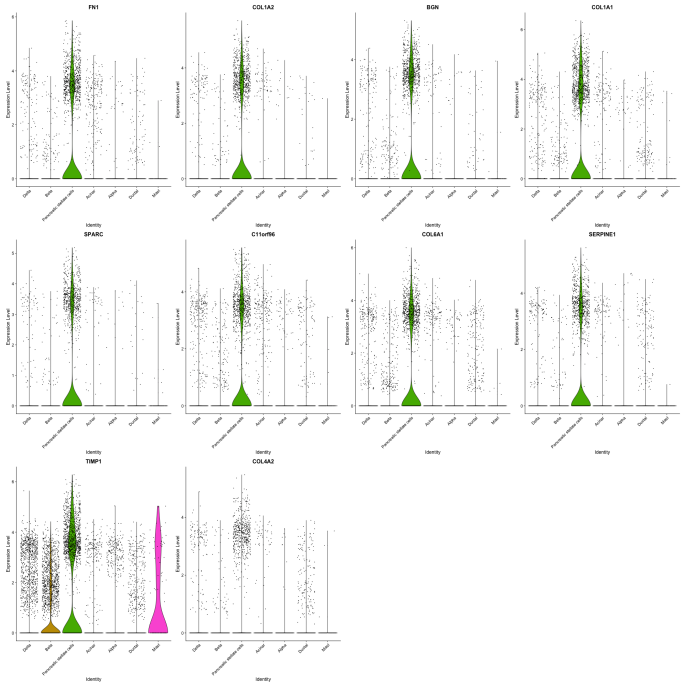
Zilionis, R., Nainys, J., Veres, A. *et al.* Single-cell barcoding and sequencing using droplet microfluidics. *Nat Protoc* 12, 44–73 (2017). https://doi.org/10.1038/nprot.2016.154

Supplementary Materials

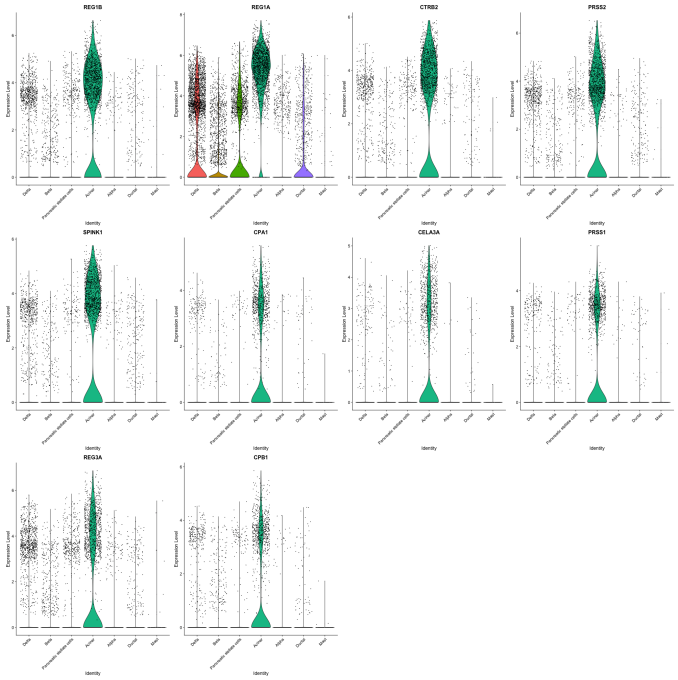
Figure S1: Violin plots showing gene expression for pancreas cell types, related to figure 8 and figure 9. Top 10 gene markers for (A) Cluster 0-Delta, (B) cluster 1-Beta, (C) cluster 2-PSC, (D) cluster 3-Acinar, (E) cluster 4-Alpha, (F) cluster 5-Ductal, (G) cluster 6-Beta, (H) cluster 7-Beta, (I) cluster 8-Alpha, (J) cluster 9-PSC, (K) cluster 10-Ductal, (L) cluster 11-Acinar, (M) cluster 12-Mast.



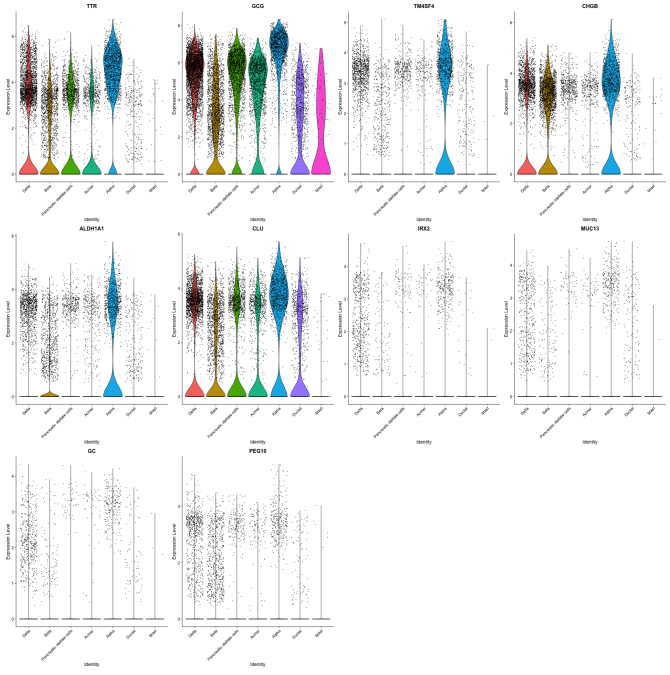
(C)Cluster2-Pancreatic Stellate Cells



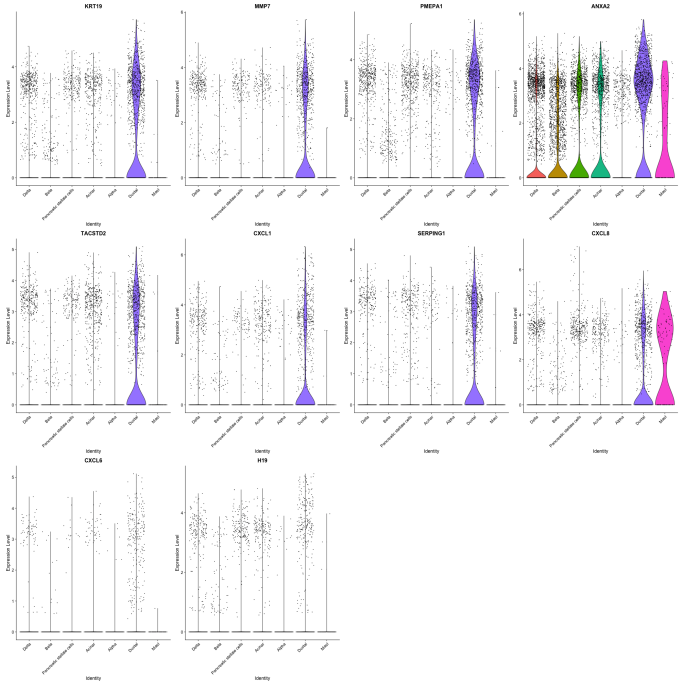
(D)Cluster3-Acinar



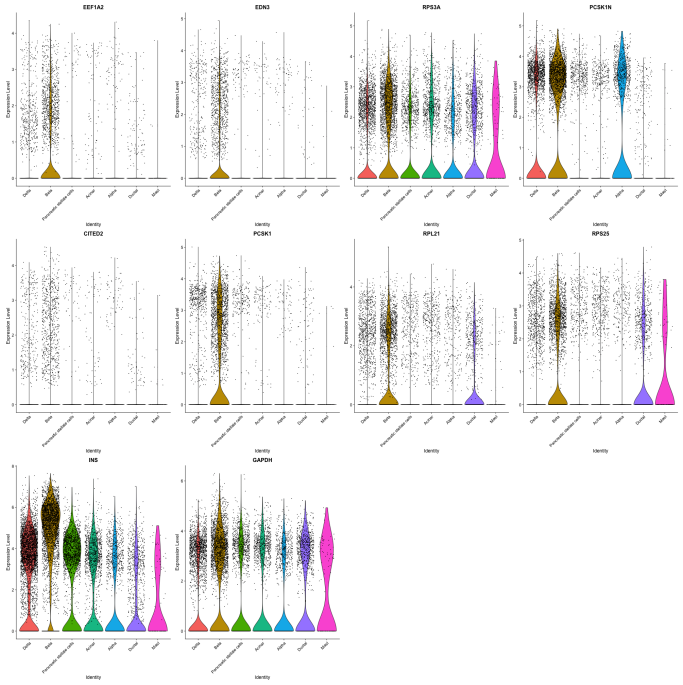
(E)Cluster4-Alpha



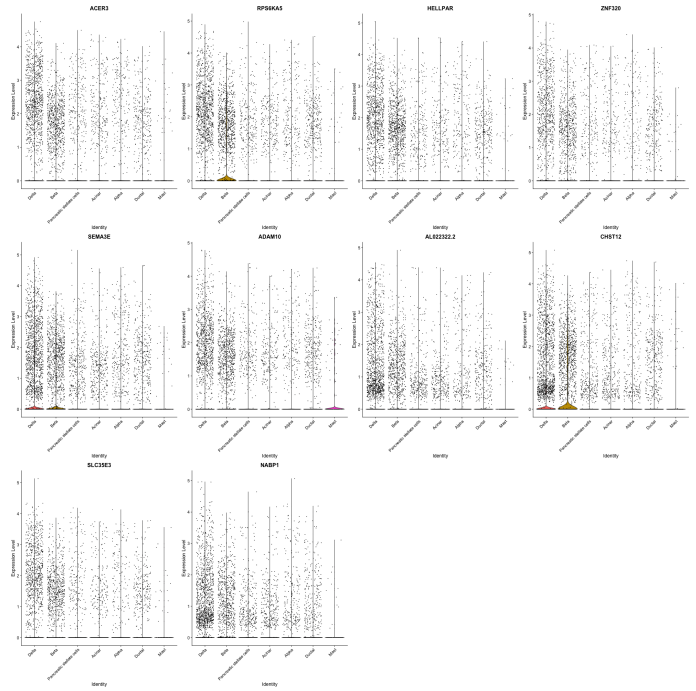
(F)Cluster5-Ductal



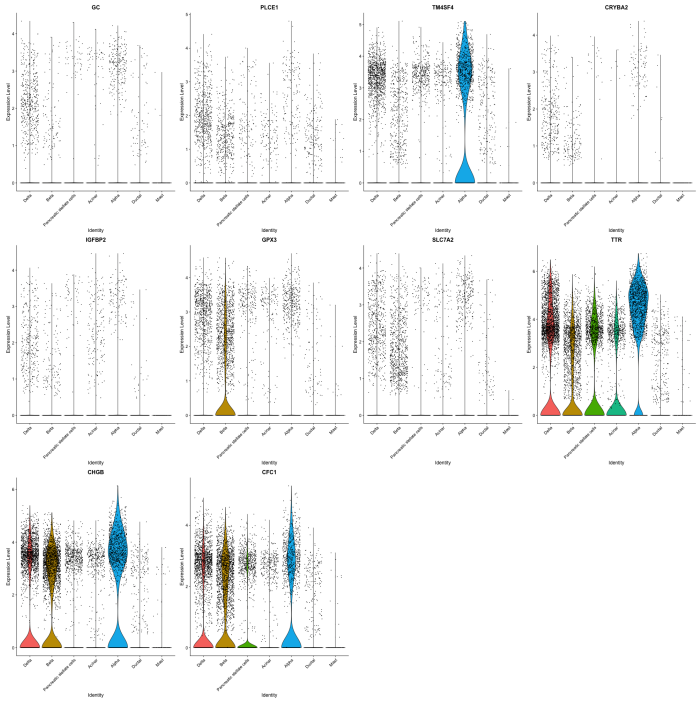
(G)Cluster6-Beta



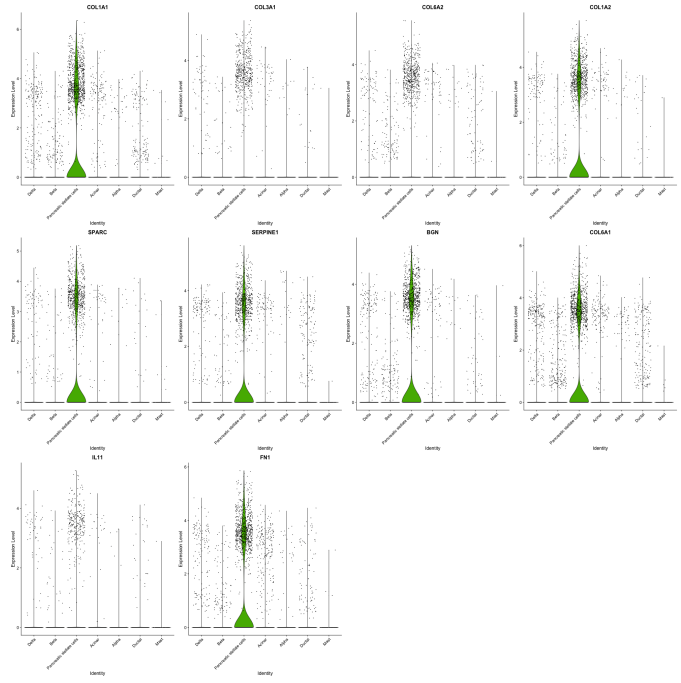
(H)Cluster7-Beta



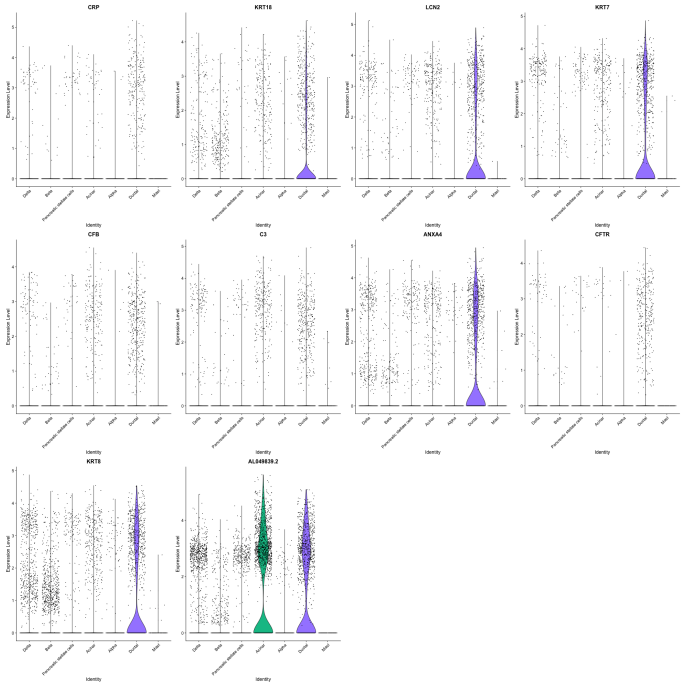
(I)Cluster8-Alpha



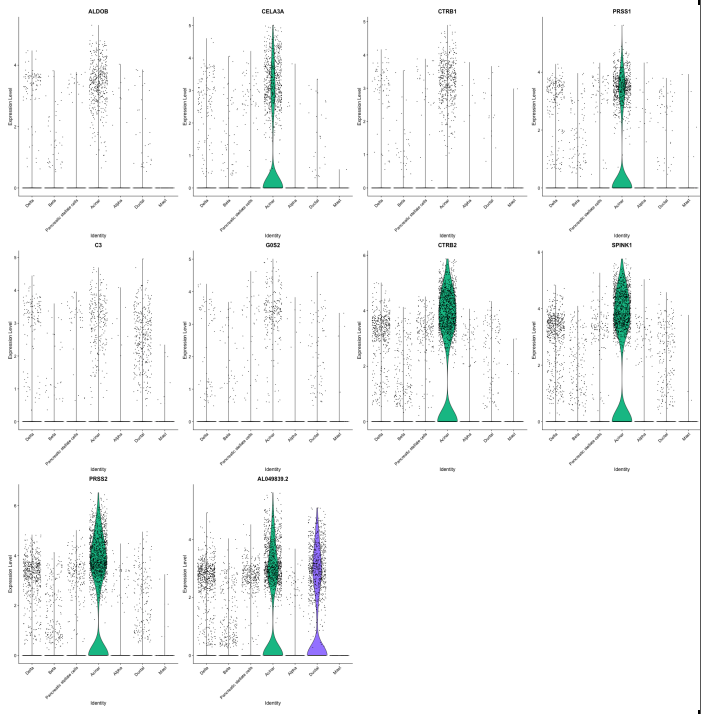
(J)Cluster9-Pancreatic stellate cells



(K)Cluster10-Ductal



(L)Cluster11-Acinar



(M)Cluster12-Mast