Project 4 - Single Cell RNA-Seq Analysis of Pancreatic Cells; replicating Baron et al. 2016

Group members: Emily Hughes, Simran Makwana, Sumiti Sandhu, and Michiel Smit
TA: Nick

Introduction

Pancreatic tissue function is characterised by the interaction of several different cell types. Identifying gene expression profiles for each of these distinct types is useful for further analysis of cell subpopulation analysis, but largely has not been completed on a granular level. Baron et al.[1] addressed these issues with their classification of single cell pancreatic samples. These researchers achieved this high resolution analysis of droplet-based single-cell RNAseq data to form cell clusters using tSNE plots. This study aims to build on the results from Baron et al. to classify different cell clusters from sequencing data of human pancreatic cells based on biological and computational analysis.

Data

For sample preparation, the cells were encapsulated into droplets on ice using inDrop platform and lysed in the 4nL microfluidic droplets using a final concentration of 0.4% NP-40. These single-cell lysates were then subjected to reverse transcription at 50°C without purification of RNA.[1] The inDrop platform, which makes use of the CEL-Seq protocol for library construction was used for cell barcoding.[3] Library preparation was carried out according to the protocol mentioned in Klein et al. with minor changes.[2] For example, random hexamers were used during the second reverse transcription step after linear amplification, eliminating the need for primer ligation. Paired-end sequencing on libraries was performed on Illumina Hiseq 2500 and single-cell RNA-Seq data was generated using the GRCh38 reference genome. The FASTQ files were acquired from the Sequence Read Archive (SRA).

Three runs (SRR files) were used from a single human pancreatic islet RNA-Seq sample belonging to a 51-year-old female donor with a BMI of 21.1. The average library size was 441,612,654 reads and the paired-end reads in the libraries had an average length of 204 bases. The number of reads per distinct barcode was counted for each run and the infrequent barcodes were eliminated by observing their cumulative frequency. This step was crucial because the probability of two cells having the same barcode is low and some barcodes might have been incorporated because of deletion errors. Thus, barcodes with ~99.8% frequency were eliminated, which resulted in a barcode whitelist. The reads were aligned and trimmed using Trimmomatic (v0.32)[4] to eliminate reads without known cell barcode, adaptor sequence (W1), or beginning of the poly-T tail.

Salmon (v1.1.0)[5] was used for sample mapping and quantification. This index and the transcript to gene map files were created using the current Gencode human reference transcriptome.[6] Alevin, a tool integrated with salmon, was run on FASTQ files of each library using the barcodes from read1 and filtered read2. The whitelisted barcodes were provided as a parameter to alevin to generate a cell-by-gene count matrix.

Methods

The initial UMI counts matrix consisted of 60233 features across 2452 samples. It was loaded into R using tximport (v1.14.2)[7] and then converted into a Seurat object using the Seurat package (v3.1.5)[8]. Genes that were detected in less than three cells or cells that were detected in less than 200 features were filtered out to remove low quality data. The filtered data then consisted of 23330 features acoss 2415 samples.

The percent of reads that map to the mitochondrial genome were calculated as a measure of cell quality and visualized with the number of features per gene in Figure 1. Cells with fewer than 200 genes were filtered out, as having too few genes may indicate a low quality cell or an empty droplet. The cutoff of 200 was chosen based on the distribution of features in Figure 1A; this would include cells in the main body of the distribution and disregard the mass of cells with very few genes. Additionally, cells with greater than 4000 features were filtered out as these may represent doublets or multiplets. Again, this cutoff was chosen to exclude those with extremely high numbers of genes based on the distribution seen in Figure 1A, since 4000 genes appears to be the upper-bound threshold for the main body of cells. Finally, cells with a mitochondrial percentage greater than five were removed as they may be low quality or dying cells. As seen in Figure 1B, there were not many cells that had an extremely high mitochondrial percentage; three cells were filtered out during this step. These filtering measures reduced the cell count to 1541 cells.

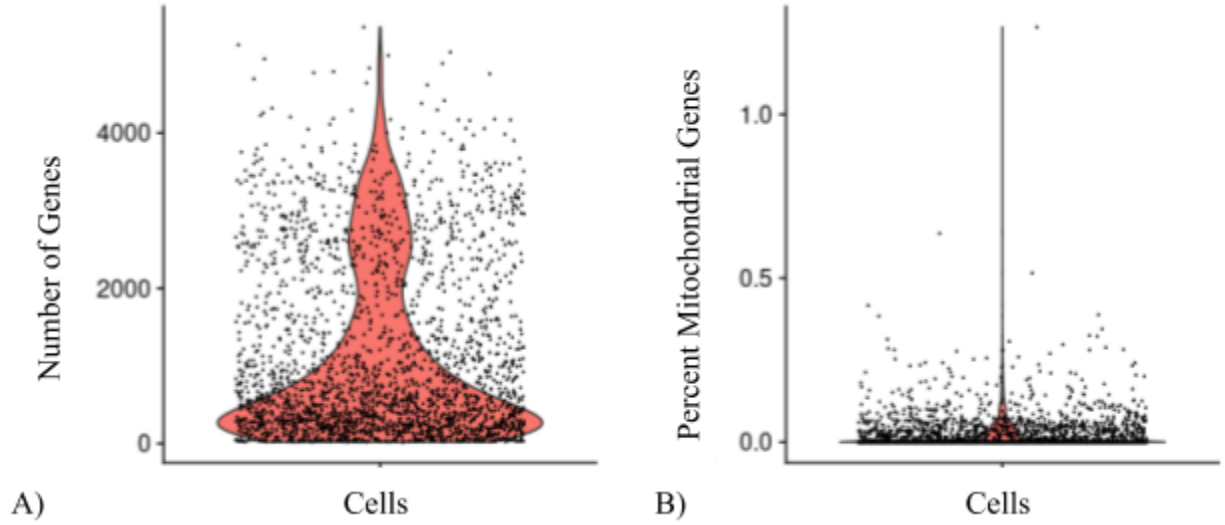

Figure 1. Visualization of quality control measures **A)** number of genes per cell and **B)** percent of mitochondrial genes per cell. The distribution of each plot informed selection of thresholds for filtering based on number of genes and percent mitochondrial content.

The resulting data was normalized using a log transformation across all samples and scaled by a factor of 10,000. The top ten percent of highly variable features were kept for further analysis and all other features were removed.

After filtering, linear dimension reduction using Principal Component Analysis (PCA) was applied to the data. An elbow plot comparing the resulting principal components to the percentage of variance explained by each one was plotted (Figure 2) to qualitatively determine the number of principal components that should be used for clustering using the Louvain algorithm.

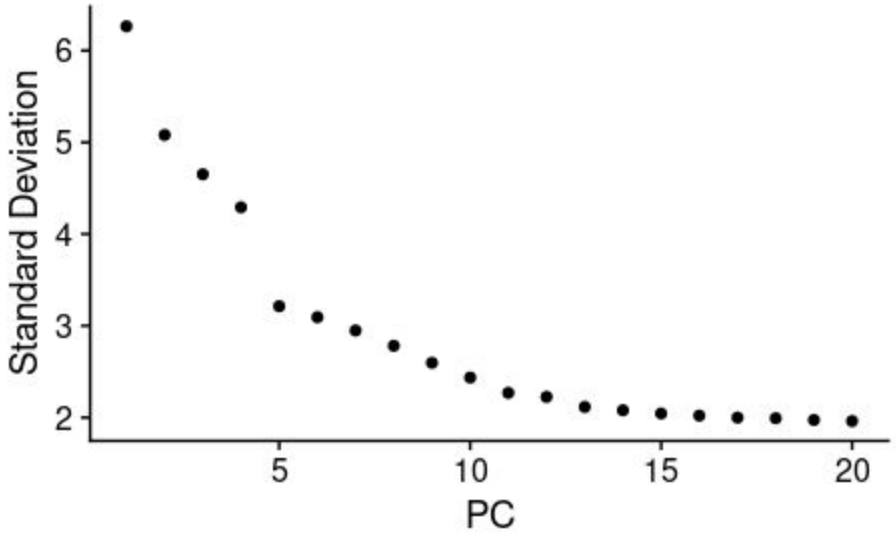

Figure 2. Elbow plot comparing principal components to their respective percentage of variance.

There is a noticeable decrease in standard deviation from principal component four (PC4) to principle component five (PC5), followed by a relatively steady decrease in deviation for the remaining PCs (Figure 2). This indicates that adding PC5 to the analysis would provide additional information to the clustering algorithm. However, adding the following PCs to the analysis would not necessarily add substantial information about the variance of the cell populations. In an attempt to replicate the number of cell clusters found by Baron et al., a resolution of 1.7 was chosen for the Louvain algorithm. Using the first five PC dimensions, sixteen clusters were identified (Figure 3).

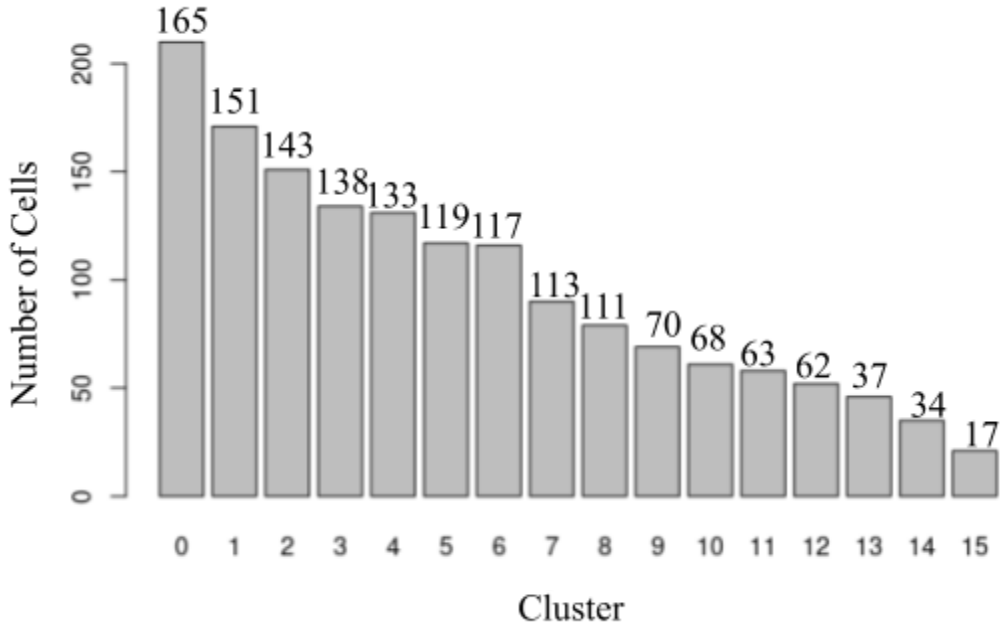

Figure 3. Histogram of identified clusters from Louvain algorithm with number of cells in each.

Using the Seurat package[8], the FindAllMarkers function was used to automate the process of finding the differentially expressed features, or cluster biomarkers. The function identifies markers of each cluster by performing a differential expression test for the given cluster against the average expression of all other clusters. A total of sixteen clusters were found and labeled 0 to 15. The expression of cluster biomarkers were compared between the given cluster and all other clusters; clusters with a log fold change greater than 0.3 and minimum percentage of 10% of cells expressing the genes were included. This resulted in clusters 10 and 14 being omitted from the data.

In order to label clusters as a cell type based on marker genes, a tab-delimited file from PanglaoDB database[9] was used as reference. If the cell type was reported in Baron et al. and the Bonferroni-adjusted p-value was significant (< 0.05), then that cell type was used as the label. Often, the cell types reported in Baron et al. were not found in this dataset. Therefore, the most significant result from PanglaoDB was used for the label. The genes of two of the clusters did not match any marker genes from the database and were thus labeled “Not found”. There were also cases where multiple clusters had the same cell type label. After combining clusters with the same label, eleven clusters were found, including a group that was labelled “Not found”.

To independently validate the computationally obtained cluster labels, the cell types of the clusters were also determined based on the associated biology. Gene Ontology (GO) terms associated with the significant marker genes ($p < 0.05$) were obtained for each cluster using the Enrichr database[10, 11]. The clusters were then labeled as a cell type based on connections between the GO terms and the biology of the cells discussed in Baron et al. An additional analysis was conducted by researching marker genes that had been previously associated with each cell type using the CellMarker database[12]. These marker genes were manually compared to the marker genes of the clusters to determine any matches. It is important to note that this database did not contain information for mast, stellate, gamma or epsilon cells, which were identified in Baron et al.

Results

Figure 4 shows a scatterplot of the marker genes labeled according to the cell types using UMAP. UMAP is a nonlinear dimension reduction technique used to visualize similar cells in low-dimensional space. Baron et al. conducted a similar dimension reduction analysis, which can be seen in their Figure 1D. However, Figure 4 differs greatly from the original analysis because the cell types and marker genes used in this analysis were not consistent with those of the original study.

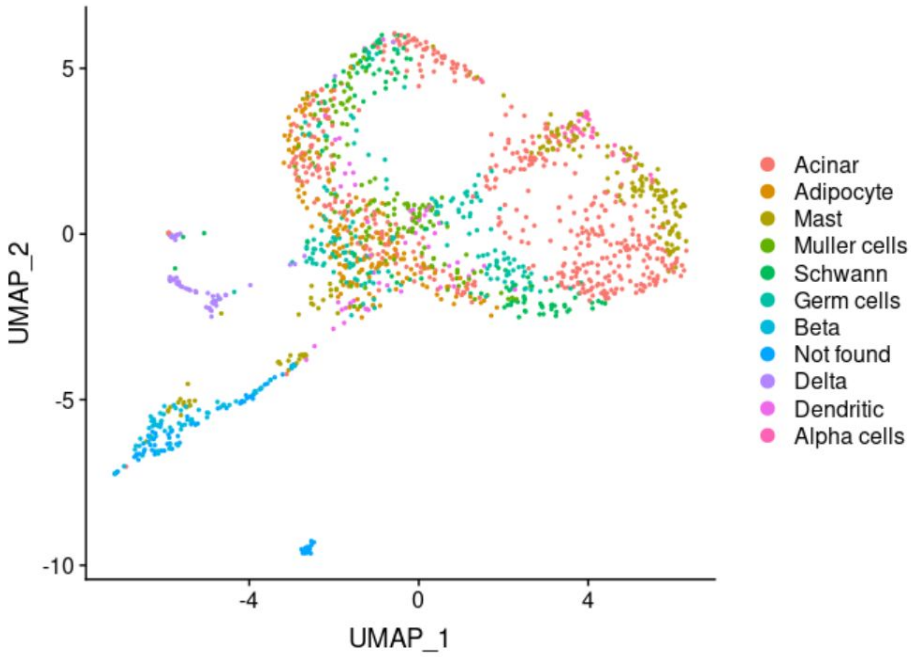

Figure 4. Scatter plot of the clustered cells using non-linear dimensional reduction (UMAP).

The top ten marker genes were chosen based on the adjusted p-value ($p < 0.05$) and were visualized using a clustered heatmap of log normalized UMI counts for those across all cells (Figure 5). This can be compared to Figure 1B of Baron et al. Although Figure 5 presents expression levels of different cell types than those presented in Figure 1B, this is simply because our data consists of different cell types and different marker genes, as explained previously.

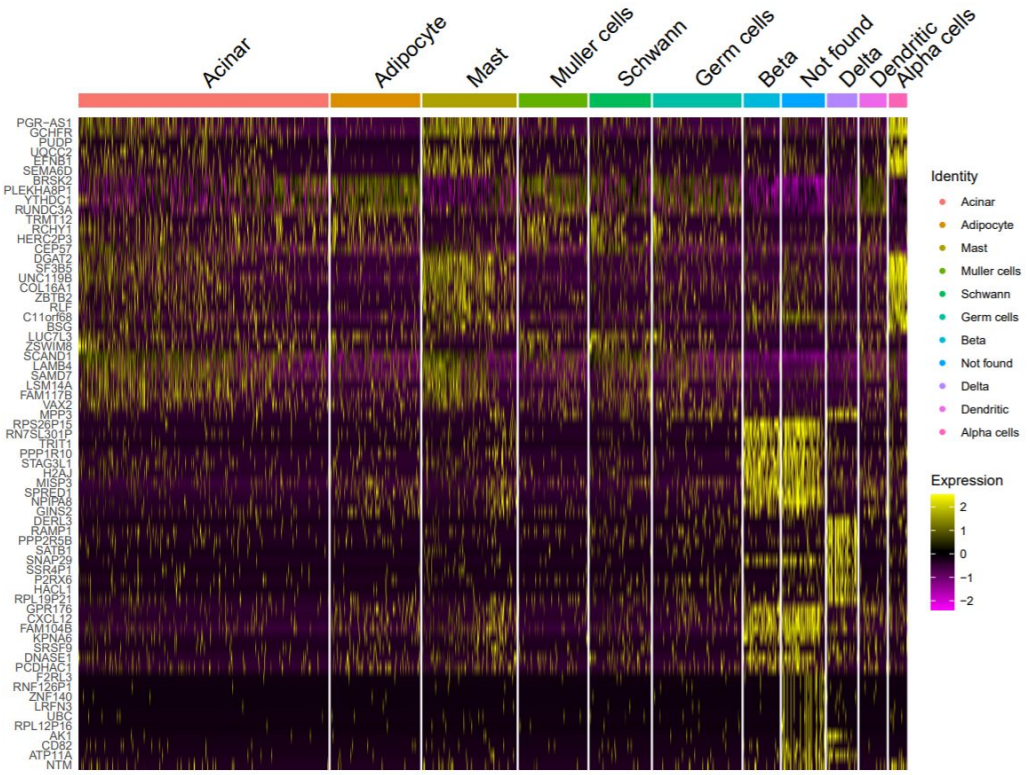

Figure 5. Heatmap of log normalized UMI counts for genes across all cells based on clusters. Marker genes are displayed on the y axis, while cell types are displayed on the x axis and color coded in the Identity legend. High expression values are colored in yellow while low expression values are colored in pink.

In addition to the marker genes chosen, each cell type also contains other differentially expressed genes. These additional genes, chosen by a cutoff of $p < 0.05$, may also be used to discriminate cell types (Table 1) and thus could be candidates for marker genes.

Table 1. Table of significant, potential marker genes per cell type ($p < 0.05$).

Cell Type	Possible Marker Genes
Acinar	PRSS2, SCAND1, UQCC2, BRSK2, PAXBP1
Adipocyte	BRSK2, PLEKHA8P1, YTHDC1, RUNDC3A, PAXBP1
Mast	HSPA13, GRB2, UNC13B, SPRED1, SRSF9
Muller cells	BRSK2, PLEKHA8P1, YTHDC1
Schwann	BCAS1, RTL8C, PRSS50, SCAND1, DISP3
Germ cells	YTHDC1, MPP3, PLEKHA8P1, BRSK2, RUNDC3A
Beta	RIMS1, SP110, ARHGDIA
Delta	MS4A8, ORC3, CPOX, COPE, CCT7
Dendritic	YTHDC1

As shown in Table 2, there were three computed cell type results that matched the biological prediction. These were the acinar, beta, and delta cells. This means that the biology-based prediction of the remaining eleven clusters did not match the computational prediction.

Table 2. Clusters with labels matching biological prediction.

Labeled cell type	**CellMarker prediction**	**Enrichr prediction**	**GO terms**
Acinar cells	Cytotoxic T, beta or delta cells	Acinar cells	SCF complex assembly (GO:0010265)
			negative regulation of catenin import into nucleus (GO:0035414)
			regulation of axon guidance (GO:1902667)
			histone H2A ubiquitination (GO:0033522)
			histone H2A monoubiquitination (GO:0035518)
			regulation of catenin import into nucleus (GO:0035412)
Beta cells	Not found	Beta cells	amylin receptor signaling pathway (GO:0097647)
			positive regulation of protein glycosylation (GO:0060050)
			calcitonin family receptor signaling pathway (GO:0097646)
			dimeric G-protein coupled receptor signaling pathway (GO:0038042)
			fatty acid alpha-oxidation (GO:0001561)
			positive regulation of glycoprotein biosynthetic process (GO:0010560)
			negative regulation of ERAD pathway (GO:1904293)
Delta cells	Alpha or beta cells	Alpha or delta cells	regulation of cAMP biosynthetic process (GO:0030817)
			entrainment of circadian clock by photoperiod (GO:0043153)
			photoperiodism (GO:0009648)
			negative regulation of cAMP biosynthetic process (GO:0030818)
			spliceosomal complex assembly (GO:0000245)

There were seven cell clusters where the biological prediction did not match the computational prediction, as shown in Table 3.

Table 3. Clusters with cell type labels differing from biological predictions.

Labeled cell type	**CellMarker prediction**	**Enrichr prediction**	**GO terms**
Acinar cells	Alpha or beta cells	Schwann cells	regulation of insulin secretion (GO:0050796)
			spinal cord motor neuron differentiation (GO:0021522)
			cardiac ventricle formation (GO:0003211)
			lymphoid progenitor cell differentiation (GO:0002320)
			cell differentiation in spinal cord (GO:0021515)
			mitral valve development (GO:0003174)
			noradrenergic neuron differentiation (GO:0003357)
			mitral valve morphogenesis (GO:0003183)
Mast cells	Gamma cells	Activated stellate cells	fibroblast growth factor receptor signaling pathway (GO:0008543)
			cellular response to fibroblast growth factor stimulus (GO:0044344)
			G2/M transition of mitotic cell cycle (GO:0000086)
			cell cycle G2/M phase transition (GO:0044839)
			regulation by virus of viral protein levels in host cell (GO:0046719)
			regulation of ERAD pathway (GO:1904292)
			regulation of protein exit from endoplasmic reticulum (GO:0070861)
Schwann cells	Cytotoxic T cells	Beta cells	regulation of ERAD pathway (GO:1904292)
			regulation of protein exit from endoplasmic reticulum (GO:0070861)
			regulation of retrograde protein transport, ER to cytosol (GO:1904152)
			entrainment of circadian clock by photoperiod (GO:0043153)
			photoperiodism (GO:0009648)
			regulation of insulin secretion involved in cellular response to glucose stimulus (GO:0061178)
Beta cells	Cytotoxic T cells	Gamma cells	regulation of ERAD pathway (GO:1904292)
			tRNA wobble uridine modification (GO:0002098)
			tRNA wobble base modification (GO:0002097)
			regulation of insulin secretion involved in cellular response to glucose stimulus (GO:0061178)
			intrinsic apoptotic signaling pathway in response to endoplasmic reticulum stress (GO:0070059)
Not found	Not found	Quiescent stellate cells	regulation of cardiac muscle cell apoptotic process (GO:0010665)
			coronary vasculature morphogenesis (GO:0060977)
			positive regulation of DNA damage response, signal

			transduction by p53 class mediator (GO:0043517)
			negative regulation of peptidyl-threonine phosphorylation (GO:0010801)
			positive regulation of signal transduction by p53 class mediator (GO:1901798)
Alpha cells	Beta cells	Macrophage	platelet dense granule organization (GO:0060155)
			nucleoside diphosphate metabolic process (GO:0009132)
			nucleoside triphosphate biosynthetic process (GO:0009142)
			entry of bacterium into host cell (GO:0035635)
			entry of bacterium into host cell (GO:0035635)

There were five clusters whose computationally predicted cell type was not one of the cell types used in the Baron et al. (Table 4). Since only the cell types discussed in the original study were used in the biological analysis, these cells could not have been predicted.

Table 4. Prediction of clusters with non-pancreatic cell type labels.

Labeled cell type	CellMarker prediction	Enrichr prediction	GO terms
Adipocyte	Cytotoxic T or acinar cells	Cytotoxic T cells	regulation of ERAD pathway (GO:1904292)
			negative regulation of NIK/NF-kappaB signaling (GO:1901223)
			negative regulation of viral-induced cytoplasmic pattern recognition receptor signaling pathway (GO:0039532)
			negative regulation of defense response to virus (GO:0050687)
			regulation of insulin secretion involved in cellular response to glucose stimulus (GO:0061178)
Muller cells	Stellate, acinar, or alpha cells	Epsilon cells	acylglycerol acyl-chain remodeling (GO:0036155)
			lipoprotein localization (GO:0044872)
			lipoprotein transport (GO:0042953)
			adipose tissue development (GO:0060612)
			fatty acid homeostasis (GO:0055089)
			nitric oxide biosynthetic process (GO:0006809)
			cellular response to fatty acid (GO:0071398)
			nitric oxide metabolic process (GO:0046209)
Germ cells	Cytotoxic T or acinar cells	Mast cells	ectoderm development (GO:0007398)
			positive regulation of response to wounding (GO:1903036)
			protein import into mitochondrial matrix (GO:0030150)
			regulation of ERBB signaling pathway (GO:1901184)
			positive regulation of wound healing (GO:0090303)
			regulation of wound healing (GO:0061041)
			ERBB2 signaling pathway (GO:0038128)
			ERBB2 signaling pathway (GO:0038128)
Germ cells	Cytotoxic T cells	Not found	sensory perception of smell (GO:0007608)
			sensory perception of smell (GO:0007608)
Dendritic cells	Acinar or alpha cells	Ductal cells	negative regulation of cellular extravasation (GO:0002692)
			negative regulation of dendritic cell apoptotic process (GO:2000669)
			coronary vasculature morphogenesis (GO:0060977)
			regulation of leukocyte mediated cytotoxicity (GO:0001910)
			induction of positive chemotaxis (GO:0050930)
			regulation of dendritic cell apoptotic process (GO:2000668)
			regulation of actin filament length (GO:0030832)
			positive regulation of calcium ion import (GO:0090280)

Discussion

Of the fourteen clusters, three of the cell type labels were confirmed with biological analysis. This improves the confidence of the computationally-determined cell type label of these clusters. However, there were eleven cell type predictions that differed. It is important to note that six of these clusters were computationally predicted to be cells that are not found in the pancreas. This is because the cell types discussed in the original study were not in the results for the computational analysis, and therefore the most significant cell type from PanglaoDB was used. This resulted in the identification of several cell types that are not found in pancreatic tissue. However, since this is a pancreatic tissue sample, these computationally-predicted cell types are incorrect. Because of this, the biologically predicted cell types provide the most logical classification.

The remaining five clusters had cell type predictions that differed between the two analyses. There are a few possible explanations for this large number of incongruous results. One explanation can be attributed to the relatively small number of significant marker genes associated with each cluster compared to those found in Baron et al. Because there are so few genes, the GO terms associated with each cluster are highly variable. For example, in Table 3, some of the GO terms associated with the Schwann cell cluster are related to photoperiodism and circadian rhythm, which are not expected attributes of pancreatic cell types. Because this cell cluster is only characterized by three significant marker genes, it is likely that one or two of those genes is skewing the GO term results.

Another explanation of the differences between the computational and biological predictions is that pancreas cells share similar biology and cannot be distinguished by GO terms. An example of this can be seen with the endocrine cells of the pancreas: alpha, beta, and delta cells[13]. Though each of these cells produces a specific molecule, they are all pancreatic endocrine cells and likely share much of the same biological characteristics. This could explain why the alpha cell cluster was biologically predicted to be beta cells, as seen in Table 3.

Conclusion

Although three cell clusters yielded the same computational and biological results that were discussed in Baron et al., the remainder of the clusters did not match or were not discussed. Based on these results, this study did reproduce the results found in the paper. While our reproduction of Baron et al. had limited success, the majority of pancreatic cell types were not accurately classified or characterized.

It is possible that the UMI counts matrix could have been filtered in a way that more accurately captured gene expression for each cell type. It is also possible that different methods of clustering, such as using a different number of PCs or a different resolution, could have resulted in more distinct cell clusters.

Most of the marker genes mentioned in the paper were not present in our data after being filtered, and therefore we relied on PanglaoDB to find the cell types. In most clusters, there are many genes with a very low adjusted p-value and they belong to various cell types but we went with the ones with the lowest p-value in each cluster. Since we are at times disregarding certain genes with a very low p-value that belong to other cell types, there is a possibility that the clusters have been wrongly labeled. Furthermore, while choosing the possible novel marker genes, only five genes were chosen for each cell type based on their adjusted p-value. Many genes were not included for some of the cell types, and so there is a possibility that we have missed some potential marker genes of interest.

When conducting the biology-based prediction analysis, it was difficult to filter through the resulting GO terms. In this study, any results that seemed too broad (ie “positive regulation of translation”) were omitted from the analysis. This was challenging because this filtering was based on the personal assessment of the results; there was no standardization to the process.

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. Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., ... & Kirschner, M. W. 2015. "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells." Cell. *161*(5), 1187-1201.
3. Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. 2012. "CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification." Cell reports, *2*(3), 666-673.
4. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120.
5. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2016). Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference. *bioRxiv*, 021592.
6. Human Release 34 (GRCh38.p13) More information about this assembly (including patches, scaffolds and haplotypes) Go to GRCh37 version of this release GTF / GFF3 files. (n.d.). Retrieved May 2, 2020, from https://www.gencodegenes.org/human/
7. Soneson C, Love MI, Robinson MD. 2015. "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences." F1000Research. 4. doi:10.12688/f1000research.7563.1.
8. Butler, A., Hoffman, P., Smibert, P. et al. 2018. "Integrating single-cell transcriptomic data across different conditions, technologies, and species." Nat Biotechnol. 36, 411–420. https://doi.org/10.1038/nbt.4096
9. Franzén, O., Gan, L.M., Bjorkegran, J.L.M. 2019. "PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data." Oxford University Press.
10. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. 2013. "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool." BMC Bioinformatics. 128(14).
11. 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 .
12. Xinxin Zhang, Yujia Lan, Jinyuan Xu, Fei Quan, Erjie Zhao, Chunyu Deng, Tao Luo, Liwen Xu, Gaoming Liao, Min Yan, Yanyan Ping, Feng Li, Aiai Shi, Jing Bai, Tingting Zhao, Xia Li, Yun Xiao. "CellMarker: a manually curated resource of cell markers in human and mouse." Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D721–D728, https://doi.org/10.1093/nar/gky900
13. VIVO Pathophysiology. (n.d.). "Functional anatomy of the endocrine pancreas." Retrieved May 2, 2020, from http://www.vivo.colostate.edu/hbooks/pathphys/endocrine/pancreas/anatomy.html