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

Group: Saxophone
TA: Jing Zhang

Jason Rose: Data Curator
Yilin Yang: Programmer
Sooyoun Lee: Analyst
Daniel Goldstein: Biologist

INTRODUCTION

Bulk transcriptomics profiling has been used to provide insight into tissues, disease, and development through RNA-seq using the entire population of cells within a sample. Single-cell sequencing is similar but for one cell at a time. First used in 2009 as a method to sequence scarce cells [1], single-cell sequencing has been a tool to provide cellular heterogeneity. This kind of specificity allows us to create cellular maps, discover new targets, or discover new cells [2]. The scaling of this technology has grown where it can be applied to the entire human body. For this project, we are using data from Baron et al. [3], a study that uses single-cell sequencing on four human and two mice pancreata.

The most well-known function of the human pancreas is the control of blood sugar levels through hormone production and to aid in digestion. Disruptions to these functions can cause severe consequences. A well-known example is the inability of the pancreas to produce sufficient insulin (diabetes mellitus), preventing glucose from being removed from the blood. A build-up of glucose in the blood will cause damage to other organs, such as the kidney or heart. Beta cells found in pancreatic inlets are associated with insulin production, necessitating a need to understand how insulin-secreting beta cells and other cell types in the pancreas interact.
As most pancreas transcriptome studies have been bulk sequenced, Baron et al. sought to use single-cell RNA sequencing to attain a molecular understanding of cell diversity.

Using the single-cell RNA sequenced data from one of the four donors, a 51-year-old female, this paper will explain our reproduced findings of Baron et al. using modern analytical methods/software.

DATA

Drop-seq (droplet sequencing) was conducted via inDrop cell encapsulation using lysis buffer, reverse transcription reagents, and barcoded oligonucleotide primers. Each droplet contains the mRNA of the lysed cell and is barcoded when complementary DNA is synthesized. The droplets are broken after barcoding and combined before sequencing. An Illumina Hiseq 2500 machine was used for paired end read sequencing. The source paper sequenced four human cadaver donor samples and two mouse samples. We are focusing on the drop-seq samples associated with the 51 year old female: SRR3879604, SRR3879605 and SRR3879606. The barcode sample data from Baron et al. was obtained from the NCBI (National Center for Biotechnology Information) database in fastq files.

METHODS

Generate UMI counts matrix

The number of reads per barcode for each SRR were counted and used to create a cumulative distribution plot to determine cutoff for a whitelist. This whitelist would allow us to focus on the barcoded data with the highest counts to ensure capture cell samples. We decided to use the top 1000 barcodes with the highest count among each sample then merged these barcodes into a single list for analysis. Among the three SRR samples, 2824 barcodes were on that list.

To generate a UMI counts matrix, Salmon and Salmon Alevin [4] software was used. Salmon was used to create an index using the v37 human transcriptome attained from the GENCODE database [5]. A transcript to gene ID mapping file was also needed to run Salmon Alevin. Data for this file was also obtained from the GENCODE database using their v37 annotation file for humans. Salmon Alevin was run on each SRR barcode 1 and 2 samples simultaneously, but on the samples in the whitelist. Custom lengths were designated; a UMI length of 6, barcode length of 18 and end length of 5.

As this process produced poor results with no reads mapped (99.9949% of reads were discarded by noise), we then used a less strict whitelist to get the counts matrix as the actual noise would be filtered in a later process. The whitelist allowed a substantially higher number of barcodes (4,189,098 barcodes) in which only around 0.01% of the reads were thrown away from noise and not mapped (240576 barcodes.

Single-cell RNA-seq data Quality Control filtering

After obtaining the UMI counts matrix, we first converted the Ensembl gene identifiers to Gene Symbols. Then, the Seurat package was downloaded and used as quality control metrics for cleaning the data. The UMI count matrix was read into R, and a pre-processing workflow for the single cell RNA-seq data was followed. We kept all genes expressed in >=3 cells and with at least 200 detected genes, and filtered cells that have unique feature counts over 2,500 or less than 200, as well as cells with >10% of reads mapped to the mitochondrial genome and used violin plots to visualize these QC metrics. After removing unwanted cells from the dataset, we normalized the gene expression measurements for each cell by the total expression, multiplied this by a scale factor 10,000, and log-transformed the result. Finally, we include only the top 2000 high variance genes for further analysis.

Dimension reduction and clustering

Next, we apply a linear transformation (‘scaling’) that is a standard pre-processing step prior to dimensional reduction techniques like PCA. We used JackStraw and an Elbow plot to determine the number of principal components for the dataset. The ScaleData function shifts and scales the expression of each gene. The clusters were visualized with a scatterplot and the relative proportions of cell numbers were visualized with a bar plot.

Visualizing the clustered cells

The cell clustered were visualized using the function of UMAP. The UMAP is the simplest function that constructs a high-dimensional graph representation of the data by optimizing the low-dimensional graph to be as structurally similar as possible while the highest-dimensional graph to be simple [6]. The RenameIdents functions in Seurat were used to join identity classes together so for this case, the seurat_dataset and the new.cluster.ids were merged as a one identity class. The DimPlot function was used to graph the output of a dimensional reduction on 2D scatter plot on each point by determining the reduction [7]. Each cell was colored by its own identity classes and labeled differently by its own cell types.

Visualizing the top marker genes per cluster

First, the differentially expressed genes were selected from the seurat_dataset and labeled as the diff_expressed_genes. From the diff_expressed_genes, the top 5 marker and the top 10 marker genes were selected and the selected genes were saved as csv files. By using these top 5 marker genes, the heatmap was created by using the DoHeatmp function to express a single cell feature expression. One additional method was used to visualize and identify each cell cluster type and this method was done by using the VlnPlot function.

Finding novel marker genes

The novel marker genes were conducted to discriminate between cell types from the other differentially expressed genes. Different seurat functions were included while finding the novel marker genes. The min.pct function was used so that genes could be pre-filtered based on their minimum detection rate across both cell groups [8]. The logfc.threshold function was used to limit the genes testing which shows an x-fold difference between the two groups of the cells. The pseudocount.use was used to add the averaged expression values when calculating the logFC. As a result, a csv file that contains p-value, avg_log2FC, pct.1, pct.2, p_val_adj, cluster and the gene names were created.

Gene set enrichment analysis

Gene set enrichment of marker genes was performed using EnrichR to verify that our predicted cell labels match the labels from Table S2 in Baron et al [3, 9]. Marker genes were subset by their cluster, then filtered by p-value and log2FC to ensure these genes were statistically significant and differentially expressed in each cluster. P-value was filtered by $p < 0.01$ for all clusters, while log2FC threshold value was chosen specifically for each cluster based on the number of genes in that cluster. Clusters 0, 1, 3, 4, 5, and 12 were not filtered by log2FC due to the low number of marker genes, cluster 6 was filtered by a fold change of 1.75, cluster 9 was filtered by a fold change of 1.5, and clusters 10 and 11 were filtered by a fold change of 2.5. Clusters 2, 7, and 8 were excluded from gene set enrichment analysis because these clusters did not match any labels from the paper. Marker genes were functionally annotated with biological process and molecular function gene ontology (GO) terms.

RESULTS

The UMI matrix contains 60232 cells across 1316041 genes within 1 assay. We filtered out low-quality cells based on the criteria (see Methods), which left 12718 cells remaining. After removing unwanted cells from the dataset, we normalized the gene expression measurements for each cell by the total expression, and filtered out the counts matrix to only include the top 2,000 variable genes for further analysis. By visualizing through a variable feature plot, we were able to see that there were initially 25718 genes with low variance (Figure 2).

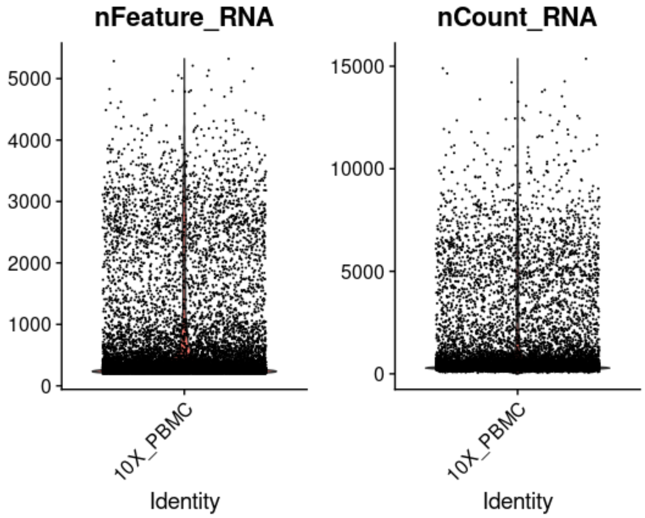

Figure 1*. *Violin plots for the number of genes, the number of molecule counts. Most of the cells have the number of genes between 1000 to 4000.

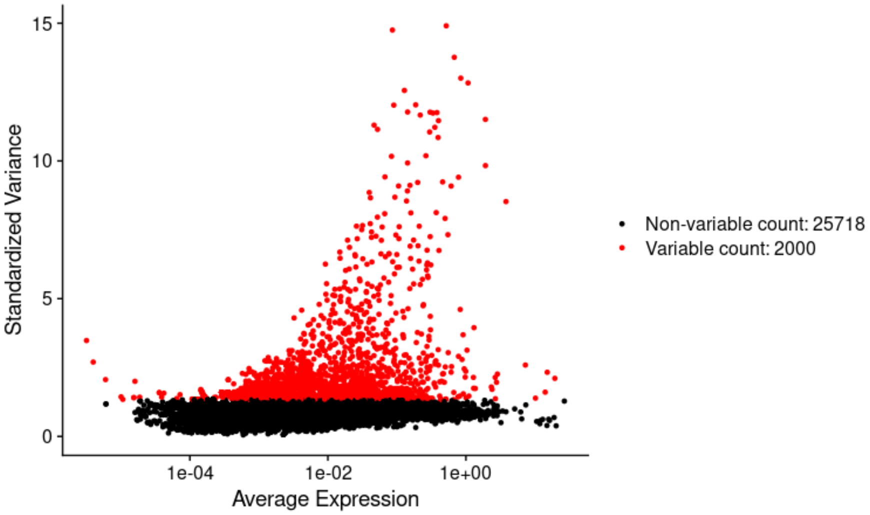

Figure 2. Scatter plot that shows the most high variable genes. 25718 genes were with low variance.

To cluster the cells, we performed principal component analysis (PCA) and also visualized with an Elbow Plot (Fig 3C), which ranks principal components based on the percentage of variance explained by each one. We observed an ‘elbow’ around PC9-10, suggesting that the majority of true signal is captured in the first 10 PCs. Our analysis resulted in 13 clusters (Fig 3D), and our nonlinear dimensionality reduction is displayed using a JackStraw plot to visualize the result (Fig 3B).

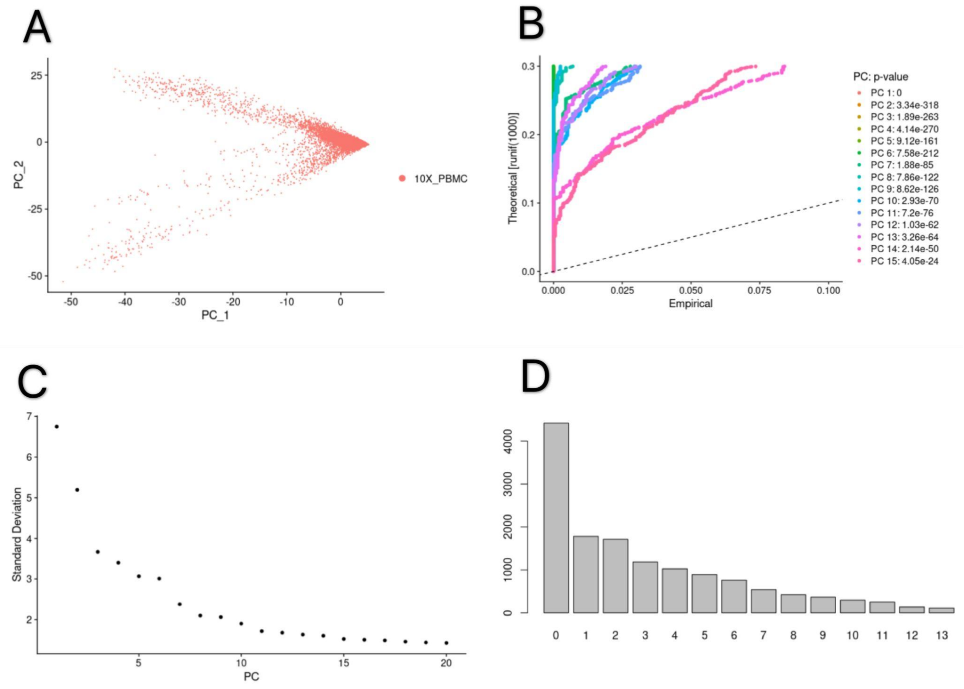

Figure 3. Dimension reduction clustering results. A) Scatter plot for gene-gene relationships. B) The JackStraw plot of 15 PC with p-values. C) The Elbow plot for 20 PC. There appears to be a sharp drop-off around PC10. D) Bar plot of relative proportions of cell numbers in each cluster, x-axis indicating number of cells, y-axis indicating each cluster. Cluster 1 has a larger proportion of cells compared to the rest.

The differential expression method of Seurat was used to analyze and exploration of single-cell RNA-seq data. By using the Seurat, we were able to identify and interpret sources of heterogeneity from single-cell transcriptomic measurements and to integrate diverse types of single-cell data [10]. First, the differentially expressed genes were selected from the dataset using the log2 fold change threshold. And from the differentially expressed genes, the top 5 and the top 10 cell markers were selected.

Cluster Level	Cell Type	Marker Gene
0	Delta/Gamma	SST, PPY
1, 6	Beta	INS
3	Acinar	CPA1
4	Alpha	GCG
5, 10	Ductal	KRT19
9	Cytotoxic T	CD3,CD8
10, 11, 12	Macrophage	CD163,CD68,IgG
12	Vascular	VWF, PECAM1, CD34

Table 1. Top 5 significant cell marker genes labeled with each cluster level and cell types. Cluster levels and the cell types that are not indicated in this table are not found in the marker gene.

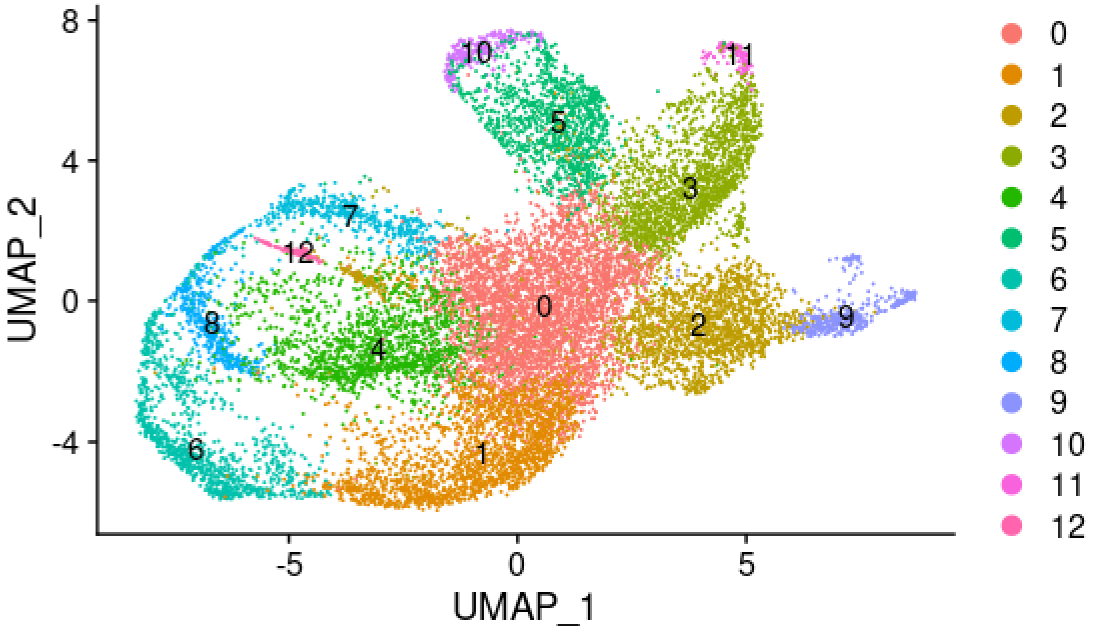

Figure 4. UMAP plot for cell clusters labeled with different colors. Clusters are labeled by specific cell types and represented with different colors. The cell types were selected from Baron et al, Table S2 [3].

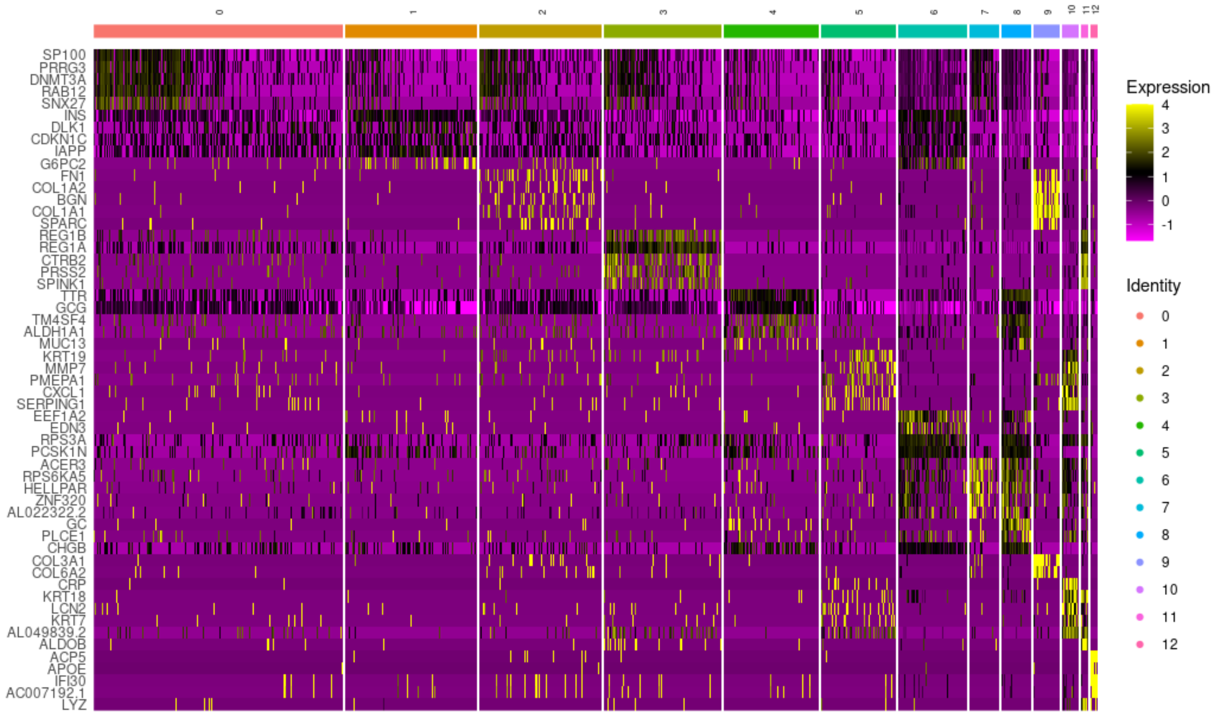

Figure 5. Clustered heatmap of top 5 marker genes. Each row represents different expression levels of each marker gene in different colors. The expression color code on the right top shows that the expression ranges from -1 to 4 and the color changes from purple to yellow as the level goes positive.

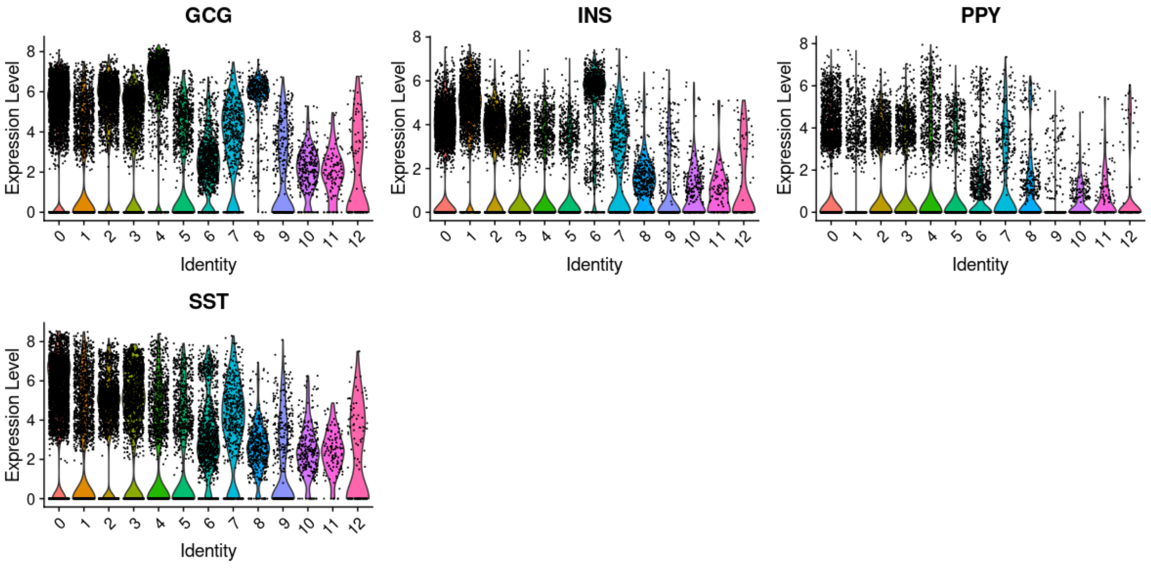

Figure 6. Violin plot for different levels of marker genes: GCG, INS, PPY, and SST of Alpha, Beta, Delta, and Gamma. In each marker gene, clusters 0, 1, 2, 3, and 4 showed relatively high expression levels compared to other clusters.

Table 2. Top 10 novel marker genes with the average log2 fold change. It represents the log ratio of the FPKM values for a pair of the expressed genes.

Cluster Level	Novel Marker Genes	avg_log2FC
0	SP100	1.38821689
1	DLK1	1.80545599
2	FN1	1.95047170
3	REG1B	3.56687298
4	TTR	2.37543952
5	CXCL1	3.05170016
6	INS	2.18324859
7	ACER3	2.60251575
8	TTR	2.88089947
9	COL1A1	4.49400796
10	CRP	2.98042440
11	ALDOB	3.96741330
12	ACP5	5.64420467

Gene Set Enrichment Analysis

For gene set enrichment analysis, certain clusters were filtered by log2FC to ensure that differentially expressed genes were annotated. Log2FC filtering reduced the number of marker genes in clusters 6, 9, 10, and 11 from 1070, 275, 1932, and 1117 to 101, 186, 97, and 155, respectively. Of the ten clusters that were labeled, seven clusters (1, 4, 6, 9, 10, 11, 12) contained marker genes that accurately matched the known functions of our predicted labels. However, gene set enrichment did not support three clusters (0, 3, 5) that were predicted to have Delta, Gamma, Acinar, and Ductal labels as there were no GO terms that represented the functions of these cell types.

Table 3. Gene Set Enrichment Analysis. The top GO terms for each cluster, its associated label(s), and the total number of marker genes after filtering by p-value and log2FC. *Top GO terms do not accurately represent the functions of the predicted cell types.

Cluster	**Label**	**Number of Marked Genes**	**Gene Set Enrichment**
0	Delta, Gamma*	32	Mitochondrial electron transport, NADH to ubiquinone; Mitochondrial ATP synthesis coupled electron transport
1	Beta	47	Insulin secretion; Insulin secretion involved in cellular response to glucose stimulus; Insulin metabolic process
3	Acinar*	84	SRP-dependent cotranslational protein targeting to membrane; protein targeting to ER; rRNA metabolic process
4	Alpha	35	Activin receptor signaling pathway; Response to glucagon; Cellular response to glucagon stimulus
5	Ductal*	81	MHC class II protein complex binding; MHC protein complex binding; antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent
6	Beta	101	Positive regulation of insulin receptor signaling pathway; Insulin secretion; Positive regulation of cellular response to insulin stimulus
9	Cytotoxic T	186	T-helper 1 type immune response; Regulation of T cell mediated immunity; Neutrophil activation involved in immune response; Regulation of immune response
10	Ductal*, Macrophage	97	Neutrophil mediated immunity; Neutrophil activation involved in immune response; Epiboly involved in wound healing

11	Macrophage	155	Antimicrobial humoral immune response mediated by antimicrobial peptide; Neutrophil mediated immunity; Innate immune response in mucosa
12	Macrophage, Vascular	163	Neutrophil degranulation; Neutrophil activation involved in immune response; Vascular endothelial growth factor receptor signaling pathway; Regulation of vascular development

DISCUSSION

Overall, the cluster cell type identification was similar to those of the original authors. Compared to the 14 clusters identified by Baron et al, we were able to identify 13 distinct clusters in the process.

The UMAP plot for the cell clusters was created by specific cell types and represented with different colors (fig 4). When comparing figure 1 with figure 1D from Baron et al, most of the cell types were represented similarly. However, we have used the method of UMAP to represent the plot while the original paper used the method of t-SNE. Also, in the original paper, some additional cell types such as activated stellate, and endothelial were included. In figure 1, the Beta cell type represents the largest range of clusters, and followed by Beta, the Acinar and Ductal cell type showed the largest cluster. Similarly, in figure 1D from the original paper, the Best cluster also represented the largest range of clusters, and then the Alpha, and Delta. Although both figures showed that the Beta is the largest cluster, the cell types that are represented in the papers are different. One possible assumption we can make is that there could be a different clustering method that the author has used while plotting the data.

The clustered heatmap of the top 5 marker genes is created in figure 5. In the figure, each row represents different expression levels of the marker gene in different colors. The color of the heatmap shows different levels of the expression of the sample for example purple represents the negative expression level while yellow represents the positive expression levels. However, when comparing figure 5 with figure 1B from the original paper, the heatmap shows a significant difference. In figure 5 of our heatmap, we have selected the top5 marker genes that were differentially expressed but a figure from Baron et al selected specific marker genes from Table S2. For example, GCG, INS, PPY, GHRL, VWF was selected from the table.

From figure 6, the marker genes from the Alpha, Beta, Delta, and Gamma were represented in the violin plot. The violin plot shows the gene expression, metrics, PC scores for each single-cell data. For each marker gene, clusters 0, 1, 3, and 4 show relatively high expression levels compared to the other clusters. One assumption we can make is that from table 1, there were few overlapping genes correlated to each cluster level such as Beta, Delta, and Gamma.

The top 5 significant cell marker genes that are labeled with each cluster level and cell types are represented in Table 1. When compared with Table S2 from Baron et al., we have few missing cell types that their gene identified. For example in Table 1, we were unable to identify the Gamma, Epsilon, Stellate, and the Mast cell types. One assumption we could make about the missing cell types is that in our table, we have selected the top 5 significant cell marker genes while the original paper did not select the most significant cell marker genes.

Seven of the ten clusters were confirmed with functional annotations from gene set enrichment analysis, including two endocrine cell types, two immune cell types, and the vascular cell type. Beta cells are endocrine cells that secrete insulin, which regulate glucose uptake in the body. Our enrichment analysis determined that the top GO terms for clusters 1 and 6 relate to insulin secretion and insulin metabolic process, which confirmed our label predictions. Cluster 4 was confirmed to be an alpha cell type as the associated GO terms, response to glucagon and cellular response to glucagon stimulus, are indicative of the role of alpha cells in glucagon secretion. The two immune cell types, cytotoxic T cells and macrophages, were also enriched with GO terms related to immune response. T-helper 1 type immune response and regulation of T cell mediated immunity supported our predicted cell type for cluster 9 as a cytotoxic T cell. While clusters 10 – 12 were predicted as macrophage cell types, the GO terms were primarily associated with another immune cell type, neutrophils; therefore, we were not as confident in our predictions for these clusters. Although cluster 12 had two predicted cell types, macrophages and vascular cells, we found enrichment for both of these cell types in our analyses, which supported this prediction.

Alternatively, there were three label predictions that were not supported by gene set enrichment analysis. We expected to find enrichment terms related to hormone secretion in cluster 0, since our predicted cell types were delta and gamma cell types, both of which are endocrine cells that produce somatostatin and pancreas polypeptide, respectively. Due to the limited total number of marker genes and the high number of mitochondrial genes, cluster 0 was enriched with mitochondrial electron transport, NADH to ubiquinone and mitochondrial ATP synthesis coupled electron transport. Similarly our cluster 3 group contained a high number of ribosomal genes, resulting in our enrichment terms for this cluster being associated with rRNA metabolic process and protein targeting to ER. While clusters 5 and 10 were predicted as Ductal cells, they were enriched processes involving MHC II protein complex binding, which is a known function of macrophages, so our predictions for these clusters may be inaccurate.

CONCLUSION

Given the lack of a complete dataset, combined with a lack of homogeneity in the processing of single-cell RNA-seq data, we cannot conclude that we were able to reproduce the findings of Baron et al with certainty. Through collaboration, we were able to utilize the tools and packages given to us to produce high-quality results that seemed to match a majority of the findings of the original paper. Using dimensional analysis and PCA, we were able to obtain 13 of the 14 clusters found in the original paper.

Using gene set enrichment analysis, we were able to confirm our label predictions for seven out of ten clusters with five cell types: alpha, beta, cytotoxic T, macrophages, and vascular cells. Due to the low number of genes in some of our clusters and the presence of mitochondrial and ribosomal genes, three of our clusters were not enriched for our predicted cell types.

We believe that we were able to reproduce the results from the Baron et al. paper with some degree of success; however, we analyzed only a subset of the data from Baron et al. If we performed our analyses on the full dataset from this paper, perhaps there would have been less of a discrepancy in our results. Additionally, due to issues running Salmon Alevin, a likely source of error is the whitelist of barcodes allowing substantially more reads with lower counts. While a good amount of these are filtered out further down the pipeline, these certainly contribute to confounding the results as it mitigates the ability to provide cellular heterogeneity.

REFERENCES

[1]Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K., Surani, M.A., 2009. mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods 6, 377–382. https://doi.org/10.1038/nmeth.1315

[2]Aldridge, S., Teichmann, S.A., 2020. Single cell transcriptomics comes of age. Nat. Commun. 11, 4307. https://doi.org/10.1038/s41467-020-18158-5

[3] 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. PMID: 27667365

[4]Srivastava, A., Malik, L., Smith, T., Sudbery, I., Patro, R., 2019. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. Genome Biol. 20, 65. https://doi.org/10.1186/s13059-019-1670-y

[5]Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D766–D773, https://doi.org/10.1093/nar/gky955

[6] "Understanding UMAP." *PAIR Page Redirection*, pair-code.github.io/understanding-umap/.

[7] "DimPlot: Dimensional Reduction Plot." *RDocumentation*, www.rdocumentation.org/packages/Seurat/versions/4.0.1/topics/DimPlot.

[8] "Seurat, Jean-Pierre." *Benezit Dictionary of Artists*, 2011, doi:10.1093/benz/9780199773787.article.b00168110.

[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(3):e90 (2021).

[10] Hoffman, Paul. "Tools for Single Cell Genomics [R Package Seurat Version 4.0.1]." *The Comprehensive R Archive Network*, Comprehensive R Archive Network (CRAN), 18 Mar. 2021, cran.r-project.org/web/packages/Seurat/index.html.