Project-4: Single Cell RNA-Seq Analysis of Pancreatic Cells

Konrad Thorner, Aishwarya Deengar, Jia Liu, Morgan Rozman

Introduction

While RNA-sequencing continues to be a powerful method in studying the transcriptome, a major shortcoming is the way in which it collectively analyzes heterogeneous populations of cells. In the study of a complex organ like the pancreas, having single-cell resolution allows for distinguishing between cell types and understanding their unique gene expression profiles. Baron et al. employed such a single-cell RNA-seq approach to human and mouse pancreatic samples, identifying multiple cell types, finding variations within types, and deconvoluting bulk RNA-seq data [1].

Our analysis seeks to replicate the original study by identifying the different cell types in the pancreas of a single donor. The unprocessed sequence data will be aligned, counted, and filtered, and cells will be clustered in a manner that should reflect their type. Marker genes for each cluster will also be determined and used to infer associated functions.

Data

Processing Sequencing Libraries to Count Matrix

In the original study samples were obtained from 4 humans and 2 mice strains, while for this analysis only the RNA-seq data of the samples obtained from the 51 year old woman were used.

The single cell RNA-seq data comprises a matrix of reads and genes. Also provided are preprocessed files containing barcodes and Unique Molecular Identifiers (UMI). Barcodes are unique cell signatures, 19 base pairs in length. UMIs correspond to the RNA transcripts and are six base pairs in length [1]. The barcodes and UMIs are extracted, a cumulative distribution is performed, and whitelisting was done. The criterion was to remove the outliers by filtering those barcodes which have a standard deviation greater than three as they are the ones that are too infrequent to be informative. A bash script was used to achieve these steps.

The number of reads before and after whitelisting are provided in Table 1 below.

Table 1. Number of reads per sample before and after whitelisting.

Samples	Reads (non-whitelisted)	Reads (whitelisted)
SRR3879604	564226059	1831
SRR3879605	392517479	1770
SRR3879606	368094423	1845

Next, Salmon analysis was performed on the whitelisted genes. Salmon is a tool for transcript quantification from RNA Seq data. It runs in two phases- Indexing and Quantification. Indexing has to be run once in the beginning while the quantification step is specific to the number of RNA-seq data and is run multiple times depending on the data.

For indexing, the transcripts were downloaded from Genecode and then indexing was performed. Next, using the salmon website as a reference the script for quantification was framed and run to get the counts matrix.

The mapping statistics have been tabulated below. The mapping rate is low which might be due to the short reads relative to the minimum required exact match length. The default is supposed to be 31 while the length of these reads were lower, as mentioned above. The unique barcodes and UMIs are also reported in the table below. Approximately two UMIs per barcode per sample are observed which is optimal and concludes that the possible occurrence of PCR amplification has been overcome by the filtering statistics used.

Table 2. Additional Salmon statistics for each sample.

Samples	Mapping Rate	Unique Barcodes	Unique UMIs
SRR3879604	33.07%	1946820	4565178
SRR3879605	28.03%	1968550	4316538
SRR3879606	28.42%	1854940	4069470

Methods

Filtering

The UMI count matrices are loaded into R using the tximport package [2]. The counts are then merged into a single Seurat object within which all further analysis occurs [3]. Filtering is performed to keep genes found in 0.1% of cells, and cells with a number of genes between 200 and a chosen maximum. Cells are also filtered based on percentage of counts corresponding to mitochondrial genes. After normalization, the analysis is limited to the top 2000 genes with the highest variation.

Clustering
Data is scaled and principal component analysis (PCA) is used for dimensionality reduction. An optimal number of dimensions is chosen using the elbow method, and the neighbors of each cell are found. Clustering is performed, with the resolution modified to get the desired number of clusters.

Identify marker genes for each cluster
Use a differential expression method to identify the marker genes in each cluster. Only select highly expressed genes and the log fold-change of the average expression greater than threshold (0.25) as marker genes for each cluster.

Find novel marker genes.
We use 0.05 as the threshold of adjusted p value of gene expression differences and only use highly expressed genes to get differential expressed genes. Then, to get the novel gene set, filter out genes that already exist in the marker gene set.

Gene Set Enrichment Analysis
Gene set enrichment was performed with enrichR [5] because this tool allows us to perform enrichment analysis on very small sets of genes. GSEA [6] was considered and tested on larger gene sets and the results are consistent with enrichR. Ultimately, we did not use GSEA due to the difficulty in analysing very small sets of genes. A filter of p-value less than 0.05 was applied to ensure only significantly expressed genes were included in analysis. For most clusters, we also applied a filter of log fold change of 1. However, for three of the clusters, this filter was too stringent and resulted in no marker genes. For these clusters, the log fold change cutoff was 0.75.

Results
The unfiltered datasets taken from Salmon contain 60,233 genes and 5,441 cells in total. When selecting criteria to filter cells by, a sign of low sample quality is reads corresponding to mitochondrial genes. The percent of reads that come from mitochondria is plotted as shown in Figure 1, and a cutoff of 20% is chosen as most cells are within that range. Similarly, when plotting the number of features (genes), a maximum of 3800 is chosen to exclude outliers that are likely doublets. There is also a recommended minimum of 200 features. 3,583 cells are ultimately used in our analysis.

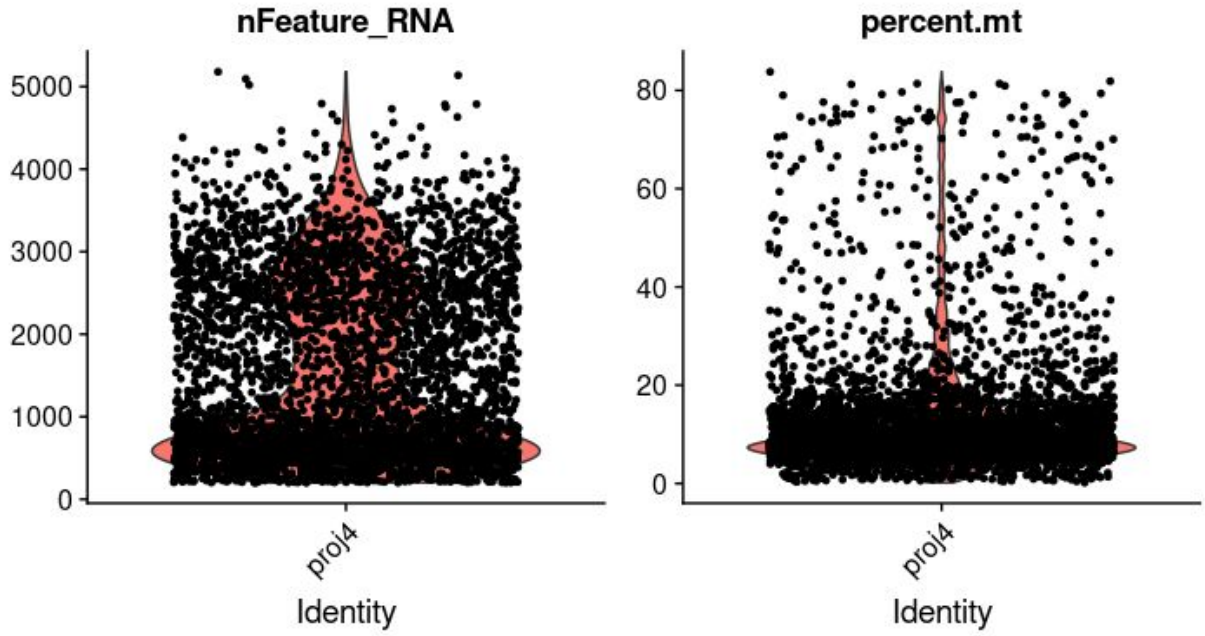

Figure 1. Plots of metadata for each cell within Serurat. "nFeature_RNA" is the number of unique genes in each cell, while "percent.mt" is the percentage of reads that align to mitochondrial genes in each cell.

In the filtering of the genes, we include only those found in a minimum of two cells, as it was found that this represents approximately 0.1% of cells within each sample. Variance is also considered, and from the 27,140 remaining genes, the 2000 with the highest variability are selected.

Cells were then clustered, with the resolution modified to generate 15 clusters as in the original paper. The percentage of cells within each of these clusters is displayed in Figure 2. A comparison of the largest and smallest clusters shows that certain cell populations can differ by a factor of ten in abundance.

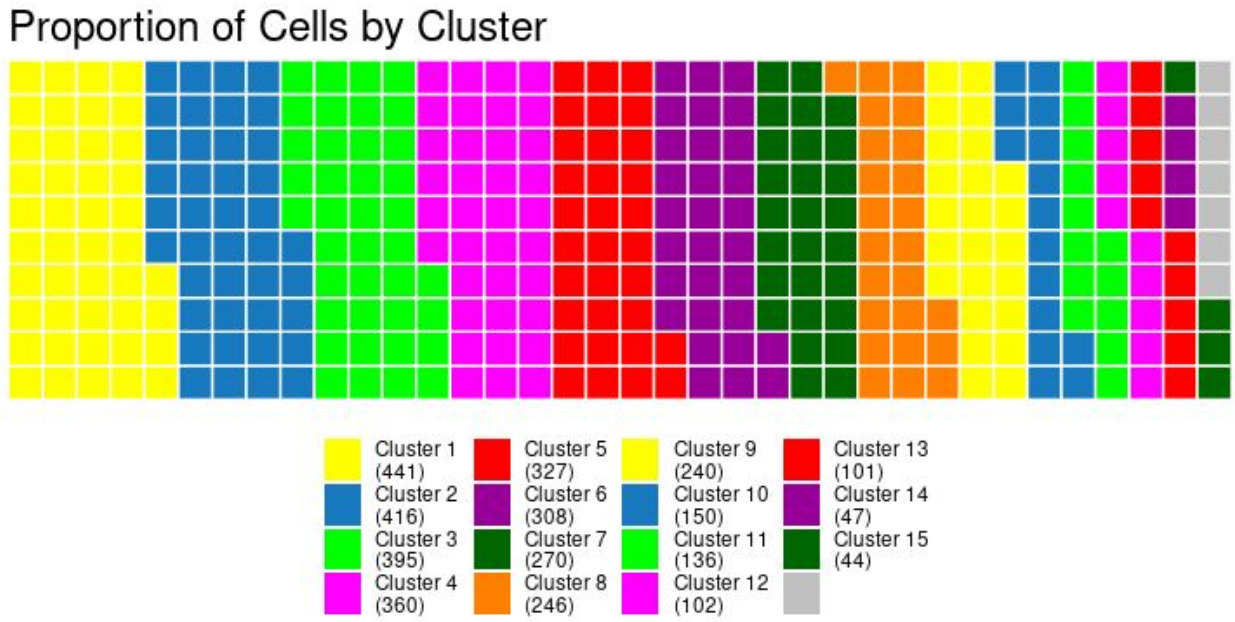

Figure 2. Waffle chart of cells in each cluster. Visualizes the proportion of the entire dataset each of the 15 clusters occupies following clustering.

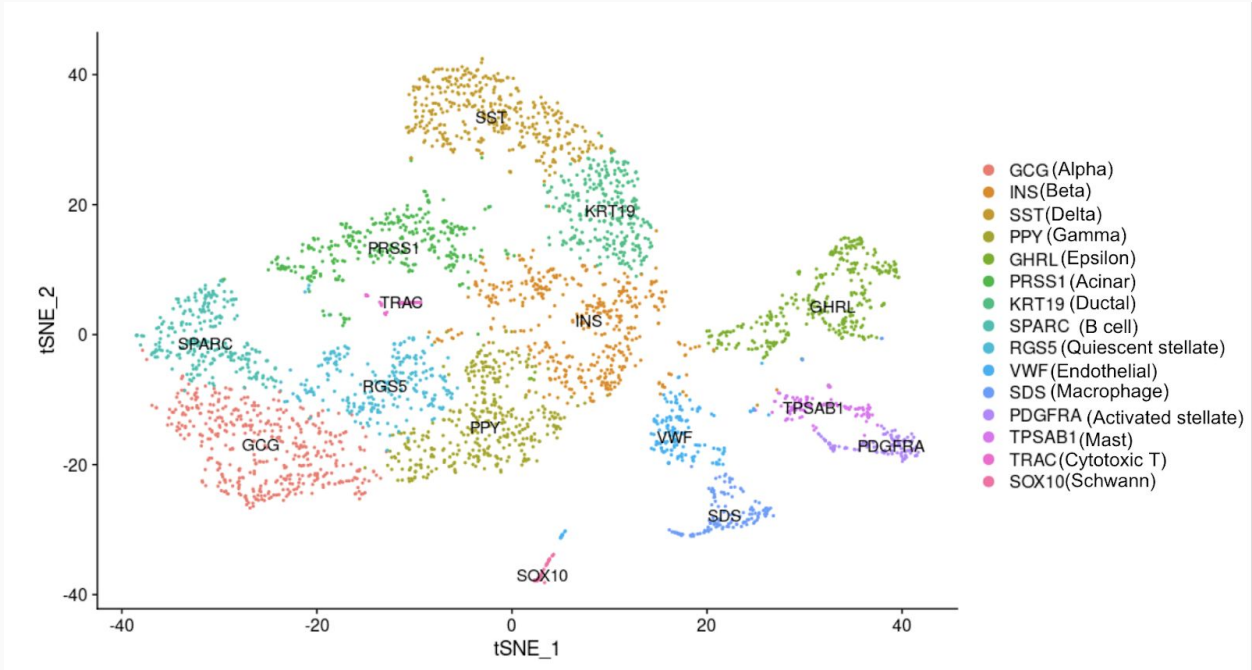

Figure 3. tSNE plot of cells marked by mark genes from the paper. The detected clusters are indicated by different colors and assigned to 15 cell types.

Figure 3 shows 15 cell types and their maker genes in tsne, all cells are from mouse or human. We use marker genes in the paper to label clusters to cell types.

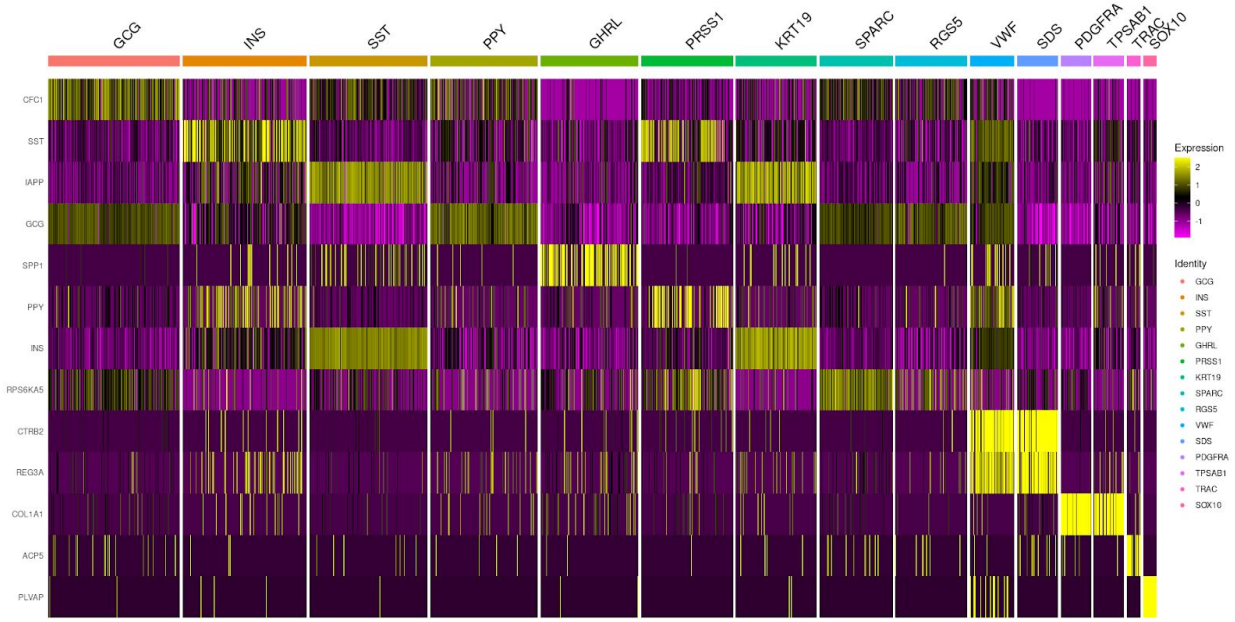

Figure 4. Clustered heatmap of log normalized UMI counts for top 1 marker genes across all cells . Marker gene expression shows differences between clusters which are tagged with marker genes. (each bar in different color marked by gene in the top of plot stands for a specific cell type).

In Figure 4, log normalized counts for top 1 marker gene expression is shown. The clusters can be separated easily because each cluster has an identical expression pattern of markers which means the quantity of clustering is pretty high.

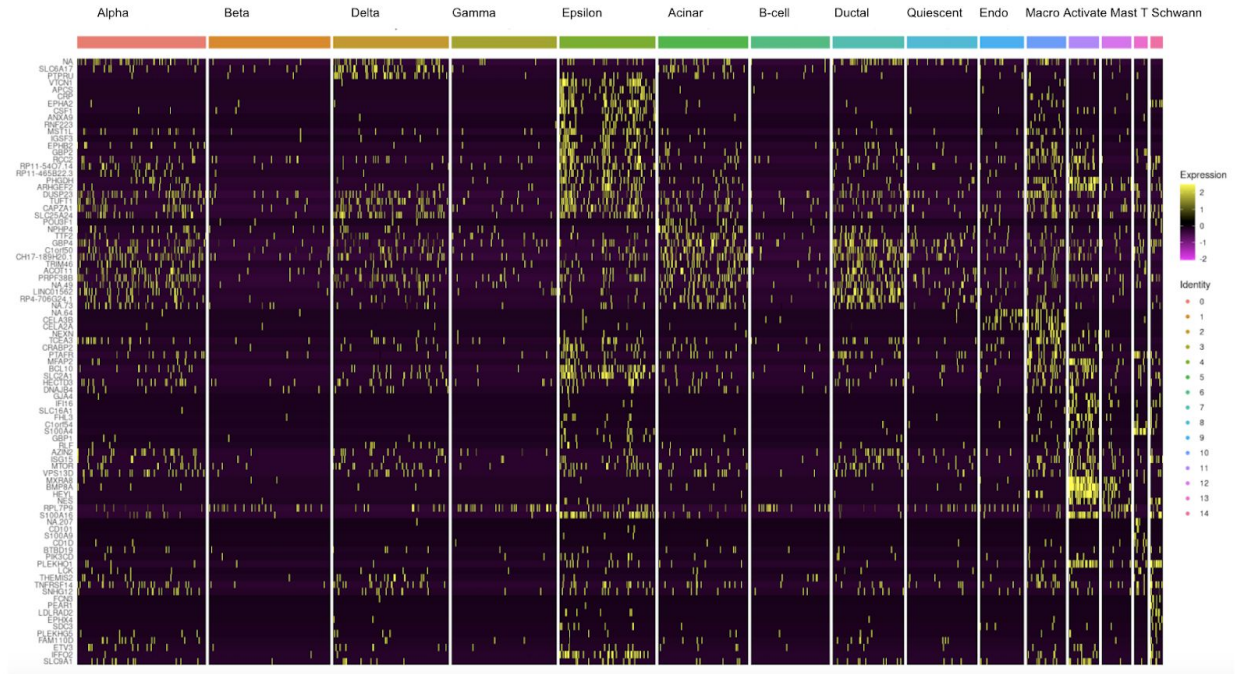

Figure 5. Clustered heatmap of log normalized UMI counts for novel marker genes across all cells. Cell types are tagged in colorful bars.

In Figure 5, it shows part of novel genes that are differentially expressed in clusters and are just as discriminative of cell type as the marker genes. Most clusters can be recognized clearly as they have different gene expression patterns.

Gene Set Enrichment Analysis

The number of marker genes in each cluster based on a p-value threshold of 0.05 and the log fold change threshold is summarized in Table 3.

Table 3. Number of marker genes in each cluster that achieve a p-value less than 0.05 and a log fold change cutoff as specified.

Marker Genes per Cluster

Cluster	Log FC Threshold	Marker Genes
Alpha	1	4
Beta	1	3
Delta	1	17
Gamma	0.75	9
Epsilon	1	98
Acinar	1	5
B cell	1	7
Ductal	0.75	24
Quiescent Stellate	0.75	4
Endothelial	1	12
Macrophage	1	100
Activated Stellate	1	154
Mast	1	49
Cytotoxic T	1	114
Schwann	1	221

Baron et al. provide markers for each cell type they identified. Based on this, the clusters that contain the given marker gene are provided in Table 4. These results provide an idea of the actual cell types of each cluster based on the marker genes they contain. These assignments disregard the thresholds set for the marker genes used in the gene set enrichment analysis because for some clusters, no marker genes are significantly expressed. Therefore, examining all differentially expressed genes with no log fold change cutoff provides cell type assignments for each of these clusters. No clusters contained marker genes GHRL, RGS5, TPSAB1, TRAC, SOX10. Note that some marker genes were found in multiple clusters and that some clusters contained multiple marker genes. This could indicate cell type overlap in the clusters or that some clusters should be combined into a single cell type cluster.

Table 4. Clusters that contain each assigned marker gene from the original paper to identify true cell type. Some clusters contain multiple marker genes and some marker genes are found in multiple clusters.

Assigned Marker Genes in Clusters

Cell Type	Marker Gene	Cluster(s)
Alpha	GCG	Alpha Gamma Ductal Quiescent Stellate Endothelial
Beta	INS	Delta B cell
Delta	SST	Beta Acinar
Gamma	PPY	Beta Acinar
Acinar	CPA1	Endothelial Macrophage
Ductal	KRT19	Epsilon Macrophage
Activated stellate	PDGFRA	Activated Stellate

Endothelial	VWF	Schwann
Macrophage	SDS	Cytotoxic T

For each cluster, enriched terms were identified. Table 5 summarizes the top notable enriched terms in each cluster. Ontology terms were taken from The Gene Ontology (GO) knowledgebase [7,8].

Table 5. EnrichR analysis of marker genes in each cluster revealed GO terms for each cluster. These enriched terms are used to determine if the cell types assigned based on marker genes in the original paper are appropriate.

Enriched GO Terms by Cluster		
Cluster	**Enriched Terms**	
Alpha	activin receptor binding (GO:0070697)	nodal signaling pathway (GO:0038092)
Beta	neuropeptide hormone activity (GO:0005184)	regulation of cell motility (GO:2000145)
Delta	hormone activity (GO:0005179)	regulation of protein localization to membrane (GO:1905475)
Gamma	activin receptor binding (GO:0070697)	nucleobase metabolic process (GO:0009112)
Epsilon	cadherin binding involved in cell-cell adhesion (GO:0098641)	regulation of nitric-oxide synthase biosynthetic process (GO:0051769)
Acinar	hormone activity (GO:0005179)	response to vitamin (GO:0033273)
B cell	insulin-like growth factor receptor binding (GO:0005159)	negative regulation of protein oligomerization (GO:0032460)
Ductal	positive regulation of histone phosphorylation (GO:0033129)	histone-serine phosphorylation (GO:0035404)
Quiescent Stellate	nucleobase metabolic process (GO:0009112)	tau protein binding (GO:0048156)

Endothelial	cobalamin metabolic process (GO:0009235)	peptidoglycan binding (GO:0042834)
Macrophage	tetrapyrrole metabolic process (GO:0033013)	peptidoglycan binding (GO:0042834)
Activated Stellate	extracellular matrix organization (GO:0030198)	collagen fibril organization (GO:0030199)
Mast	extracellular matrix organization (GO:0030198)	collagen fibril organization (GO:0030199)
Cytotoxic T	MHC protein complex binding (GO:0023023)	immunoglobulin mediated immune response (GO:0016064)
Schwann	transforming growth factor beta binding (GO:0050431)	regulation of cell migration (GO:0030334)

The primary function of alpha cells is to make and release glucagon in response to glucose levels. Clusters Alpha, Gamma, Ductual, Quiescent Stellate, and Endothelial all contain the marker gene for alpha cells. Enrichment analysis revealed that clusters Alpha, Ductual, Quiescent Stellate, and Endothelial are enriched for the alpha cell type in the ARCHS4 Tissues database [9]. Clusters Alpha and Gamma are enriched for activin receptor binding. Brown et. al. found that activin proteins were primarily found in alpha cells, which provides further evidence that these clusters are appropriately assigned to the alpha cell type [10]. Cluster Ductual is enriched for histone-serine phosphorylation which has previously been linked to endothelial cells [11]. This indicates that cluster Ductal may be inappropriately labeled as an alpha cell type.

Beta cells synthesize and secrete insulin and amylin, hormones that regulate blood glucose levels. Cluster Delta is enriched for hormone activity and cluster B-cell is enriched for insulin-like growth factor receptor binding. This is evidence that clusters Delta and B-cell are appropriately assigned as beta cells.

Delta cells are endocrine cells that secrete the hormone somatostatin, while gamma cells produce polypeptides. Clusters Beta and Acinar were assigned to both of these cell types based on the marker genes they contain. Both clusters are enriched for hormone activity. This is evidence that the delta cell classification may be more accurate for these clusters. Interestingly, cluster Beta is enriched for regulation of cell motility. Delta cells are responsible for regulating alpha and beta cells and their motility has been cited as one mechanism for how they reach their target [12].

Acinar cells are exocrine cells that produce enzymes and transport them to aid in digestion. Cluster Endothelial is enriched for tetrapyrrole metabolic processes, which has previously been linked to pancreatic function. According to Dwarka et. al., digestion of bilirubin, a tetrapyrrole, is dependent on enzymes secreted from the pancreas [13]. This could be evidence that cluster Endothelial is truly the acinar cell type instead of the alpha cell type. Cluster Macrophage is enriched for many of the same terms as cluster Endothelial, including those in Table 5. Ductal cells are specialized epithelial cells that deliver the enzymes produced by acinar cells to the duodenum. Therefore, these two cell types are very closely related, which could explain why cluster Macrophage contains the marker gene for both. Cluster Epsilon is enriched for general terms including cadherin binding involved in cell-cell adhesion. There is limited evidence in the results linking cluster Epsilon to ductal cells.

Cluster Activated Stellate contains the PDGFRA marker gene for the activated stellate cell type. Activated stellate cells regulate the extracellular matrix. Cluster Activated Stellate is enriched for extracellular matrix organization, confirming that this cluster most likely corresponds to activated stellate cells. Cluster Mast does not contain any of the marker genes that correspond to the cell types in the original paper. However, cluster Mast is enriched for many of the same terms as cluster Activated Stellate, indicating that it also most likely corresponds to the activated stellate cell type.

Endothelial cells transport nutrients to the pancreas and aid in regulation of beta cell function. Cluster Schwann is enriched for transforming growth factor beta binding. TGF-B has been shown to promote endothelial cell survival [14]. This is evidence that cluster Schwann may be associated with endothelial cells, but further analysis would be required to be confident that this cell type label fits. Finally, Macrophages are white blood cells known to express histocompatibility molecules (MHC-II) [15]. Cluster Cytotoxic T is enriched for MHC protein complex binding and immunoglobulin mediated immune response, confirming that this cluster is most likely macrophage cells.

Discussion

In the initial filtering of the data, it is evident that many cells have reads in excess of 20% from mitochondrial genes, suggesting issues with the sample preparation. This, combined with the fact that the samples came from a single individual, are factors to consider in the explanatory power of our results.

Using cell type assignments that are based on whether a cluster contains the marker genes specific for each cell type in the original paper yields just nine distinct cell types from the 15 clusters. Gene set enrichment analysis provides evidence that most of these clusters are appropriately assigned using this method. Notably, there are four clusters that all fit under the alpha cell type and two clusters that fit under the beta, delta, gamma, acinar, and ductal cell types. These cell types with multiple clusters could be due to overclustering of the original data. For example, clusters Alpha, Gamma, Ductal, and Quiescent Stellate may truly belong to the

same cell type and should be clustered together. On the other hand, assigning cell types in this way is crude, and may not be appropriate for our data set. One way to do so would be to analyze the marker genes independently of the original paper's definitions to assign cell types, which could result in different results. For example, cluster Endothelial contains the marker gene for the alpha cell type, but enrichment analysis revealed that it is more appropriately assigned as an acinar cell.

Compared to the result in the paper in Figure 1d, our Beta cell in figure 3 looks more separate and has less cell number. That may be because we have two sources of cells (mouse and human) and Figure 1d in paper just shows human cells, so the differences between our data is greater than paper. Besides that, the number of cells is lower may be because we have a higher threshold and filtered more cells.

Our heatmap in Figure 4 has different marker genes from Figure 1b in the paper, but it works pretty well as we can identify different clusters by differential gene expression pattern. And each cluster in Figure 4 has one or more highly regulated expression patterns just like the paper.

Conclusion

The marker genes in our result agree with those in the paper. We used marker genes in the paper to label clusters to cell types, but the B cell marker hasn't been mentioned in their results. So we find our own marker for B cells in the marker gene dataset. And by comparing high regulated gene expression, we can find our own marker genes which can also identify clusters.

References

[1] Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res*. 2017;27(3):491-499. doi:10.1101/gr.209601.116.

[2] 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.

[3] Soneson C, Love MI, Robinson MD (2015). "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences." *F1000Research*, 4.

[4] Stuart, Tim, et al. "Comprehensive Integration of Single-Cell Data." *Cell*, vol. 177, no. 7, 2019, doi:10.1016/j.cell.2019.05.031.

[5] 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).

[6] Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550) and Mootha, Lindgren, et al. (2003, Nat Genet 34, 267-273).

[7] Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. May 2000;25(1):25-9.

[8] The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. Jan 2019;47(D1):D330-D338.

[9] Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, Silverstein MC, Ma'ayan A. Massive mining of publicly available RNA-seq data from human and mouse. Nature Communications 9. Article number: 1366 (2018), doi:10.1038/s41467-018-03751-6.

[10] Brown ML, et al. Effects of activin A on survival, function and gene expression of pancreatic islets from non-diabetic and diabetic human donors. Islets. 2014;6:e1017226. doi: 10.1080/19382014.2015.1017226.

[11] Alghamdi T. A., Batchu S. N., Hadden M. J., Yerra V. G., Liu Y., Bowskill B. B., et al. (2018). Histone H3 serine 10 phosphorylation facilitates endothelial activation in diabetic kidney disease. Diabetes 67, 2668–2681. 10.2337/db18-0124.

[12] Drigo R. A. E., et al. Structural basis for delta cell paracrine regulation in pancreatic islets. Nat. Commun. 10, 1–12 (2019).

[13] Dwarka D., et al. New insights into the presence of bilirubin in a plant species strelitzia nicolai. AJTCAM. 2017. 14(2): 253–262.

[14] Viñals F., Pouysségur J. Transforming growth factor beta1 (TGF-beta1) promotes endothelial cell survival during in vitro angiogenesis via an autocrine mechanism implicating TGF-alpha signaling. Mol. Cell. Biol. 2001;21:7218–7230. doi: 10.1128/MCB.21.21.7218-7230.2001.

[15] Carrero J. A., McCarthy D. P., Ferris S. T., Wan X., Hu H., Zinselmeyer B. H., et al. (2017). Resident macrophages of pancreatic islets have a seminal role in the initiation of autoimmune diabetes of NOD mice. Proc. Natl. Acad. Sci. U.S.A. 114 E10418–E10427. 10.1073/pnas.1713543114.