Nitsueh Kebere - Data curator
Xinyu Sun - Programmer
Will Mischler - Analyst
Reva Shenwai - Biologist

TA: Dakota Hawkins

Introduction

The pancreas is a vertebrate-specific organ that plays an important role in converting the food we eat into fuel and maintaining general energy homeostasis. There are a variety of disorders of the pancreas including pancreatitis, pancreatic cancer, and Cholangitis (Baron *et al*., 2016). Additionally, the dysfunction of this organ is also clinically linked to type 1 (T1D) and type 2 diabetes (T2D). In order to evaluate pancreatic diseases, several efforts have been made in characterizing the gene expression profiles of cells in the pancreas using different RNA-seq approaches. However, these previous efforts have been limited by the difficulty of analyzing many human cells at a high-throughput scale (Baron *et al*., 2016).

The demand for analyzing human cells at a high-throughput scale has led to the development of many single-cell RNA-seq (sc-RNA-Seq) platforms, one of which is the inDrop sc-RNA-seq platform (Klein *et al*., 2015). This platform uses high-throughput droplet microfluidics that allows the capturing of thousands of cells without pre-sorting. Baron *et al*. (2016) makes use of this platform to provide a detailed look at the gene expression profile of over 12,000 pancreatic cells from four human donors and two mice strains. This approach led to the discovery of 14 cell clusters in human samples and 13 cell clusters in mice samples that matched previously characterized cell type. Additionally, the authors also found subpopulations of ductal and beta cells with distinct expression profiles (Baron *et al*., 2016).

Data

For this study, human islet cells were obtained from The National Disease Research Interchange (NDRI), and mice strains were obtained from Jackson Laboratories. Mice islet cells were isolated by perfusion of the common bile duct, and purified by Histopaque gradient (Sigma) centrifugation (Baron *et al*., 2016). Sample cells were encapsulated into droplets, and 80% of encapsulated cells were barcoded (Baron *et al*., 2016). Library preparation was carried out by using random hexamers after linear amplification, and paired end sequencing was performed on Illumina Hiseq 2500 machine. Reads lacking any of the expected sequences in

read 1 (known cell barcode, adaptor sequence (W1), or beginning of the poly-T tail) were removed (Baron *et al*., 2016).

This project used the sc-RNA-seq data associated with a 51-year-old female donor. There are 3 libraries (SRR3879604, SRR3879605, SRR3879606) associated with this particular sample, and each of these libraries have fastq files generated from a paired-end sequencing run. The barcode for each cell and UMIs for each library was provided in a convenient format, where raw read 1 barcode for each library has already been processed.

For each library, the number of reads per distinct barcode was calculated, and a cumulative distribution plot was generated (Fig. 1) to see how the reads are distributed among barcodes. Figure 1 shows that greater than 95% of the barcodes have read numbers (counts) less than 1000 across all three libraries. Based on this distribution plot, a read threshold was set to filter out infrequent barcodes. Barcodes with read number (count) over two standard deviations of the mean count were used to generate the whitelist barcodes, and were kept for further analysis.

To generate the UMI count matrix for each library, salmon alevin 1.2.0 (Srivastava, A. *et al.*, 2019) was used. Before running salmon, an index file containing the salmon index of the reference transcriptome (Release 33 (GRCh38.p13)), and a transcript to gene map file that contains mapping of each transcript were created. Salmon was then run using the ISR library with the fastq files and the whitelist barcodes for each library. The UMI count matrix generated for each of the libraries was merged for further downstream analysis.

From the salmon statistics, all three libraries showed a low mapping rate (Table 1). The mapping rate for each of the three libraries is much lower than expected, since many RNA-seq experiments have mapping rates greater than 50%. This low mapping rate could be due to the short read length (~ 55 bp) across all three libraries.

Important sample and library information from the salmon statistics is shown in Table 1&2.

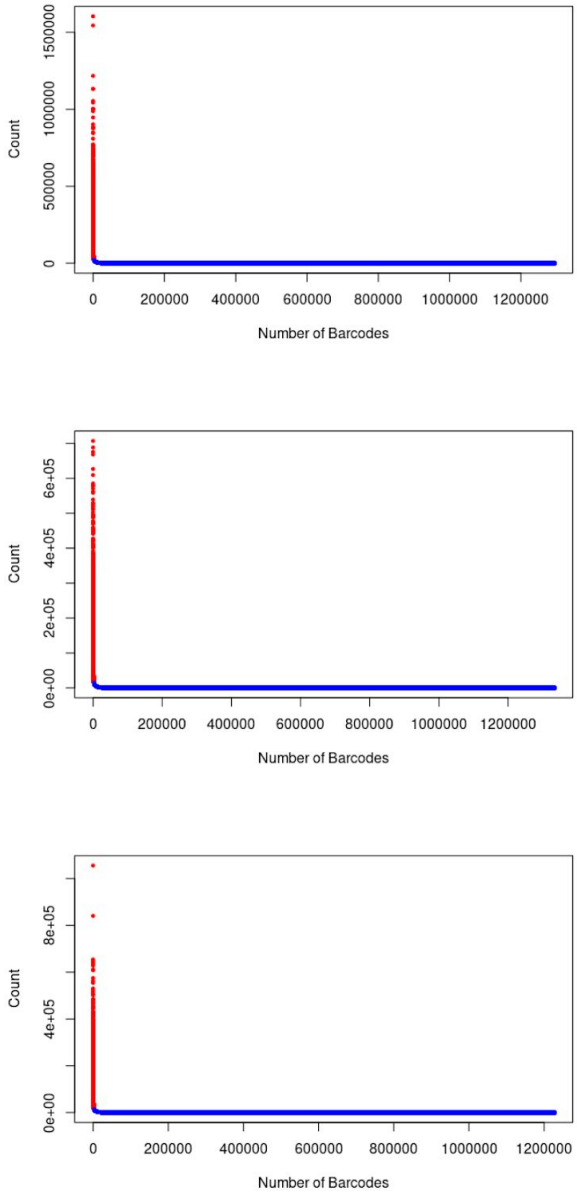


Figure 1. Cumulative Distribution Plot. Distribution shows how reads are distributed among barcodes for each of the libraries (SRR3879604, SRR3879605, SRR3879606). About >95% of the barcodes have count less than 1000 across all three libraries. Threshold= mean(count) + 2*SD(count). Red dots show barcodes that have count greater than the threshold (kept for further analysis). Blue dots show barcodes that have count less than the threshold (not kept for further analysis).

Table 1. Library information from salmon statistics. The table reports informative statistics generated by salmon.

	Library 1	**Library 2**	**Library 3**
Total Reads	564226059	392517,479	368094423
Used Reads	341839633	234889698	215478094
Total Barcodes	1946820	1968550	1854940
Used Barcodes	33839	34243	32676
Whitelist Informative Barcodes	2454	2955	2864
Mapping Rate (%)	33.92	29.36	29.75

Table 2. Overall sample information after merging UMI count matrix for three of the libraries. The information in this table is generated by salmon.

Total Number of Cells	8240
Mean Number of UMIs per Cell	1650
Mean Number of Genes per Cell	850
Average Mapping Rate (%)	31.01

Methods

The UMI matrix generated had Ensembl gene identifiers (ENSG ID) as row name. It is often more helpful to convert them to gene symbols, thus ENSG IDs were converted to gene symbols using the biomaRt package in R. When the conversion was done, there were some ENSG IDs mapped to the same gene symbols, so there were 60194 genes remaining.

The count matrix with gene symbols was used to create a Seurat object, and low quality cells were filtered by excluding unique feature counts over 2,500 or less than 200. We selected

this cutoff to minimize the effect of doublets and empty droplets respectively. After filtering, there were 26864 genes and 4178 cells remaining (Table 3). In addition, we filtered genes that appeared in less than 3 cells, so that only representative genes which are present in multiple cells can be included in further analysis.

The Seurat object was then filtered out for low variance genes. In order to include only highly variable features that are likely to be informative, we selected top 2000 features, by using "vst" in FindVariableFeatures function. "Vst" method first fits a curve to predict the variance of each gene as a function of its mean, by calculating a local fitting of polynomials of degree 2. Then, it standardizes the feature values by subtracting the observed mean and expected variance and then dividing by expected standard deviation of a feature from the fitting. Top 2000 feature variance is then selected (Stuart, 2018). Vst method accounts for the mean-variance relationship that is inherent to single cell RNA-seq, so its highly variable feature selections are more reliable.

PCA algorithms and linear dimensionality reduction methods were used to maximize the variance in the dataset using the previously determined 2000 highly variable feature. An elbow plot was then generated to determine the number of principal component vectors used in the clustering. The elbow plot generated from our dataset showed an "elbow" around PC10, so the first 10 PCs were used in clustering (Fig. 4).

A KNN graph was constructed based on the euclidean distance in PCA space, and the edge weights between any two cells were refined based on Jaccard similarity. To cluster cells, Louvain algorithm was applied to iteratively group cells together. During clustering, resolution was set to 0.6, and 14 cell clusters were obtained. The detailed distribution of those clusters is shown in the pie chart (Fig. 5).

Based on the obtained clusters, marker genes were identified using the Seurat package differential expression analysis, which uses a Wilcoxon Rank Sum test by default. Other tests were performed, but the Wilcoxon test gave the most accurate data to recreate the study. Specific marker genes from Baron *et al*. (2016) were then applied to the clusters in order to assign cell types to each cluster. PanglaoDB (Franzén *et al*., 2019) was used to identify clusters that were not assigned a cell type through the previous list of marker genes. A heatmap and tSNE plot were created using the Seurat package as well.

Table 3. Table reports the number of genes and cells at each stage of filtering. Eventually, about half of all genes and cells were filtered out. The table reports the number of genes and cells at each stage of filtering.

Filtering Stage	Genes	Cells
Original UMI matrix	60233	8240
Converting ENSG to gene symbol	60194	8240
Filtering out low-quality cells	26864	5214
Filter out low variance genes	26864	4178

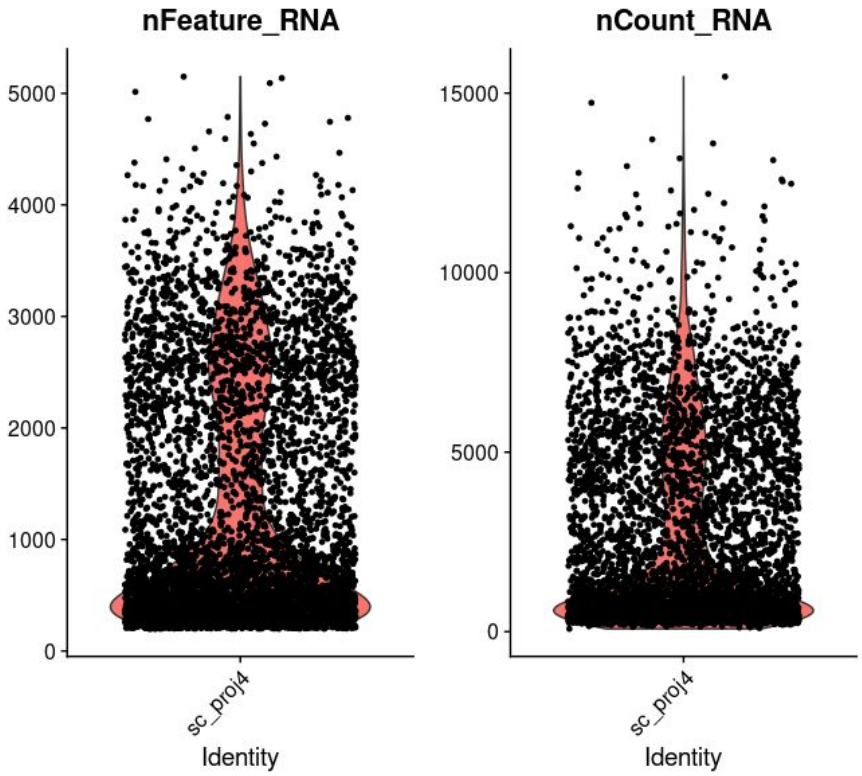


Figure 2. QC metrics as a violin plot. nFeature_RNA is the number of genes detected in each cell. nCount_RNA is the total number of molecules detected within a cell

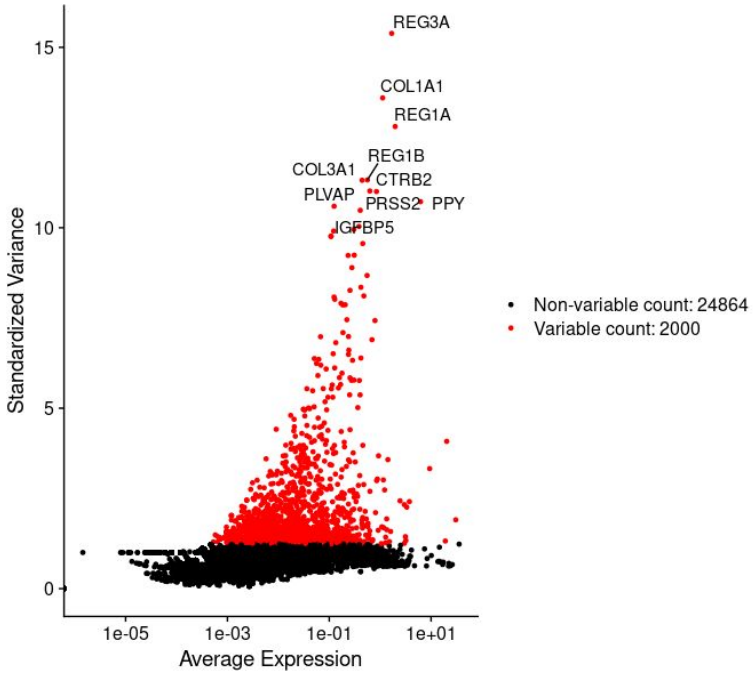


Figure 3. 2000 variable features are highlighted red in the plot. Among 26864 genes, the top 2000 variable features are selected for downstream analysis. Top 10 variable features are labeled.

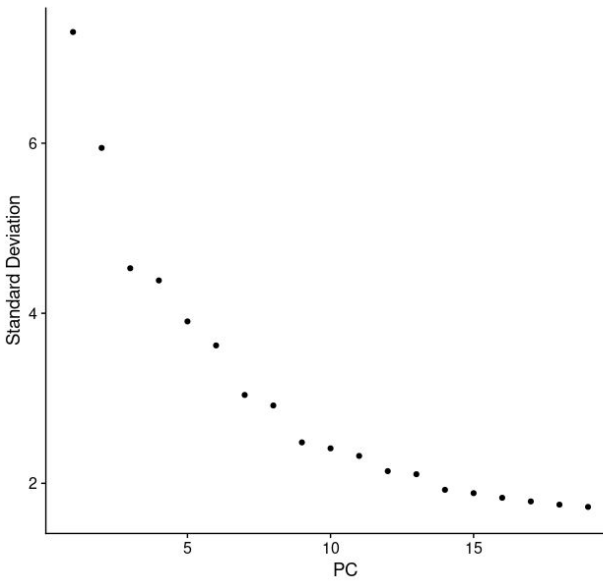


Figure 4. The plot becomes flattened around PC 10, and an 'elbow' is observed around PC9-10 . A ranking of principal components based on the percentage of variance explained by each component.

Results

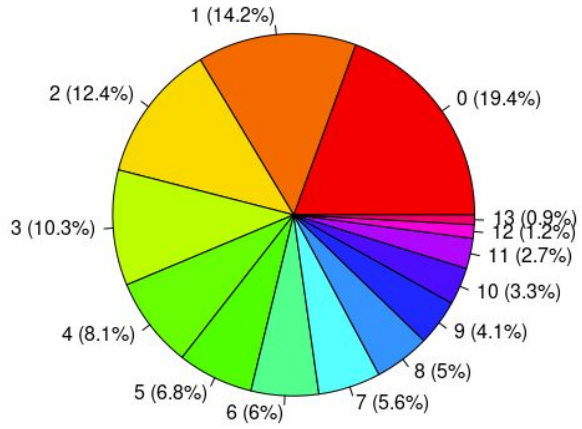
After filtering for low quality cells and low variance genes, which is explained in the previous section, 14 clusters were identified. Among those clusters, clusters from 0-5 had more cells than the other clusters. However, clusters from 8-13 were less abundant than the other clusters (Fig. 5).

Of the clusters identified, all 14 were assigned cell types that coincided with cell types identified in the Baron *et al.* (2016) paper. The assigned cell types are shown in Table 4 with the cluster they were assigned to and their marker genes. The top 10 marker genes for each cluster are illustrated in a heatmap (Fig. 6). This shows the expression of each marker gene across all of the different clusters. Almost all clusters show a high difference in expression level for their marker genes except for clusters 2, 3 and 4. The tSNE plot shows the dimensionality of the clusters and separation between different cell types (Fig. 7). This plot does not match the tSNE plot from Baron *et al*. (2016) (Fig. 1D). This could be due to differences in clustering methods and because there was less data involved in the clustering for this project versus the clustering for Baron *et al*. (2016).

Table 4. Cell types and marker genes for each cluster. Cell types were assigned based on marker gene expression. Marker genes were chosen based on Baron *et al*. (2016) Table S2.

Cluster label	Cell Type	Marker Gene
0	Delta	SST
1, 8	Beta	INS, ADCYAP1
2,3,4	Alpha	GCG
5	T Cells	RPS6KA5
6	Ductal	KRT19
7	Gamma	PPY
9, 11	Acinar	REG1B, CPA1
10	Stellate	PDGFRB
12	Macrophage	CD68
13	Vascular	PECAM1

A



B

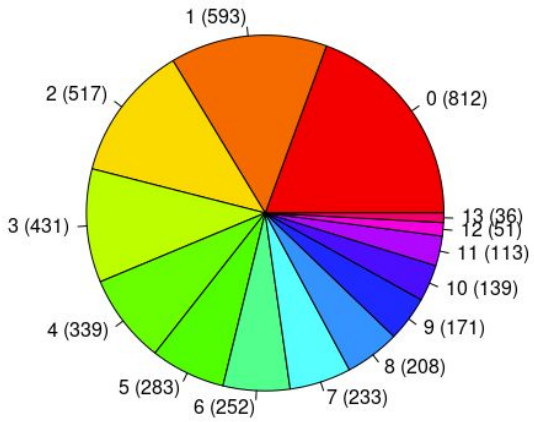


Figure 5. Number and Percentage of cells in each of 14 clusters. (A) Shows the percentage of total cells in each cluster. (B) shows the number of cells in each cluster.

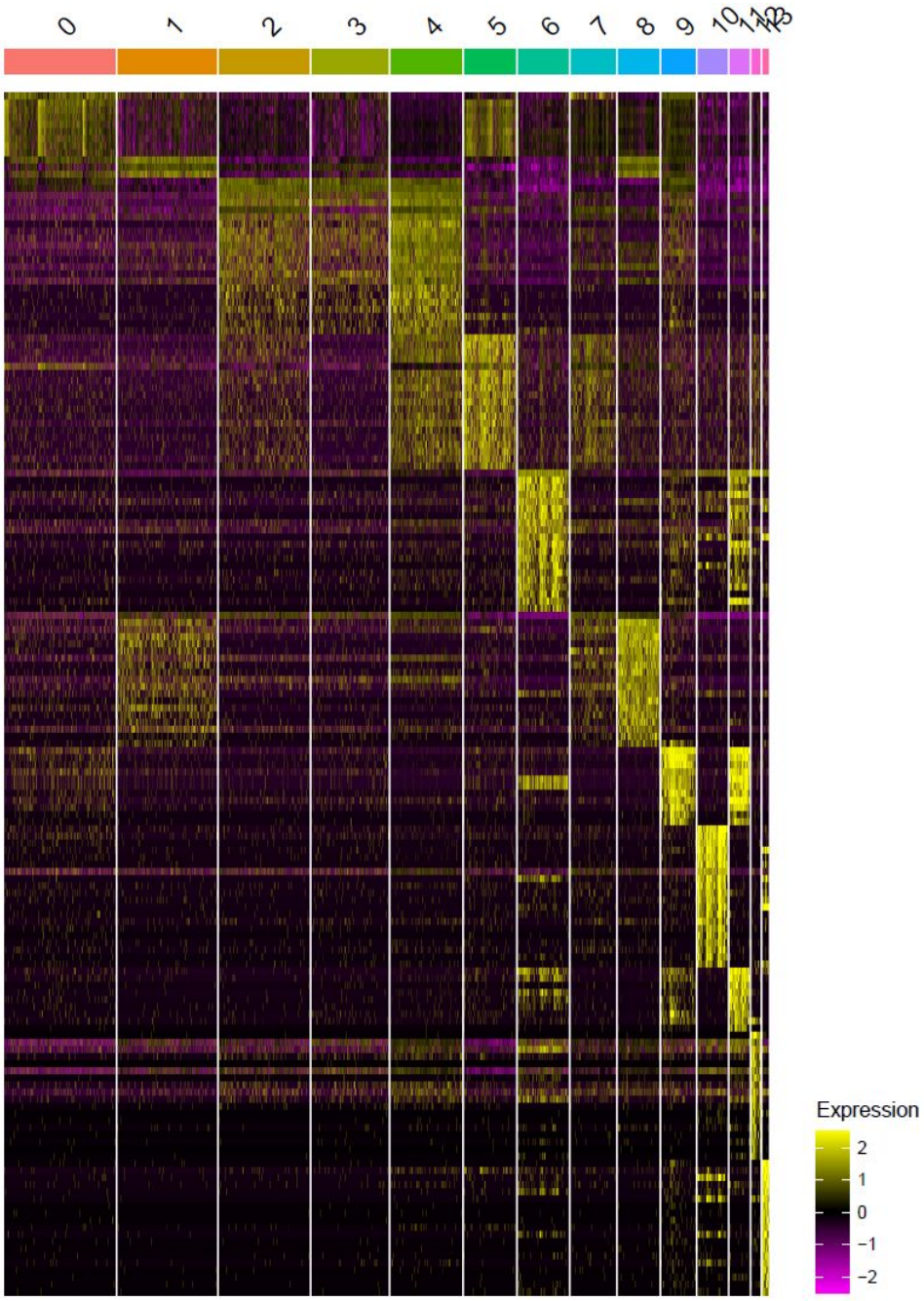


Figure 6. Heatmap of the top 10 marker genes for each cluster. Each row shows the expression level of one marker gene and the columns are grouped by their respective clusters. Yellow color indicates a higher level of expression, while purple shows a decrease in expression level.

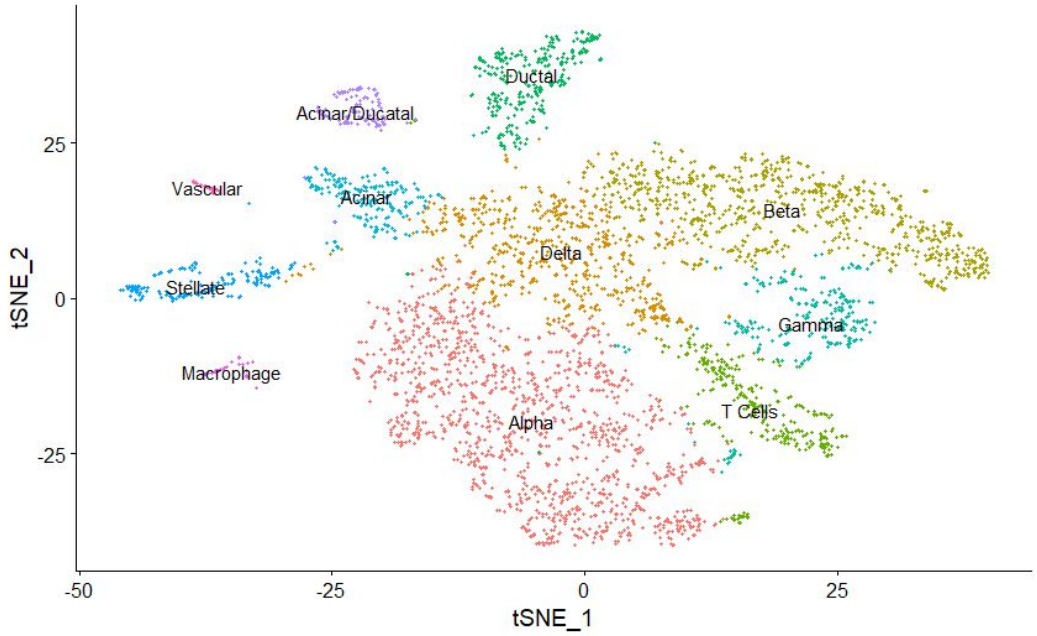


Figure 7. Dimensionality of cell clusters using tSNE plot. Clusters labeled by specific cell type assigned. Cell types were assigned using marker genes from Baron *et al*. (2016).

Gene set enrichment was performed on the obtained clusters using DAVID (Huang *et al*., 2009), enrichR (Chen *et al.*, 2013; Kuleshov, 2016) and metascape (Zhou *et al.*, 2019). Marker genes were filtered by adjusted p-value of less than 0.05, and then the top 4 or 5 genes were generally used for analysis. All significant genes were used for analysis in cluster 0, since it had a large number of mitochondrial genes, which yielded fewer results through most of the abovementioned enrichment analysis tools. About 8 marker genes were considered for cluster 10 since these were all equally significant with an adjusted p-value of 0. The gene set enrichment results are summarized (Table 5), along with gene ontology (GO) terms and proposed cell types. Only GO terms with adjusted p-value of less than 0.05 were included (Table 5).

Table 5. Summary of proposed cell types for each cluster and marker genes used for gene enrichment analysis. We proposed 10 cell types from our 14 obtained clusters, including 4 endocrine cells (alpha, beta, delta, gamma), both exocrine cells (ductal, acinar), 2 immune cells (macrophage, T cells), stellate cells and vascular cells.

Cluster name	Label	Marker genes	Geneset enrichment terms
Delta	0	SST, MT.ND4, MT.ND1, MT.CYB, MT.ND2, MT.CO1, MT.ATP6, MT.CO2, MT.CO3, INS, MT.ND3, MT.RNR2, IAPP, MT.ND5, MTATP6P1, MT.RNR1, H4C3	➔ Hormone ➔ Hormone activity
Beta	1	INS, GNAS, IAPP, HADH	➔ Insulin secretion ➔ Gene expression regulation in pancreatic beta cells ➔ Pancreatic beta-cell development regulation ➔ Beta-Cell Death in Diabetes Mellitus Type2 ➔ Insulin Synthesis in beta-Cell
Alpha	2	TTR, GCG, CLU, TM4SF4, MALAT1	➔ Pancreatic islet ➔ Secretory granule lumen (GO:0034774) ➔ Maturity-onset diabetes of the young, type 3 (disorder) ➔ Autonomic neuropathy
Alpha	3	TTR, GCG, TM4SF4, SERPINA1, CRYBA2	➔ Alpha cell ➔ Pancreatic islet ➔ Maturity-onset diabetes of the young, type3 (disorder) ➔ Reactive hypoglycemia
Alpha	4	GC, IRX2, FXYD3, FXYD5, EGFL7	---
T Cells	5	RPS6KA5, ACER3, GPR82, SLC35E3	➔ Macrophage bone marrow 0hr (mouse)
Ductal	6	TACSTD2, KRT7, KRT19, TINAGL1, CFTR	➔ Gastric epithelial cell ➔ Gastric tissue (bulk)
Gamma	7	DPYSL3, AQP3, CASR, RBP4, PEG10	---
Beta	8	ADCYAP1, MAFA, SAMD11, PDX1,GLP1R	➔ Pancreatic islet ➔ Maturity onset diabetes mellitus in young ➔ Peptide hormone secretion (GO:0030072) ➔ Gene expression regulation in pancreatic

			beta cells ➔ Hyperglycemia
Acinar	9	REG1B, CTRB2, CELA3A, CTRB1, REG1A	➔ Pancreatic secretion ➔ Pancreatic juice ➔ Pancreatitis ➔ PancreaticIslet
Stellate	10	COL3A1, COL1A2, COL6A2, SPARC, PDGFRB, COL5A1, COL15A1, FMOD	➔ SNU1105 Central nervous system TenPx31 ➔ A172 Central nervous system TenPx24 ➔ U118MG Central nervous system TenPx26 ➔ Brain SF-539 BTO:0004214 SF539 NCI60 ➔ Brain SNB-75 BTO:0004222 P001561
Acinar	11	ALDOB, ALB, RARRES2, ANPEP, DUOXA2	➔ Gastric tissue (bulk) ➔ Pancreatic cancer DOID-1793 human GSE16515 sample 475
Macrophage	12	LAPTM5, TYROBP, FCER1G, RUNX3	➔ Microglia Pathogen Phagocytosis Pathway WP3937 ➔ THP1 Haematopoietic and lymphoid tissue TenPx24 ➔ Natural Killer Cell Activation through ITAM-containing Receptors ➔ MP:0008042 abnormal NK T cell physiology ➔ Natural killer cells ➔ Lymphoblastoid BTO:0000773 X131.126 HM48.GM07346
Vascular	13	PLVAP, F2RL3, ESM1, RGCC, VWF, ACVRL1, SEMA3F, ROBO4, KDR	➔ Blood coagulation Homo sapiens P00011 ➔ regulation of blood vessel endothelial cell migration (GO:0043535) ➔ Vasculature ➔ Thrombospondin complex ➔ platelet-derived growth factor complex

Discussion

We attempted to recreate the single cell RNA-seq analysis conducted by Baron *et al*. (2016) to verify the cell types they identified. Overall, we were able to reproduce several, but not all components of their results. As shown in Table 5, through our clustering analysis we found 10 of the 15 cell types identified by Baron *et al*. (2016). These include 4 endocrine cells (alpha, beta, delta, gamma), both exocrine cells (ductal, acinar), 2 immune cells (macrophage, T cells), stellate cells, and vascular cells. We did not have sufficient information to conclude whether the

stellate cell type identified was quiescent or activated. The gene enrichment analysis and resulting GO terms allowed us to be reasonably confident in our findings for clusters 1-3 and 8-12. These include both beta cell clusters (1,8), two (2,3) of the three alpha cell clusters, one of the two acinar cell clusters (9, 11), and the stellate cell (10), macrophage (12) and vascular cell clusters (13).

We were able to find limited gene enrichment associated with clusters 0, 6 and 11, but the GO terms provided some support for the assigned cell types. GO terms for cluster 0 indicated hormone activity, which is reasonable given that delta cells release the hormone somatostatin (Rorsman & Huising, 2018). GO terms for cluster 6 indicated gastric epithelial tissue, which is also reasonable given that ductal cells form the epithelial lining of the pancreatic duct for delivery of enzymes to the duodenum (Grapin-Botton, 2005). Additionally, GO terms for cluster 11 indicated both pancreatic and gastric tissue, which is also reasonable given that acinar cells release enzymes that are transported to the duodenum to aid in digestion (Williams, 2010). The lower number of gene enrichment terms make us less confident in our assigned cell types, but similar results obtained from PanglaoDB (Franzén *et al*., 2019) reinstate confidence in our assigned cell types. Cluster 5 also produced limited gene enrichment, although the GO terms indicated similarity to mouse immune cells, which supports our cell type assignment of T cells.

We were unable to find significant gene enrichment for clusters 4 and 7 to support the cell type assignments. For these clusters, we used the marker genes reported in the paper, which may not be the most accurate way to determine cell type. PanglaoDB (Franzén *et al*., 2019) reports different marker genes for each cell type listed and reports that there may be some overlap in some of the marker genes used. This could lead to the incorrect assignment of cell type to cluster, and therefore a lack of enrichment found in the pathways for that cell type.

The lower gene enrichment results may be explained by incorrect labeling of cells, errors in clustering or cell labelling. We were unable to find four cell types that were identified in the Baron *et al*. (2016) study, namely an islet cell type (epsilon cells), two immune cell types (mast and B cells), and Schwann cells. Unlike Baron *et al*. (2016), we were also unable to determine whether our stellate cells were quiescent or activated types, due to there being no marker gene to determine that. Our analysis produced one less cluster than the Baron *et al*. (2016) study, which can explain some of the discrepancies. We found 14 clusters, while Baron *et al*. (2016) found 15. This can also be explained by the difference in data used to cluster the cell types. We used one sample while Baron *et al*. (2016) had multiple samples to draw data from.

In conclusion, the analysis conducted by Baron *et al*. (2016) was largely reproducible, and we successfully identified 10 of the 15 cell types from their original single cell RNA-seq analysis. The discrepancies in our results are likely explained by errors in our cell labeling and clustering steps, or since we used a subset of their data for our analysis. Repeating this analysis with the complete dataset from the Baron *et al*. (2016) study would prove more useful in measuring the reproducibility of their results.

References

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S., Klein, A. M., Melton, D. A., & Yanai, I. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell systems, 3(4):346–360.e4. https://doi.org/10.1016/j.cels.2016.08.011

Chen E.Y., Tan C.M., Kou Y., Duan Q., Wang Z., Meirelles G.V., Clark N.R., Ma'ayan A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics, 128(14).

Franzén, O., Gan, L.M., Björkegren, J.L.M. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data, Database, 2019:baz046, doi:10.1093/database/baz046

Grapin-Botton, A. (2005). Ductal Cells of the Pancreas. The International Journal of Biochemistry & Cell Biology, 37(3):504–510. doi:10.1016/j.biocel.2004.07.010

Huang D.W., Sherman B.T., Lempicki R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc, 4(1):44-57.

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

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., & Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell, 161(5):1187–1201. https://doi.org/10.1016/j.cell.2015.04.044

Kuleshov M.V., Jones M.R., Rouillard A.D., Fernandez N.F., Duan Q., Wang Z., Koplev S., Jenkins S.L., Jagodnik K.M., Lachmann A., McDermott M.G., Monteiro C.D., Gundersen G.W., Ma'ayan A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Research, gkw377.

Rorsman, P., & Huising, M. O. (2018). The somatostatin-secreting pancreatic δ-cell in health and disease. Nature reviews. Endocrinology, 14(7), 404–414. https://doi.org/10.1038/s41574-018-0020-6

Srivastava, A., Malik, L., Smith, T. *et al.* Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biol* 20, 65 (2019). https://doi.org/10.1186/s13059-019-1670-y

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W.M., Stoeckius, M., Smibert, P., Satija R. (2018). Comprehensive integration of single cell data. bioRxiv 460147. doi: https://doi.org/10.1101/460147

Williams J. A. (2010). Regulation of acinar cell function in the pancreas. Current opinion in gastroenterology, 26(5), 478–483. https://doi.org/10.1097/MOG.0b013e32833d11c6

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., Benner, C., & Chanda, S. K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature communications, 10(1), 1523. https://doi.org/10.1038/s41467-019-09234-6