Introduction

In their 2016 paper, Ma'ayan et. al. elucidated the complex cell types present in the mammalian pancreas using droplet-based single-cell RNA-seq. Previously, gene expression data had been reported for only mixtures of cell types, which lacked specificity and would have yielded noisier and less informative data. The authors were able to distinguish between 15 distinct clusters to match previously known cell types. In addition, they were able to identify new subtypes of cells, based on their distinct expression profiles. This large dataset will be helpful in the future, as it generated a wealth of information for more cell-type-specific gene expression analyses.

Data

In order to reproduce the paper's findings, we used the GEO accession number that was provided in the paper. Metadata files for single cell RNA-seq were collected from the NCBI GEO with accession number GSE84133. A single- cell transcript of over 12,000 cells from four human donors and two mouse strains were generated using iDrop. For our analysis, three files associated with the 51 years old female donor were selected. Sample SRR3879604, SRR3879605, and SRR3879606 were downloaded for our analysis.

Once we obtained the single-cell RNA-seq data, Sample IDs from the female donors were examined and calculated the number of reads per distinct barcode. For our analysis, we used awk to extract barcodes and count the number of reads by barcodes, our sample counts indicated 1.2 to 1.3 million cell barcodes per sample. further, cumulative distribution plot of reads among the barcodes are plotted and shown in (Supplementary 1). With aim of eliminating reads with infrequent barcodes, we performed filtering and sorting the cell barcodes against a whitelist of expected cell barcodes and 44 to 50 thousand frequent barcodes were discovered. Whitelist informative barcodes were transferred to generate UMI matrix using salmon alevin.

In order to generate a unique molecular identifier (UMI) matrix, an index of human reference transcriptome was generated. This reference transcriptome is given to salmon in the form of FASTQ files. For our experiment, We used the latest GENCODE version 37 human reference transcriptome. The module salmon and reference transcriptome were submitted within a script as a qsub to generate an index. Finally, running salmon alevin once an index has been generated to produce alevin output files. The output file of salmon alevin command contains the quantification RNA-seq data. Overall, our salmon output reads didn't show any concerning mapping rate.

Methods

The salmon alevin counts file was loaded into R using tximport (Charlotte). The alevin files take a while to load, but this load time was cut much shorter after importing fishpond (Zhu). Ensembl gene ids were used in the UMI counts matrix, which had to be mapped to gene symbols for our further analysis. Using Ensembl's library, EnsDb.Hsapiens.v79, the gene symbols were mapped to the count matrix.

The Bioconductor package, Seurat, was used as a tool for single cell sequencing quality control (Stuart). The UMI matrix was used to create our Seurat object. Seurat allows us to store the data as well as analyses of the data in the object. The initial object was 20987 features across 2753 samples within 1 assay.

We then performed quality control, QC, measures by filtering the data. The filtering criteria was decided after generating a violin plot to visualize our QC metrics (Figure 1). The mitochondrial QC metrics were calculated to help eliminate low quality and dying cells. This included the percentage of mitochondrial, mt, genes. The low quality cells were filtered, removing any cells with greater than 35 percent mt. We then subset our Seurat object, pbmc, to only greater than 200 features but less than 3500 features. After filtering our object was made of 20987 features across 2543 samples.

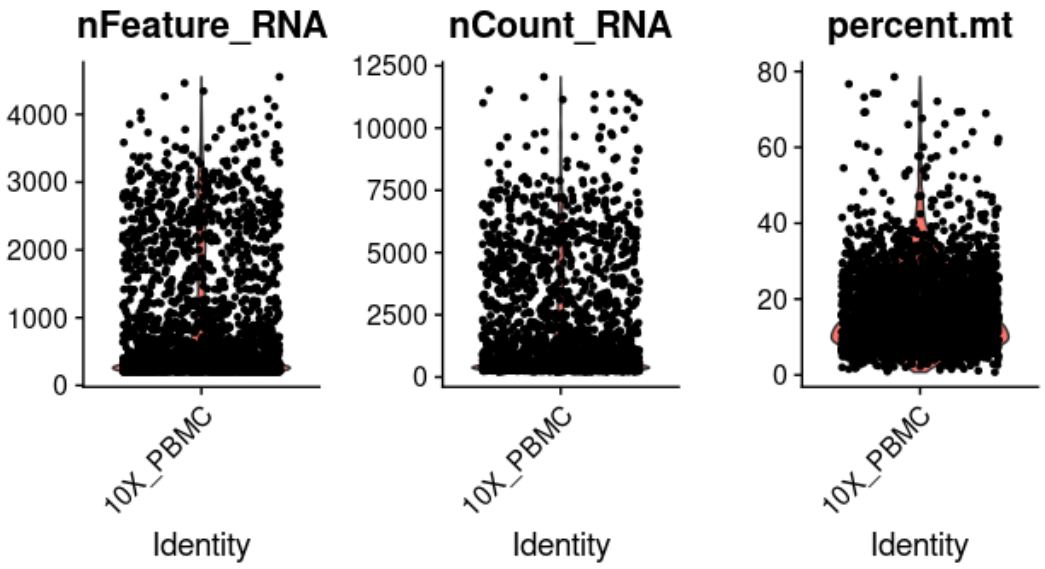


Figure 1. Violin Plot visualizing quality control metrics before filtering

Once unwanted cells were removed from our pbmc dataset, the data then was normalized. We normalized the data using the LogNormalize method. Next, we filtered out the low variance genes in the dataset. To do this, we included the identified highly variable features in our dataset to use for further analysis. The top 10 most variable features are identified and labeled in the plot below (Figure 2). After this filtering, the dataset still consisted of 20987 features across 2543 samples.

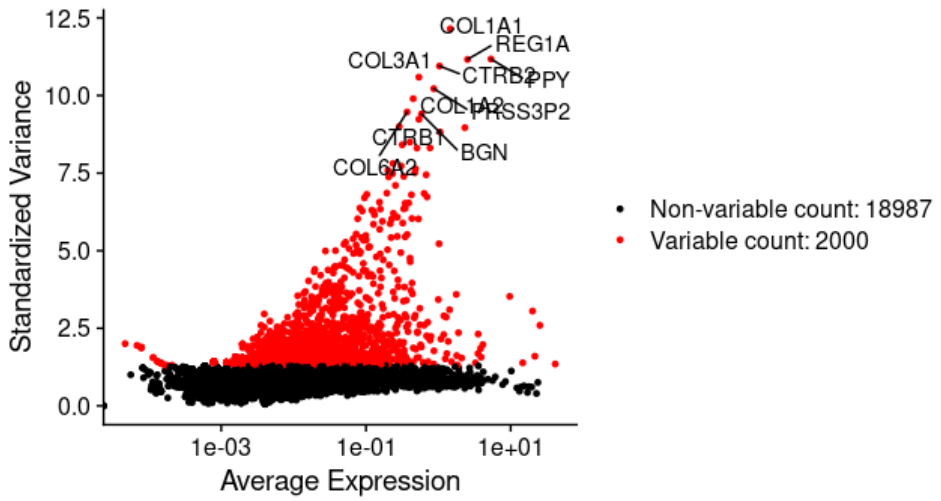


Figure 2. The Variable Features

The data was then scaled as it is a necessary step prior to performing PCA, a dimensional reduction technique. To determine the dimensionality of the dataset we used JackStraw scores as well as visually observing an elbow plot (Figure 3). The plot shows a bend between the 10-12 principle components, suggesting we may have around 10 clusters.

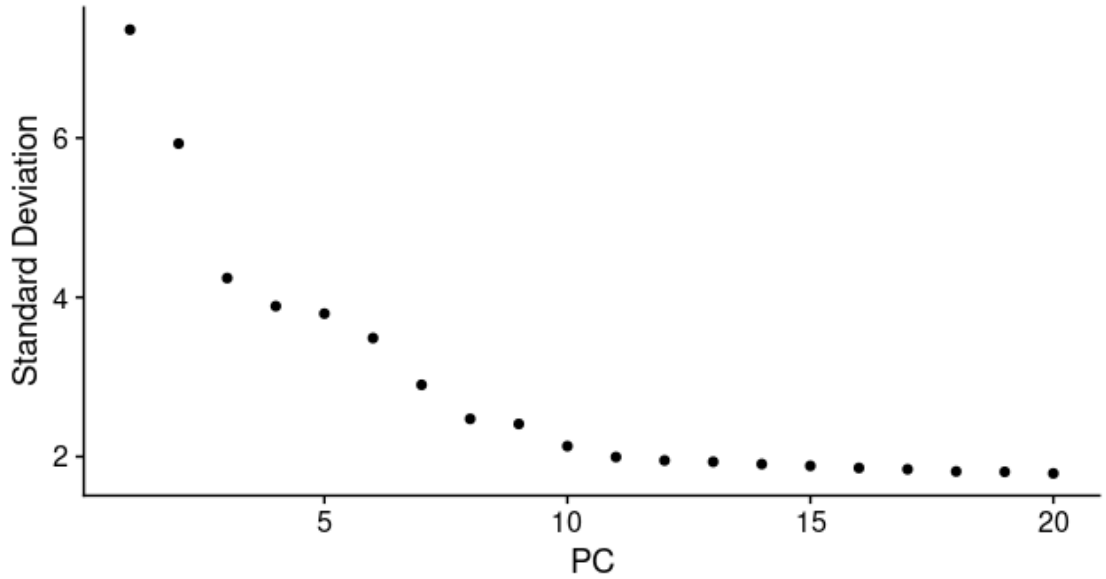


Figure 3. Elbow plot

Finally, we were able to cluster the cells using the FindNeighbors and FindClusters functions in R. In total there were 11 clusters found. The relative proportions of cell numbers of the clusters are visually shown in the pie chart and bar plot below (Figure 4 and 5).

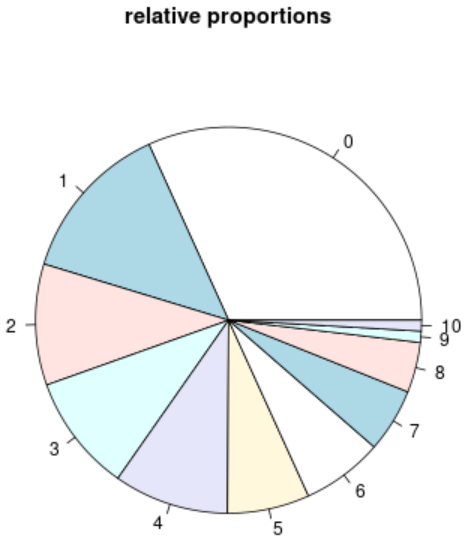


Figure 4. Pie chart showing relative proportions of cell numbers of clusters

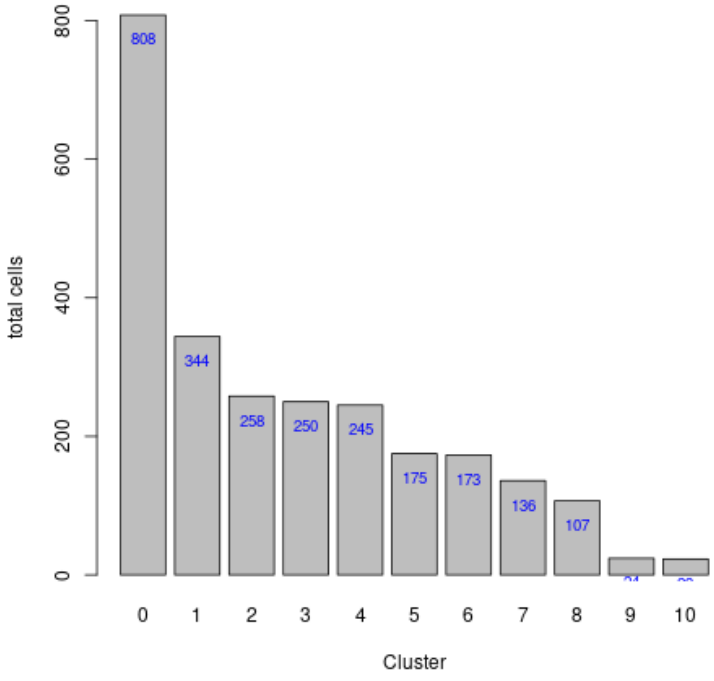


Figure 5. Bar plot showing our clusters found

Through the Seurat package (version 4.0.1) the FindAllMarkers tool was utilized to identify the marker genes within each cluster. To detect the marker genes, we set the minimum percentage of expressed genes and log fold change threshold to 0.1 respectively. Once marker genes were identified, their respective cluster groups were renamed to their corresponding cell type using RenameIdents. Following, we generated: a UMAP of the all clustered cell types, individual UMAPs of marker genes, a heatmap and violinplots of gene expressions, and identified the top three and top five differentially expressed genes in each cluster.

Gene set enrichment was performed using the enrichR package (version 3.0), an R interface to the Ma'ayan lab online EnrichR tool. Functional terms were imported from GO databases for Molecular Function, Cellular Component, and Biological Process (2015 version). Differential gene expression results were split by cell cluster, then filtered by average log2 fold change (greater than 0.5) or adjusted p-value (less than 0.05), separately, and the top ten significant results according to adjusted p-value are reported.

Results

Once positive marker genes were identified through the use of FindAllMarker tool and filtering, we generated a UMAP projection displaying the relatedness between the clusters(Figure 6). As shown in the figure, Alpha cells are closely associated with Gamma, Acinar, Delta, Quiescent steliate, and Epsilon cell types. Endothelial and Macrophage cells were found to be the least related to cell types (Figure 6). When looking at individual marker gene expression across the clusters, it was found that GCG was highly expressed in multiple cell types (Figure 7). In addition to GCG; INS, SST, and PPY showed a similar pattern. The marker gene found to be highly specific in terms of gene expression within cell types was found to be SDS (Figure 7). SDS is highly expressed only in macrophage cells (Figure 7). This is further supported by the generated violin plots (Figure 8). Furthermore, the generated heatmap of the top five genes within each cluster provides insight into which genes are highly expressed within each cell type starting with Alpha cells ending with Macrophage cells (Figure 9). Highly expressed genes are represented by yellow markers within the figure (Figure 9). The top three and top five expressed within each cell type were also generated (Table 1, Table 2). These tables provide additional insight into highly expressed genes within each cell type cluster.

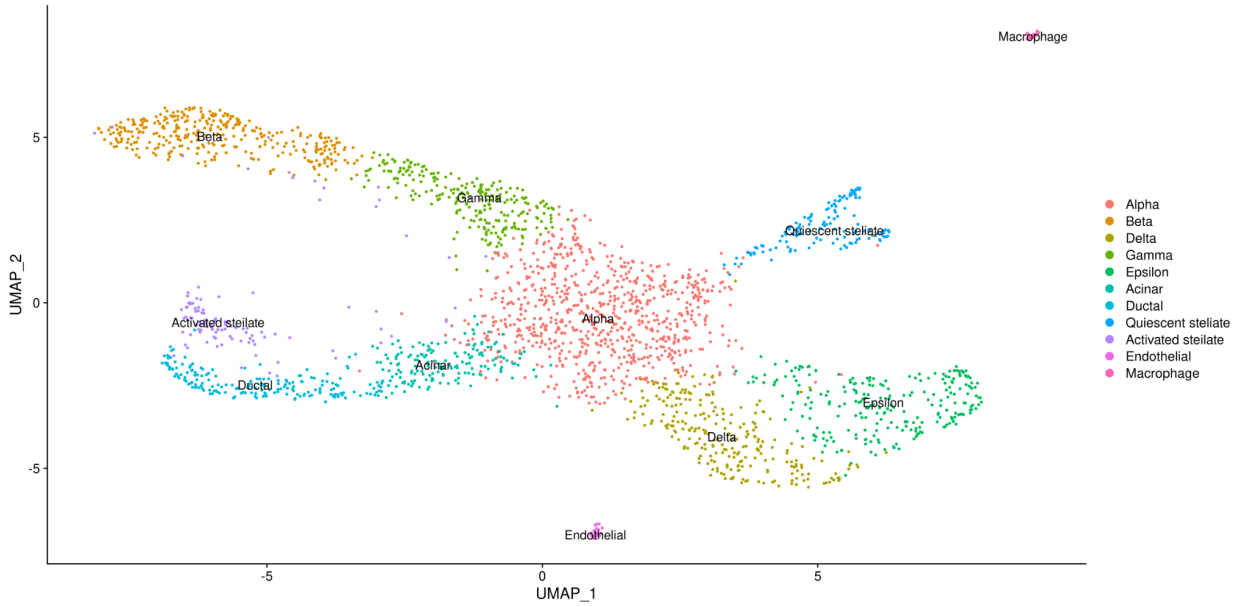


Figure 6: UMAP demonstrates the relatedness between cell types

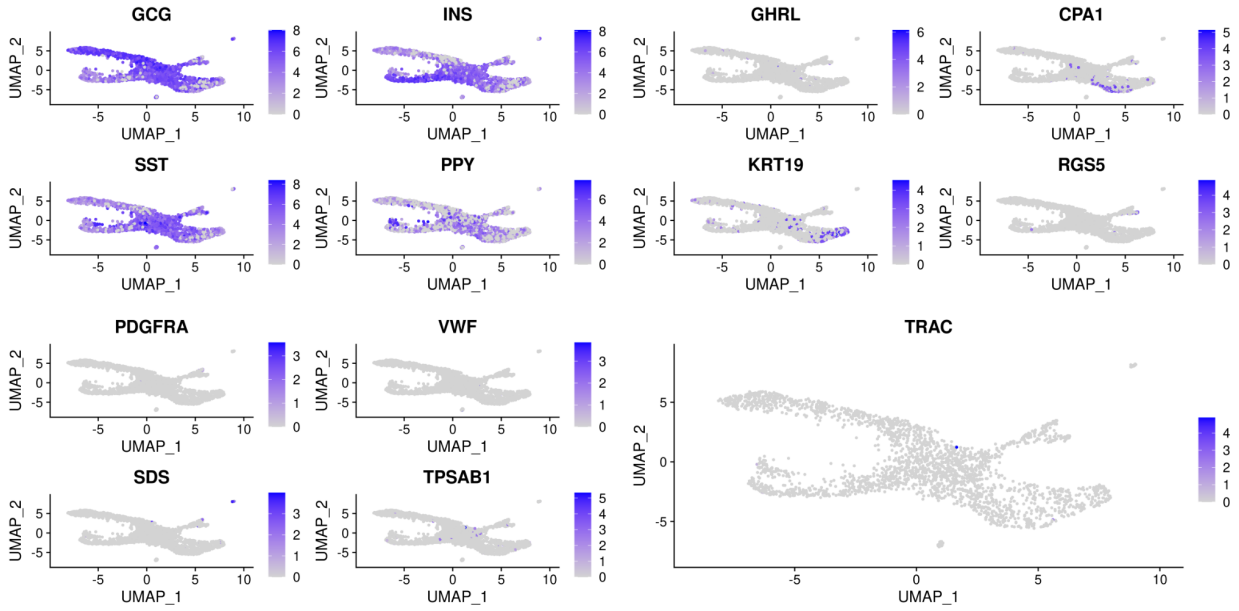


Figure 7: UMAP gene expression within cell types

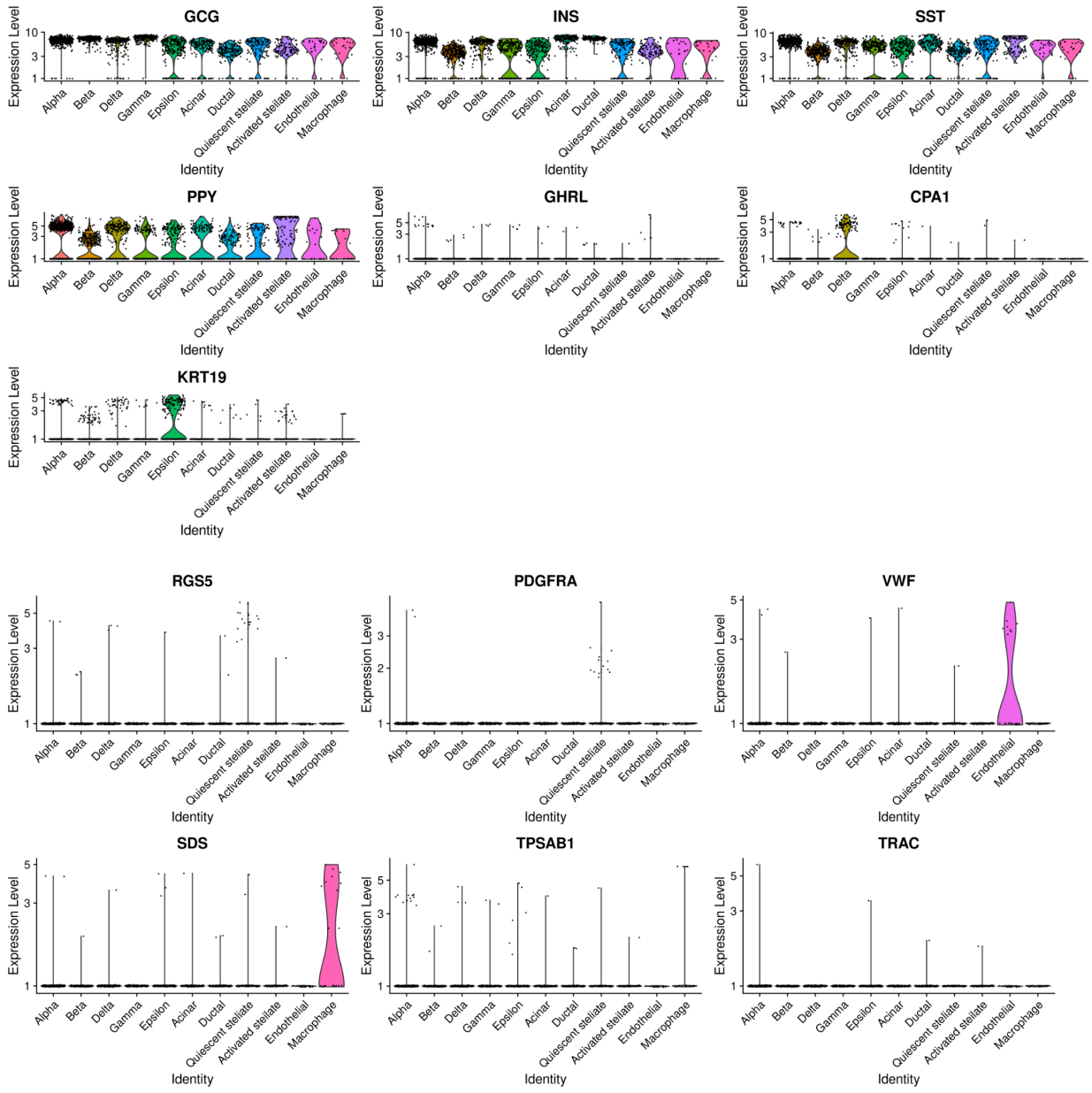


Figure 8: Gene expression within cell types

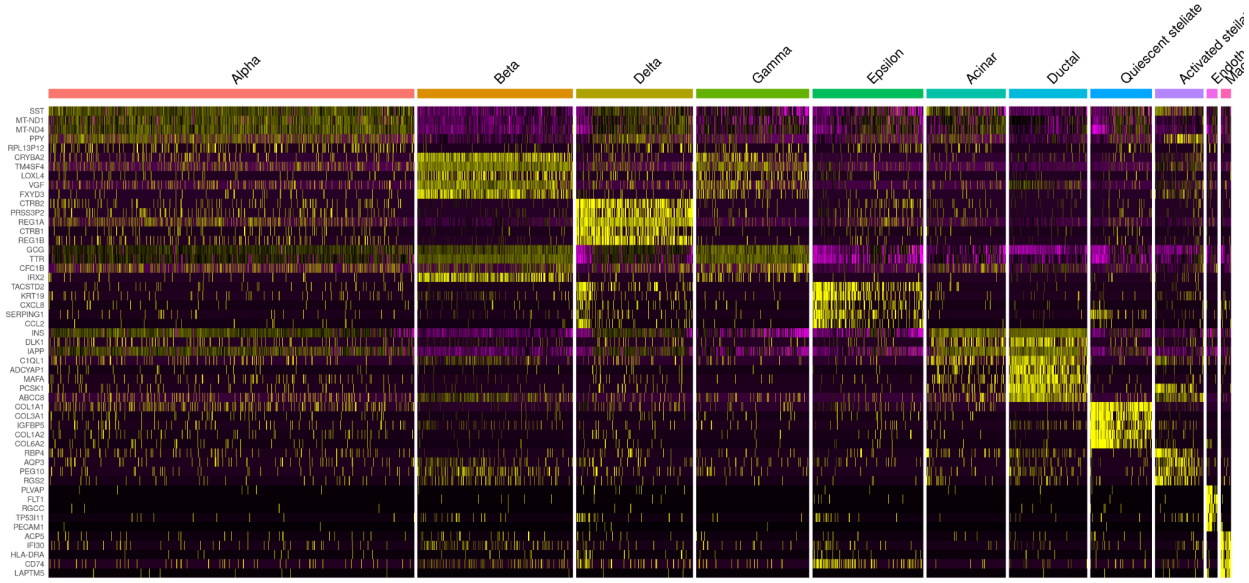


Figure 9: Heatmap of gene expression within each cell type cluster

	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene
1	2.477610e-143	1.2926883	0.983	0.896	5.199760e-139	Alpha	SST
2	5.882769e-142	1.0391789	0.998	0.990	1.234617e-137	Alpha	MT-ND1
3	5.286549e-154	0.9643158	1.000	0.995	1.109488e-149	Alpha	MT-ND4
4	1.581748e-169	1.7839936	0.846	0.131	3.319615e-165	Beta	CRYBA2
5	3.215127e-116	1.6002803	0.980	0.333	6.747587e-112	Beta	TM4SF4
6	6.205114e-113	1.5703506	0.541	0.074	1.302267e-108	Beta	LOXL4
7	1.114947e-153	3.9823643	0.694	0.095	2.339940e-149	Delta	CTRB2
8	9.080237e-122	3.8602144	0.632	0.102	1.905669e-117	Delta	PRSS3P2
9	1.385115e-142	3.7696462	0.926	0.303	2.906942e-138	Delta	REG1A
10	1.160453e-70	1.5380713	0.996	0.954	2.435442e-66	Gamma	GCG
11	7.326476e-07	1.4812995	0.224	0.127	1.537607e-02	Gamma	LOXL4
12	1.931702e-79	1.4100466	1.000	0.909	4.054062e-75	Gamma	TTR
13	6.093398e-119	2.8421830	0.522	0.057	1.278821e-114	Epsilon	TACSTD2
14	9.528721e-76	2.5353496	0.469	0.082	1.999793e-71	Epsilon	KRT19
15	1.429051e-57	2.5059679	0.318	0.045	2.999149e-53	Epsilon	CXCL8
16	1.682391e-46	1.8606672	0.989	0.893	3.530833e-42	Acinar	INS
17	1.069639e-13	1.6809230	0.326	0.137	2.244852e-09	Acinar	DLK1
18	2.915931e-29	1.6699835	0.931	0.811	6.119665e-25	Acinar	IAPP
19	6.265273e-138	2.1176712	0.694	0.065	1.314893e-133	Ductal	MAFA
20	3.272611e-133	1.9562684	0.699	0.071	6.868228e-129	Ductal	PCSK1
21	1.212070e-49	1.8789420	0.977	0.808	2.543771e-45	Ductal	IAPP
22	6.028456e-106	4.7310790	0.824	0.160	1.265192e-101	Quiescent steliate	COL1A1
23	1.760437e-164	4.5937248	0.654	0.042	3.694628e-160	Quiescent steliate	COL3A1
24	1.086803e-99	4.0804658	0.640	0.082	2.280874e-95	Quiescent steliate	IGFBP5
25	1.860480e-06	2.9226896	0.748	0.556	3.904590e-02	Activated steilate	PPY
26	1.243697e-43	2.4350894	0.505	0.086	2.610148e-39	Activated steilate	RBP4
27	1.061466e-44	1.8168795	0.542	0.098	2.227699e-40	Activated steilate	AQP3
28	2.437183e-119	4.7451198	0.583	0.010	5.114915e-115	Endothelial	PLVAP
29	2.650725e-220	4.4711308	0.792	0.008	5.563077e-216	Endothelial	FLT1
30	2.219247e-190	3.8981094	0.667	0.006	4.657533e-186	Endothelial	RGCC
31	1.737694e-59	5.3084302	0.696	0.035	3.646899e-55	Macrophage	ACP5
32	1.643697e-34	4.7598331	0.783	0.090	3.449627e-30	Macrophage	IFI30
33	7.084000e-74	4.4993557	0.652	0.023	1.486719e-69	Macrophage	HLA-DRA

Table 1: Top three marker genes within each cell type cluster

	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene
1	2.477610e-143	1.2926883	0.983	0.896	5.199760e-139	Alpha	SST
2	5.882769e-142	1.0391789	0.998	0.990	1.234617e-137	Alpha	MT-ND1
3	5.286549e-154	0.9643158	1.000	0.995	1.109488e-149	Alpha	MT-ND4
4	4.030889e-46	0.8822608	0.668	0.516	8.459627e-42	Alpha	PPY
5	7.242278e-08	0.8611667	0.163	0.096	1.519937e-03	Alpha	RPL13P12
6	1.581748e-169	1.7839936	0.846	0.131	3.319615e-165	Beta	CRYBA2
7	3.215127e-116	1.6002803	0.980	0.333	6.747587e-112	Beta	TM4SF4
8	6.205114e-113	1.5703506	0.541	0.074	1.302267e-108	Beta	LOXL4
9	8.279152e-112	1.5435453	0.913	0.255	1.737546e-107	Beta	VGF
10	4.967870e-188	1.5343748	0.744	0.083	1.042607e-183	Beta	FXYD3
11	1.114947e-153	3.9823643	0.694	0.095	2.339940e-149	Delta	CTRB2
12	9.080237e-122	3.8602144	0.632	0.102	1.905669e-117	Delta	PRSS3P2
13	1.385115e-142	3.7696462	0.926	0.303	2.906942e-138	Delta	REG1A
14	3.243576e-124	3.7194810	0.512	0.054	6.807294e-120	Delta	CTRB1
15	3.654855e-152	3.6989235	0.659	0.081	7.670444e-148	Delta	REG1B
16	1.160453e-70	1.5380713	0.996	0.954	2.435442e-66	Gamma	GCG
17	7.326476e-07	1.4812995	0.224	0.127	1.537607e-02	Gamma	LOXL4
18	1.931702e-79	1.4100466	1.000	0.909	4.054062e-75	Gamma	TTR
19	4.401052e-11	1.3298117	0.588	0.462	9.236488e-07	Gamma	CFC1B
20	8.239862e-04	1.2876778	0.192	0.133	1.000000e+00	Gamma	IRX2
21	6.093398e-119	2.8421830	0.522	0.057	1.278821e-114	Epsilon	TACSTD2
22	9.528721e-76	2.5353496	0.469	0.082	1.999793e-71	Epsilon	KRT19
23	1.429051e-57	2.5059679	0.318	0.045	2.999149e-53	Epsilon	CXCL8
24	1.542315e-69	2.4664396	0.420	0.070	3.236856e-65	Epsilon	SERPING1
25	7.688071e-62	2.3825358	0.306	0.037	1.613496e-57	Epsilon	CCL2
26	1.682391e-46	1.8606672	0.989	0.893	3.530833e-42	Acinar	INS
27	1.069639e-13	1.6809230	0.326	0.137	2.244852e-09	Acinar	DLK1
28	2.915931e-29	1.6699835	0.931	0.811	6.119665e-25	Acinar	IAPP
29	3.991467e-09	1.4482539	0.343	0.196	8.376891e-05	Acinar	C1QL1
30	5.528308e-05	1.3278037	0.137	0.063	1.000000e+00	Acinar	ADCYAP1
31	6.265273e-138	2.1176712	0.694	0.065	1.314893e-133	Ductal	MAFA
32	3.272611e-133	1.9562684	0.699	0.071	6.868228e-129	Ductal	PCSK1
33	1.212070e-49	1.8789420	0.977	0.808	2.543771e-45	Ductal	IAPP
34	5.207168e-66	1.8620535	0.809	0.236	1.092828e-61	Ductal	ABCC8
35	1.605812e-154	1.8289338	0.566	0.032	3.370117e-150	Ductal	ADCYAP1
36	6.028456e-106	4.7310790	0.824	0.160	1.265192e-101	Quiescent steliate	COL1A1
37	1.760437e-164	4.5937248	0.654	0.042	3.694628e-160	Quiescent steliate	COL3A1
38	1.086803e-99	4.0804658	0.640	0.082	2.280874e-95	Quiescent steliate	IGFBP5
39	9.971683e-148	4.0284497	0.640	0.046	2.092757e-143	Quiescent steliate	COL1A2
40	3.555860e-152	4.0236305	0.588	0.034	7.462684e-148	Quiescent steliate	COL6A2
41	1.860480e-06	2.9226896	0.748	0.556	3.904590e-02	Activated steliate	PPY
42	1.243697e-43	2.4350894	0.505	0.086	2.610148e-39	Activated steliate	RBP4
43	1.061466e-44	1.8168795	0.542	0.098	2.227699e-40	Activated steliate	AQP3
44	1.933380e-29	1.5759333	0.458	0.102	4.057584e-25	Activated steliate	PEG10
45	1.785474e-42	1.4226861	0.430	0.064	3.747174e-38	Activated steliate	RGS2
46	2.437183e-119	4.7451198	0.583	0.010	5.114915e-115	Endothelial	PLVAP
47	2.650725e-220	4.4711308	0.792	0.008	5.563077e-216	Endothelial	FLT1
48	2.219247e-190	3.8981094	0.667	0.006	4.657533e-186	Endothelial	RGCC
49	6.641236e-48	3.6974822	0.625	0.037	1.393796e-43	Endothelial	TP53I11
50	4.644577e-162	3.6016730	0.542	0.004	9.747573e-158	Endothelial	PECAM1
51	1.737694e-59	5.3084302	0.696	0.035	3.646899e-55	Macrophage	ACP5
52	1.643697e-34	4.7598331	0.783	0.090	3.449627e-30	Macrophage	IFI30
53	7.084000e-74	4.4993557	0.652	0.023	1.486719e-69	Macrophage	HLA-DRA
54	1.269605e-26	4.4898720	0.826	0.131	2.664521e-22	Macrophage	CD74
55	1.073637e-205	4.3933435	0.783	0.008	2.253242e-201	Macrophage	LAPTM5

Table 2: Top five marker genes within each cell type cluster

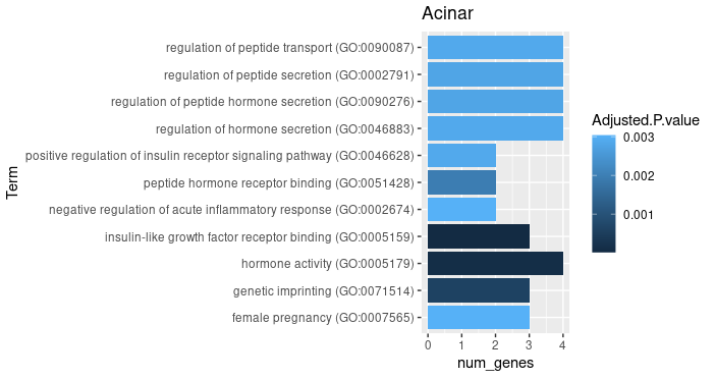
Discussion

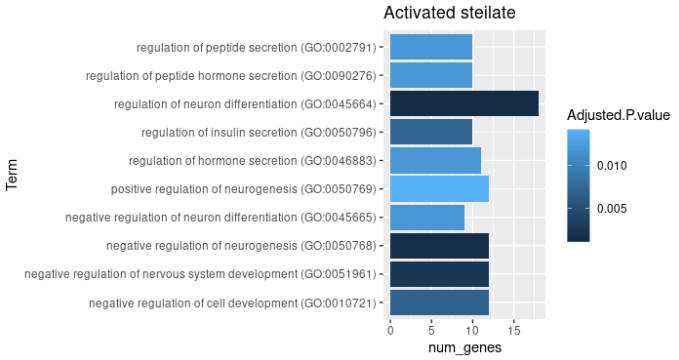
Though our analysis was generated using the same data, we were unable to identify all clusters found in the original paper. While we were able to identify most types, Cytotoxic T cells and mast cells were missing from our results. As we were unable to include these clusters we lost marker genes CD3, CD8, TPSAB1, KIT, and CPA3. To investigate this, we generated a number of figures from the paper. Our UMAP (Figure 6) and TSNE plots (Supplemental Figure 2) show different results from that in the original paper. However, our heatmap of marker gene expression within cell types is similar to that of Baron et. al. Again, the differences in the heatmap are due to the loss of Cytotoxic T cells and mast cells from the data.

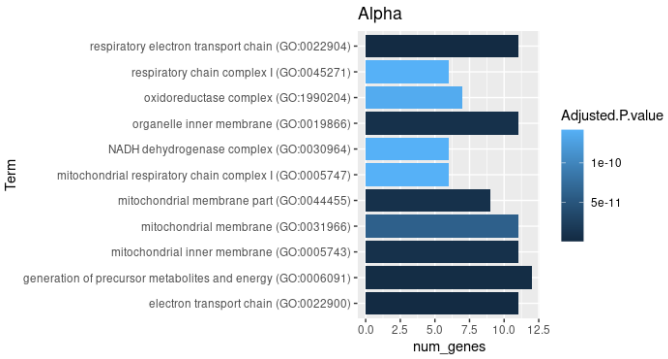
Gene set enrichment analysis of differentially expressed genes revealed functional differences between cell types that generally aligned with their known function. The following plots represent the functional enrichment present in marker gene sets with the highest average log2 fold change and then the top 10 results with smallest adjusted p-values, for each cluster.

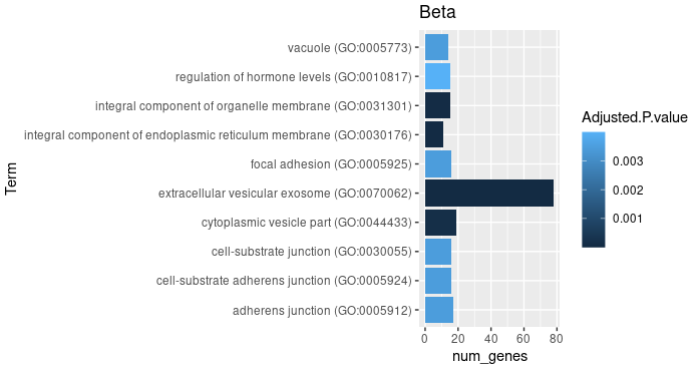
Term	Cluster	Adjusted P-value	Gene Count	Source
insulin-like growth factor receptor binding (GO:0005159)	Acinar	0.00022798	2	Molecular Function
regulation of insulin secretion (GO:0050796)	Acinar	0.00217143	3	Biological Process
regulation of peptide hormone secretion (GO:0090276)	Acinar	0.00217143	3	Biological Process
regulation of neuron differentiation (GO:0045664)	Activated steilate	0.00080576	18	Biological Process
regulation of insulin secretion (GO:0050796)	Activated steilate	0.00604952	10	Biological Process
respiratory electron transport chain (GO:0022904)	Alpha	6.48E-16	10	Biological Process
electron transport chain (GO:0022900)	Alpha	6.48E-16	10	Biological Process
generation of precursor metabolites and energy (GO:0006091)	Alpha	1.16E-12	11	Biological Process
mitochondrial inner membrane (GO:0005743)	Alpha	8.19E-12	10	Cellular Component
extracellular vesicular exosome (GO:0070062)	Beta	6.18E-07	78	Cellular Component
focal adhesion (GO:0005925)	Beta	0.00344812	16	Cellular Component
extracellular vesicular exosome (GO:0070062)	Delta	1.35E-12	39	Cellular Component
digestion (GO:0007586)	Delta	0.00028854	6	Biological Process
glutathione transferase activity (GO:0004364)	Delta	0.00040044	4	Molecular Function
serine-type endopeptidase activity (GO:0004252)	Delta	0.00040044	7	Molecular Function
viral transcription (GO:0019083)	Ductal	2.72E-20	22	Biological Process
translational elongation (GO:0006414)	Ductal	3.62E-20	24	Biological Process
angiogenesis (GO:0001525)	Endothelial	4.24E-23	38	Biological Process
regulation of vasculature development (GO:1901342)	Endothelial	3.81E-14	27	Biological Process
regulation of angiogenesis (GO:0045765)	Endothelial	2.53E-13	25	Biological Process
regulation of endothelial cell migration (GO:0010594)	Endothelial	3.69E-13	19	Biological Process
extracellular vesicular exosome (GO:0070062)	Epsilon	3.17E-59	199	Cellular Component
anchoring junction (GO:0070161)	Epsilon	3.30E-24	54	Cellular Component
adherens junction (GO:0005912)	Epsilon	2.39E-23	52	Cellular Component
hormone metabolic process (GO:0042445)	Gamma	8.22E-05	5	Biological Process
peptide hormone processing (GO:0016486)	Gamma	0.00055114	3	Biological Process
response to dietary excess (GO:0002021)	Gamma	0.00226985	2	Biological Process
regulation of cell activation (GO:0050865)	Macrophage	1.12E-19	42	Biological Process
regulation of leukocyte activation (GO:0002694)	Macrophage	1.91E-17	38	Biological Process
regulation of lymphocyte activation (GO:0051249)	Macrophage	8.01E-15	33	Biological Process
activation of immune response (GO:0002253)	Macrophage	1.22E-13	37	Biological Process
extracellular matrix (GO:0031012)	Quiescent steliate	7.57E-52	72	Cellular Component
extracellular vesicular exosome (GO:0070062)	Quiescent steliate	1.75E-29	140	Cellular Component

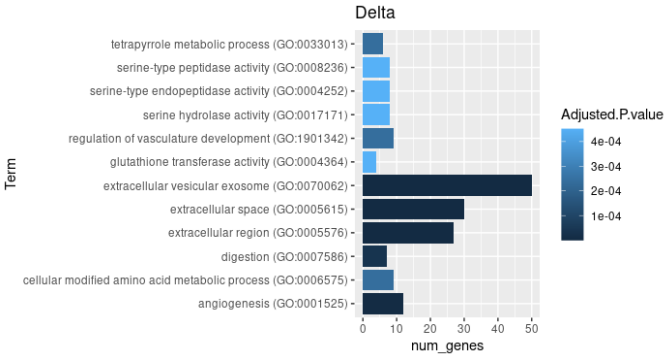
Table 3. Top gene set enriched terms by cluster

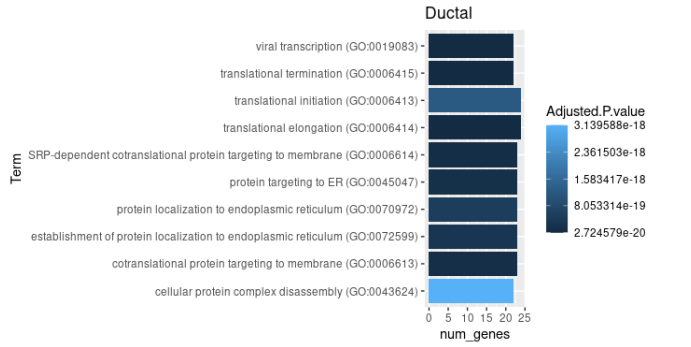

Acinar
regulation of peptide transport (GO:0090087)
regulation of peptide secretion (GO:0002791)
regulation of peptide hormone secretion (GO:0090276)
regulation of hormone secretion (GO:0046883)
positive regulation of insulin receptor signaling pathway (GO:0046628)
peptide hormone receptor binding (GO:0051428)
negative regulation of acute inflammatory response (GO:0002674)
insulin-like growth factor receptor binding (GO:0005159)
hormone activity (GO:0005179)
genetic imprinting (GO:0071514)
female pregnancy (GO:0007565)
Term
0 1 2 3 4
num_genes
Adjusted.P.value
0.003
0.002
0.001

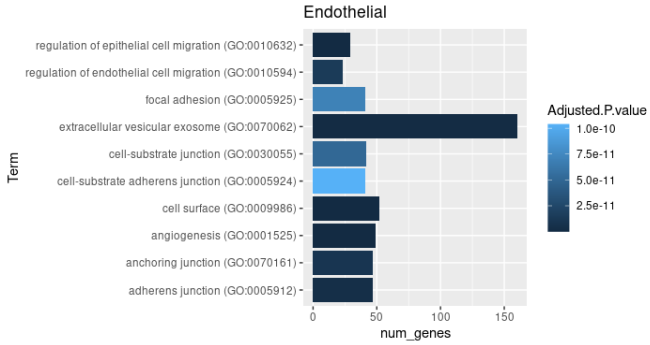

Activated steilate
regulation of peptide secretion (GO:0002791)
regulation of peptide hormone secretion (GO:0090276)
regulation of neuron differentiation (GO:0045664)
regulation of insulin secretion (GO:0050796)
regulation of hormone secretion (GO:0046883)
positive regulation of neurogenesis (GO:0050769)
negative regulation of neuron differentiation (GO:0045665)
negative regulation of neurogenesis (GO:0050768)
negative regulation of nervous system development (GO:0051961)
negative regulation of cell development (GO:0010721)
Term
0 5 10 15
num_genes
Adjusted.P.value
0.010
0.005

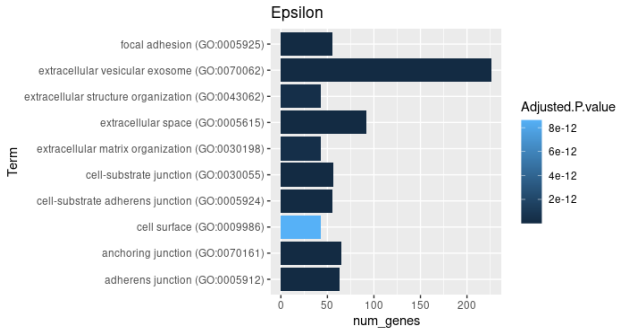

Alpha
respiratory electron transport chain (GO:0022904)
respiratory chain complex I (GO:0045271)
oxidoreductase complex (GO:1990204)
organelle inner membrane (GO:0019866)
NADH dehydrogenase complex (GO:0030964)
mitochondrial respiratory chain complex I (GO:0005747)
mitochondrial membrane part (GO:0044455)
mitochondrial membrane (GO:0031966)
mitochondrial inner membrane (GO:0005743)
generation of precursor metabolites and energy (GO:0006091)
electron transport chain (GO:0022900)
Term
0.0 2.5 5.0 7.5 10.0 12.5
num_genes
Adjusted.P.value
1e-10
5e-11


Beta
vacuole (GO:0005773)
regulation of hormone levels (GO:0010817)
integral component of organelle membrane (GO:0031301)
integral component of endoplasmic reticulum membrane (GO:0030176)
focal adhesion (GO:0005925)
extracellular vesicular exosome (GO:0070062)
cytoplasmic vesicle part (GO:0044433)
cell-substrate junction (GO:0030055)
cell-substrate adherens junction (GO:0005924)
adherens junction (GO:0005912)
Term
0 20 40 60 80
num_genes
Adjusted.P.value
0.003
0.002
0.001


Delta
tetrapyrrole metabolic process (GO:0033013)
serine-type peptidase activity (GO:0008236)
serine-type endopeptidase activity (GO:0004252)
serine hydrolase activity (GO:0017171)
regulation of vasculature development (GO:1901342)
glutathione transferase activity (GO:0004364)
extracellular vesicular exosome (GO:0070062)
extracellular space (GO:0005615)
extracellular region (GO:0005576)
digestion (GO:0007586)
cellular modified amino acid metabolic process (GO:0006575)
angiogenesis (GO:0001525)
Term
0 10 20 30 40 50
num_genes
Adjusted.P.value
4e-04
3e-04
2e-04
1e-04


Ductal
viral transcription (GO:0019083)
translational termination (GO:0006415)
translational initiation (GO:0006413)
translational elongation (GO:0006414)
SRP-dependent cotranslational protein targeting to membrane (GO:0006614)
protein targeting to ER (GO:0045047)
protein localization to endoplasmic reticulum (GO:0070972)
establishment of protein localization to endoplasmic reticulum (GO:0072599)
cotranslational protein targeting to membrane (GO:0006613)
cellular protein complex disassembly (GO:0043624)
Term
0 5 10 15 20 25
num_genes
Adjusted.P.value
3.139588e-18
2.361503e-18
1.583417e-18
8.053314e-19
2.724579e-20


Endothelial
regulation of epithelial cell migration (GO:0010632)
regulation of endothelial cell migration (GO:0010594)
focal adhesion (GO:0005925)
extracellular vesicular exosome (GO:0070062)
cell-substrate junction (GO:0030055)
cell-substrate adherens junction (GO:0005924)
cell surface (GO:0009986)
angiogenesis (GO:0001525)
anchoring junction (GO:0070161)
adherens junction (GO:0005912)
Term
0 50 100 150
num_genes
Adjusted.P.value
1.0e-10
7.5e-11
5.0e-11
2.5e-11


Epsilon
focal adhesion (GO:0005925)
extracellular vesicular exosome (GO:0070062)
extracellular structure organization (GO:0043062)
extracellular space (GO:0005615)
extracellular matrix organization (GO:0030198)
cell-substrate junction (GO:0030055)
cell-substrate adherens junction (GO:0005924)
cell surface (GO:0009986)
anchoring junction (GO:0070161)
adherens junction (GO:0005912)
Term
0 50 100 150 200
num_genes
Adjusted.P.value
8e-12
6e-12
4e-12
2e-12

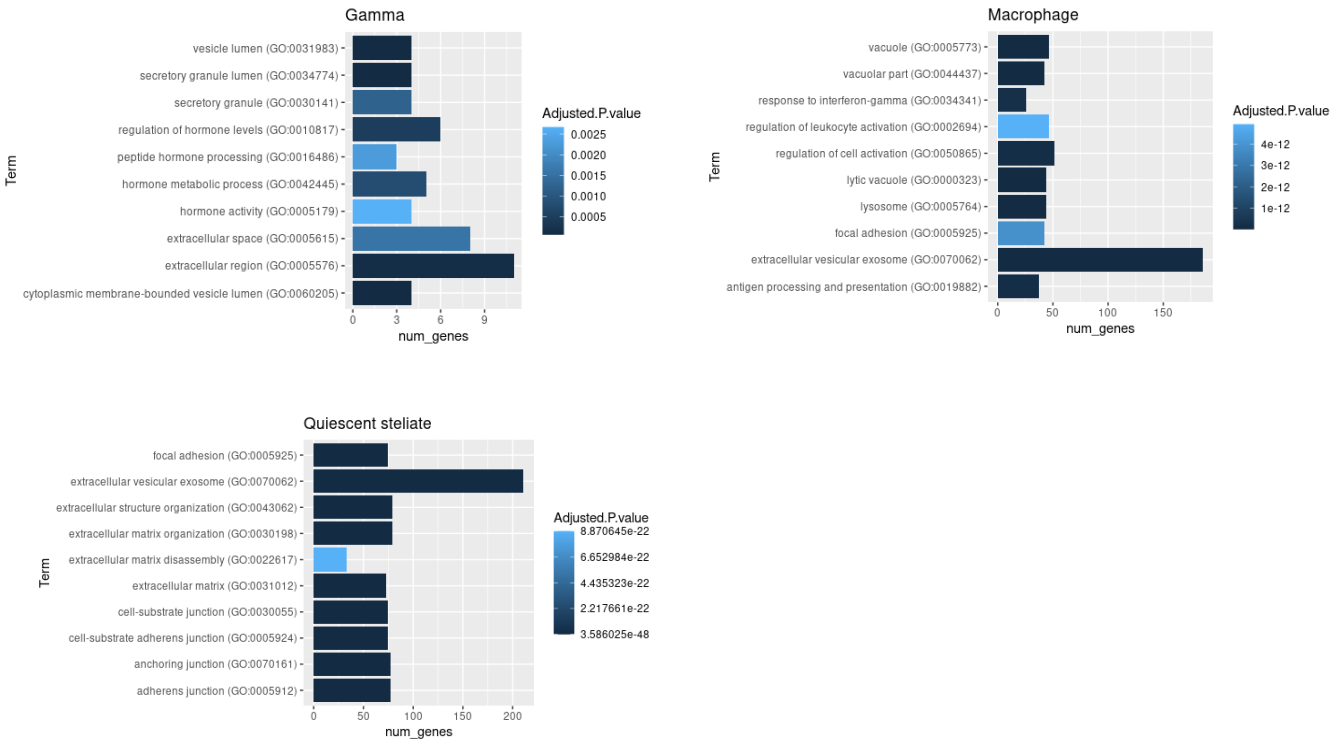


Figure 10: Gene enrichment by cluster

Cell type	Primary function
Acinar	Digestive enzyme production
Activated steilate	Involved in signaling pathways, cell growth pathways
Alpha*	Secrete glucagon
Beta*	Secrete insulin
Delta*	Secrete somatostatin
Ductal	Bicarbonate secretion
Endothelial	Part of pancreatic blood vessels
Epsilon*	Secrete ghrelin
Gamma*	Secrete pancreatic polypeptide
Macrophage	Immune cells
Quiescent stellate	Associated with lipid storage & transport

Table 4. General function of cells found in pancreatic samples, by cluster. Clusters not found in our analysis were not included. Information from Baron, 2016. * denotes a pancreatic islet cell subtype. Islets all contain endocrine cells that secrete hormones involved in glucose homeostasis

A comparison of the known cell function and gene set enrichment data shows significant agreement. Acinar cells, known to be involved in digestive enzyme production, presented enrichment in peptide transport and secretion genes, peptide-hormone receptor binding, and hormone activity, which together suggest that these cells are involved in a complex hormone-regulated digestive enzyme secretion system. Activated stellate cells, which are involved in pancreatic fibrosis and cell growth show enrichment neurogenesis and nervous system development. Baron et. al. identified a subset of activated stellate cells of neural crest origin, which they suggest are pancreatic Schwann cells responding to injury (the same stimulus that would have activated regular stellate cells). No other functions related to cell growth were identified for this group. Alpha cells show strong enrichment of mitochondrial activity and cellular respiration, and given their secretion of glucagon, which involves an increase in glycolysis and mitochondria activity (Quesada, 2008). Beta cells, which are responsible for producing insulin, do not show any indication of that function in their enrichment terms. However, they demonstrate strong enrichment in extracellular vesicular exosome related genes, which have been demonstrated to play a role in Type-1 diabetes (Cianciaruso, 2017). Delta cells demonstrate functional enrichment of extracellular related genes, which may relate to their secretion of somatostatin. Ductal cells demonstrate strong enrichment of translation and protein synthesis, which invites further investigation. Endothelial cells are highly enriched for angiogenesis and vascular development genes (which is expected, given their function as vasculature of the pancreas). Epsilon cells also demonstrate this extracellular and cell junction enrichment. Gamma exhibit enrichment of secretory vesicle and hormone-related gene functions, which suggests their polypeptide secretion function. Macrophages exhibit immune-related functions like leukocyte activation, interferon-gamma response, and antigen processing. Quiescent stellate cells also exhibit enrichment in cell-junction genes.

While a number of these cell types have their known function clearly aligned with their functional enrichment results, some do not. While this could represent some new findings of pancreatic cell function, it could also be an artifact of any upstream processing errors (whether in the process of isolating cells the in lab, or during a computational processing step), or more likely, a bias introduced by filtering criteria. While we chose to filter the differential expression results based on log2 fold change and then adjusted p-value (selecting top ten terms by adjusted p-value), it is possible that the gene functions most classically associated with a certain subtype are tied to genes that are not expressed at a very high level (and would be filtered out by a high log2 fold change threshold, or a low p-value). To attempt to solve this issue, we attempted to filter based only on low adjusted p-value, but the results remained the same.

Conclusion

Ultimately, we were able to replicate the results of the paper to a degree. We successfully identified 11 cell clusters using scRNA-seq data, but were missing some of the immune cells identified by the authors. Gene set enrichment expression analysis confirmed the identity and function of these cells, along with some new findings that may be artifacts of quality filters applied to the results, or might suggest subpopulations of the cell types with more specialized functions, as the authors discovered themselves.

A major problem we encountered was missing a large amount of genes from the Seurat object. We found that this problem dealt with the mapping of gene symbols to the Seurat object dataset. Depending on the mapping approach, certain genes were missing from the data. These issues also caused for the percent mt to be 0 for all the data. After trying multiple mapping approaches, we used Ensembl v. 79 to map the gene symbols. It did not remove half the genes from the dataset and it found the mitochondrial data and percentages. By problem solving together, as the analyst and programmer, we were able to fix this issue. An additional problem we encountered was missing the two clusters Cytotoxic T and Mast cells. Due to limited time we were unable to resolve this issue. If there was additional time the programmer and analyst would have worked together to fix this issue as well.

References

Charlotte Soneson, Michael I. Love, Mark D. Robinson (2015): Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research.

Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A.
Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013; 128(14).

Cianciaruso, C., Phelps, E. A., Pasquier, M., Hamelin, R., Demurtas, D., Alibashe Ahmed, M., Piemonti, L., Hirosue, S., Swartz, M. A., De Palma, M., Hubbell, J. A., & Baekkeskov, S. (2017). Primary Human and Rat β-Cells Release the Intracellular Autoantigens GAD65, IA-2, and Proinsulin in Exosomes Together With Cytokine-Induced Enhancers of Immunity. Diabetes, 66(2), 460–473. https://doi.org/10.2337/db16-0671

Hao and Hao et al. Integrated analysis of multimodal single-cell data. bioRxiv (2020) [Seurat V4]

Johannes Rainer (2017). EnsDb.Hsapiens.v79: Ensembl based annotation package. R package version 2.99.0.

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A.
Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Research. 2016; gkw377.

Quesada, I., Tudurí, E., Ripoll, C., & Nadal, Á. (2008). Physiology of the pancreatic α-cell and glucagon secretion: role in glucose homeostasis and diabetes, Journal of Endocrinology, 199(1), 5-19. Retrieved May 3, 2021, from https://joe.bioscientifica.com/view/journals/joe/199/1/5.xml

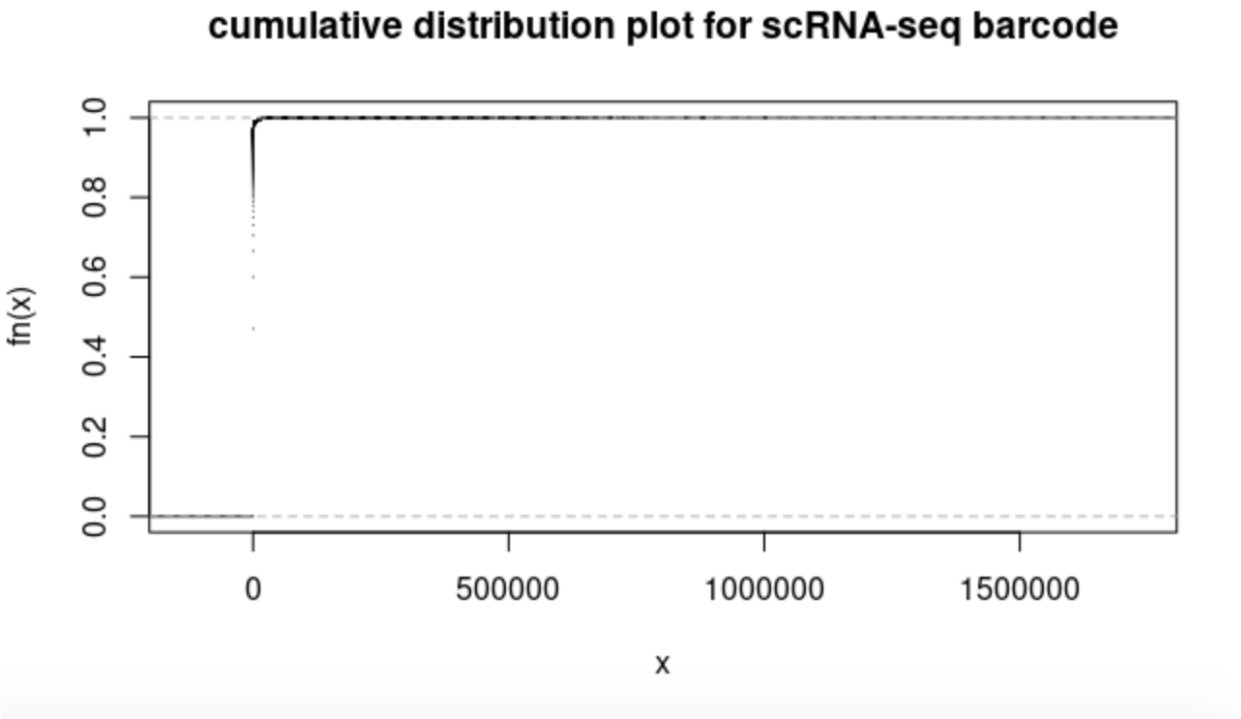
Stuart and Butler et al. Comprehensive integration of single cell data. bioRxiv (2018).

Wang X and Cairns MJ (2014). SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing. Bioinformatics, 30(12):1777-9.

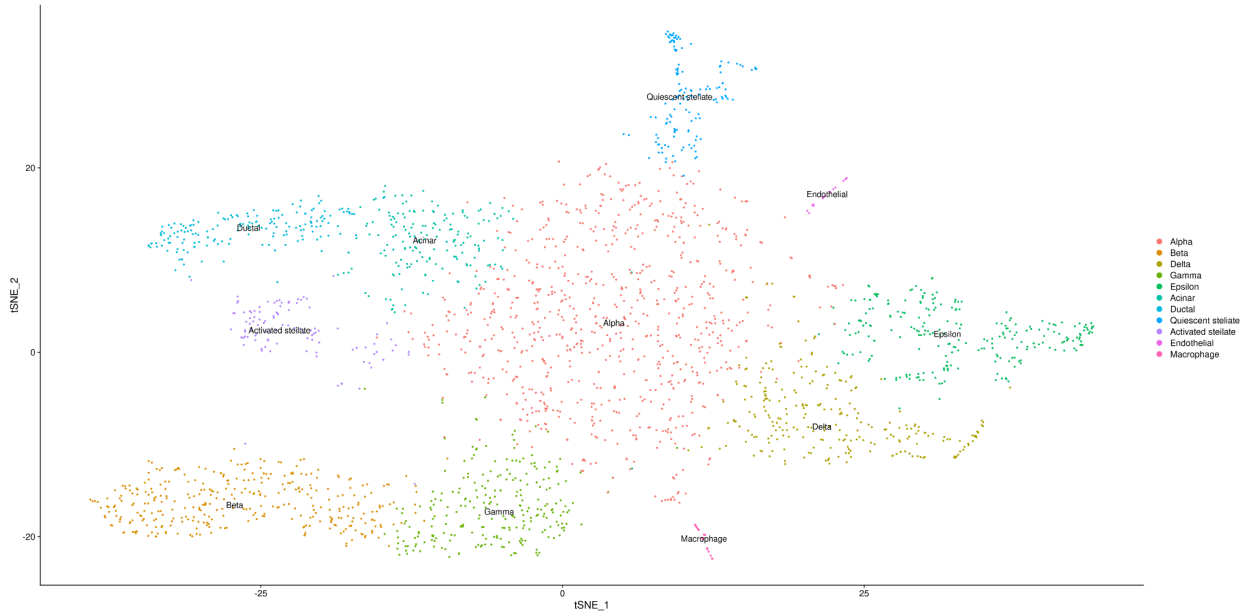
Xie Z, Bailey A, Kuleshov MV, Clarke DJB., Evangelista JE, Jenkins SL, Lachmann A, Wojciechowicz ML, Kropiwnicki E, Jagodnik KM, Jeon M, & Ma'ayan A.
Gene set knowledge discovery with Enrichr. Current Protocols, 1, e90. 2021. doi: 10.1002/cpz1.90

Zhu, A., Srivastava, A., Ibrahim, J.G., Patro, R., Love, M.I. Nonparametric expression analysis using inferential replicate counts Nucleic Acids Research (2019).

Supplementary Data:



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



Supplementary Figure 2: tSNE map displaying relatedness between cell types