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

Group Roles: Lucas Zhang (Data Curator), Benyu Zhou (Programmer), Mano Ranaweera (Analyst), Sriramteja Veerisetti (Biologist)

Introduction

The study of pancreatic cells has great clinical importance to the study of Type ID, Type II Diabetes Melitus, pancreatitis, and cancer as the pancreas plays an important role in energy homeostasis achieved by secreting digestive enzymes and metabolic hormones. Thus, the goal of the original paper was to utilize a novel technique in order to study human and mouse pancreatic cells in high resolution. Characterizing the transcriptomes of individual cells has typically been done using RNA-seq due to its ability to detect transcripts on a large genomic scale, but challenges present itself when trying to detect human pancreatic cells at a high throughput scale. Klein et al. presents the inDrop platform as an unique solution in order to enable high resolution cell detection without the need of pre-sorting, utilizing high-throughput microfluidics to barcode RNA from thousands of its constituent cells.

This project seeks to replicate certain elements of the original paper using sequencing data from a single human donor in order to perform cell-by-gene quantification and quality control on UMI counts as well as identify clusters marker genes for distinct cell type populations.

Data

The data used in this analysis was from a human female donor with an age of 51 years, part of GEO (Gene Expression Omnibus) sample GSM2230758, and located in samples SRR3879604, SRR3879605, and SRR3879606. This sequencing data was obtained using the inDrop platform, a high-throughput droplet microfluidics system which barcodes the RNA from thousands of individual cells while utilizing a linear amplification method for single cell transcriptome library construction. The inDrop method utilizes the Illumina HiSeq 2500 instrument, the Illumina platform, and a RNA-seq assay. The sequencing data was generated with the human reference genome GR38.82, producing paired end reads for downstream analysis.

Methods

Preprocessing the human female sample data began with identifying the SRR runs associated with this donor in the SRA database (Sequence Read Archive). The resulting files contained compressed paired-end fastq reads as well as a read file containing the UMI barcodes. The first step in processing this data was counting the number of reads per distinct barcode, generating a cumulative distribution graph displaying the distribution of read counts. The barcodes were then filtered by read counts based on the results of the cumulative distribution graph, with the cutoff being around 700,000 reads as this was approximately the peak of the slope of the distribution graph. The number of barcodes after read count filtering was approximately three thousand, which were written to a text file to be used with Salmon Alevin (v1.1.0).

The next step was producing a transcript ID map to be used to identify the transcripts during the Salmon Alevin run, which was obtained through Ensembl.org using the latest version of human reference genome GR38. At the same time, a Salmon index was created for the run by obtaining the human map transcript file from the Gencode website and using the index function of the Salmon suite.

After producing the index, transcript map, and Salmon index, Salmon Alevin was run to create the UMI counts matrix, with additional options specifying barcode length, UMI length, and the position of the UMI sequence in order to use a custom whitelisted barcode file as input.

The programming session is mostly the application of Seurat package to process the Alevin UMI matrix and cluster the cell type subpopulations (Hao and Hao et al., 2021, [Seurat V4]). Another important package used here is tximport, which can read the alevin type UMI matrix file into a large list in R (Soneson C, Love MI, Robinson MD, 2015). The matrix count in the list was what was used to create the Seurat object for processing and clustering. When creating the Seurat object, cell minimum of 3 was set so that features that only existed in individual cells were not included.

With the UMI matrix imported and transformed into a Seurat object, multiple filtering steps were performed. First, mitochondria percentage was calculated because low-quality or dying cells often exhibit extensive mitochondrial percentage. So, cells with a mitochondrial percentage over 5% were filtered out. Also, cells that have unique feature counts over 2,500 or less than 200 were also filtered out as those are low-quality cells or cell multiplets.

After the low quality cells were filtered out, feature variance was examined. Using the FindVariableFeatures function in the Seurat package, top 2000 features (genes) were identified to carry on to the PCA analysis. Before the PCA analysis, necessary normalization and scaling were performed on the Seurat object. With the dimension cut-off identified by PCA, UMAP was used to cluster the cells and visualize the dataset. The UMAP graph and a pie chart displaying the number of cells in each cluster were produced.

The Seurat package was also used to find marker genes for each of the clusters from the GSM2230760 RDA file. The data from the file was precomputed. Each cluster is intended to be identified by cell type, labeled by referring to the Panglao database to match the top marker genes found per cluster to the cell type found in the database. Top marker genes considered for each cluster were found in terms of the top log2 fold change values.

The clusters were then visualized using the UMAP projection method, with each cell type being distinguished, and the top 2 expressed genes in each cluster were visualized on a heatmap to see differences in expression levels across different clusters, with the corresponding clusters seen on top of the map. UMAP reduction has the purpose of reducing dimensions of the data, similar to the tsne method.

With each cluster being labeled by a cell type, some other genes were found in each cluster that were also differentially expressed, but can still be related to the cell type based on the function of that gene. Additional research was done by referring to other published papers to find what novel marker genes there could be.

In the biologist section, the goal was to utilize the marker genes obtained through the unbiased clustering process in order to confirm the cell type and function of the different clusters. A gene set enrichment analysis was performed on the marker genes in each cluster via the bioinformatics tool DAVID. The marker genes were filtered via two criteria: p_val_adj < 0.05 and avg_log2FC > 0. The generated GO terms were then compared to the cell type labels for each cluster.

Results

Important run statistics were gathered for the Salmon Alevin run, such as mapping rate, reads discarded, and number of transcripts. A total of 108296 transcripts were found by Alevin, with a total of 2837 barcodes of which 48.41% reads were discarded due to noisy barcodes. The mapping rate was also 26.843%.

16256 features across 2540 samples existed in the seurat object before filtering. Figure 1 and Figure 2 are visualizations of the quality control metrics. With a strict filtering criteria presented in the Data curator part, the feature counts and RNA counts mostly fall into the good sample metrics (>200 and <2500 unique features). The mitochondrial percentages are very low (close to 0) across cells. There is no correlation between RNA counts and mitochondrial percentage observed. A strong correlation between RNA counts and unique features per cell is observed, which indicate good quality.

Both FIgure 1 and Figure 2 show that the dataset provided by the data curator had good quality even before filtering.

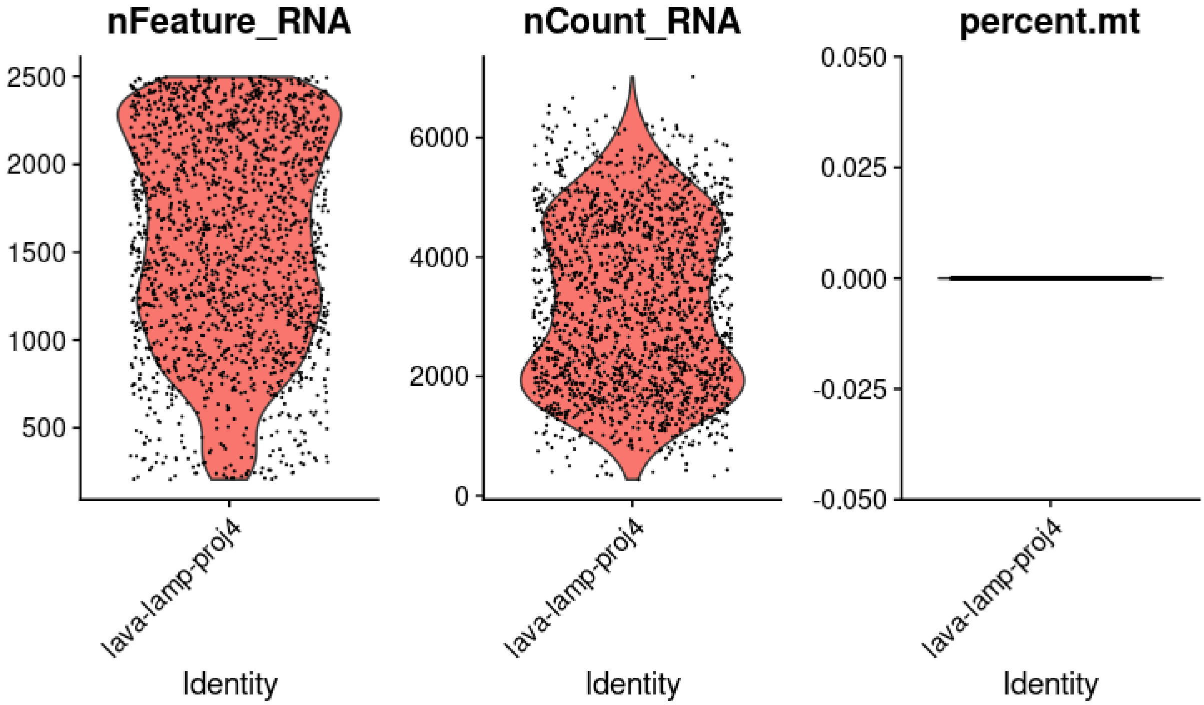

Figure 1 violin plot for feature numbers, RNA counts and mitochondria percentage

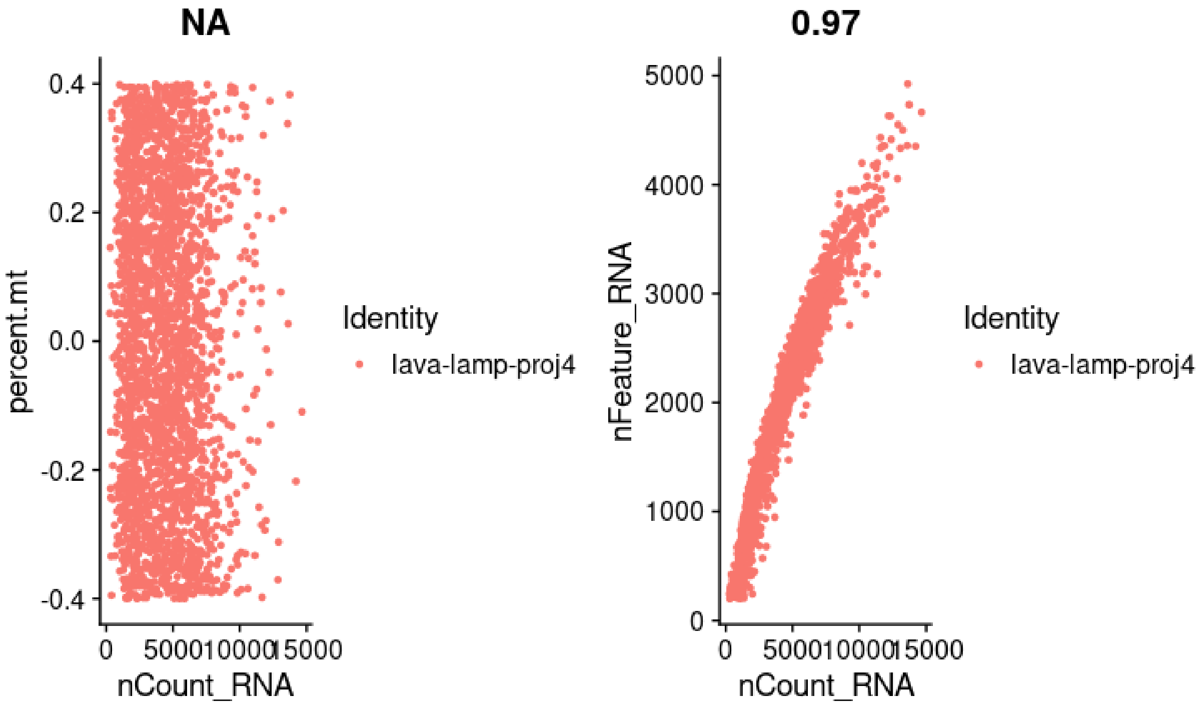

Figure 2 Scatterplot visualizes correlation between RNA counts and mitochondrial percentage (left). Scatterplot visualizes correlation between RNA counts and unique feature numbers in a cell (right).

After the first filteration out of low-quality cells, 16256 features across 1806 samples were left. Then feature variance is checked using the FindVariableFeatures function in the Seurat package. Top 2000 features were chosen by default. From figure 3 below, it can be seen that the top 2000 features (genes) indeed have significant variance above 0 and should be enough for the PCA analysis. Top 10 most highly variable genes are also displayed in table 1.

Rank	1	2	3	4	5	6	7	8	9	10
Gene symbol	REG1A	REG3A	COL1A1	CTRB2	PPY	REG1B	COL3A1	PRSS3P2	IGFBP5	COL1A2

Table 1 Top 10 most highly variable genes

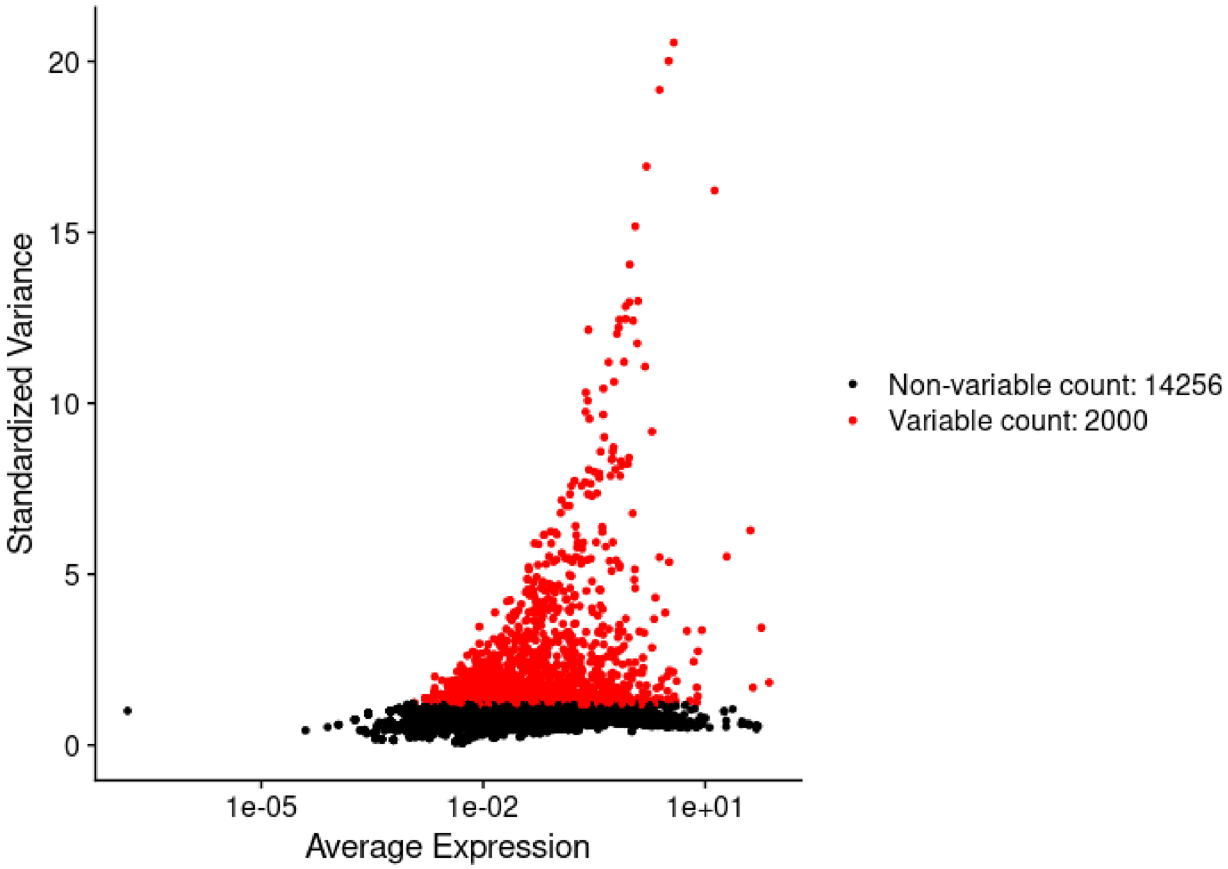

Figure 3 Scatter plot displaying variable features. X-axis is Average Expression and Y-axis is Standardized variance.

The top 20 principal components identified by PCA are visualized in JackStaw plot and ElbowPlot in figure 4 and 5 below. From figure 4, it can be seen that a significant p-value gap occurred at PC15, but weirdly the p-value became smaller than PC15 for PC 16-18. This could be due to certain features related to those 3 PCs being rare but significant. Same trend can be also observed in the Elbow plot. Additionally, it can be observed in the Elbow plot that the first 8 PCs have the most rapid standard deviation decrease, although the decrease from PC9 to PC19 was also significant. So the dimension cut-off can be 15 or 19.

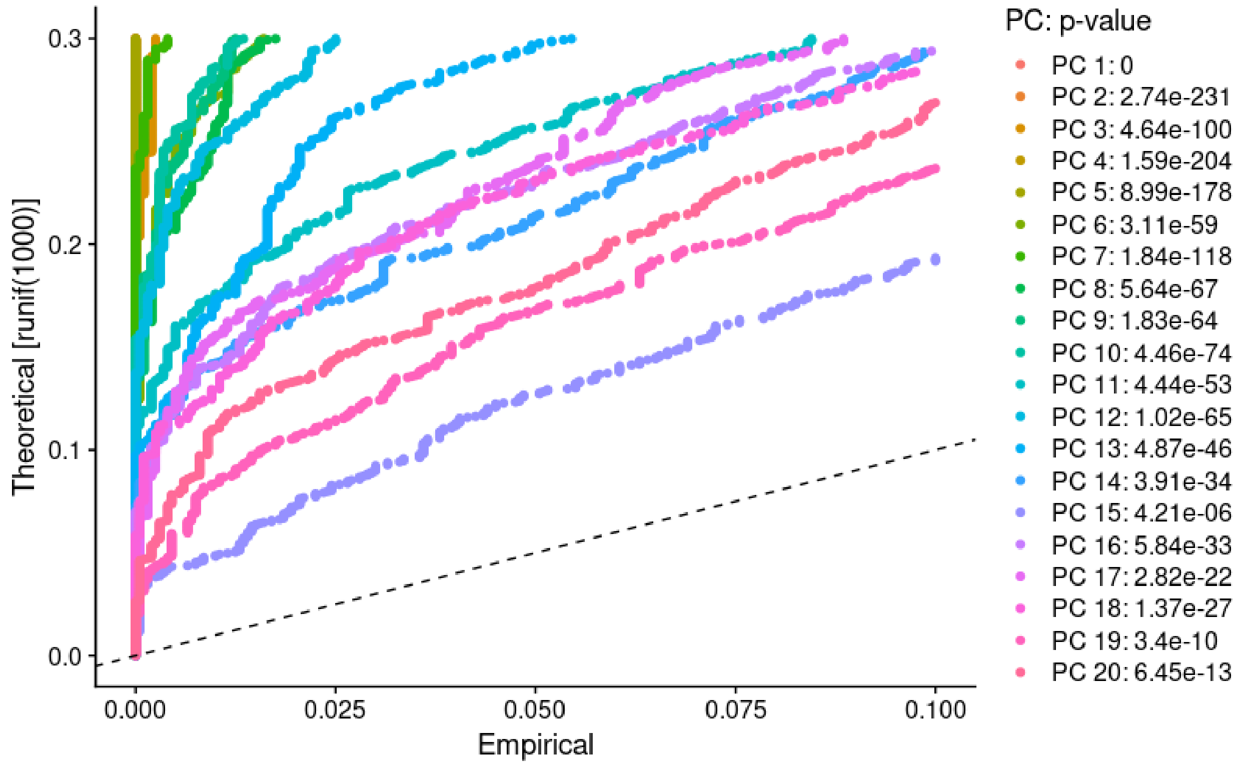

Figure 4 JackStrawPlot comparing p-value for each principal component.

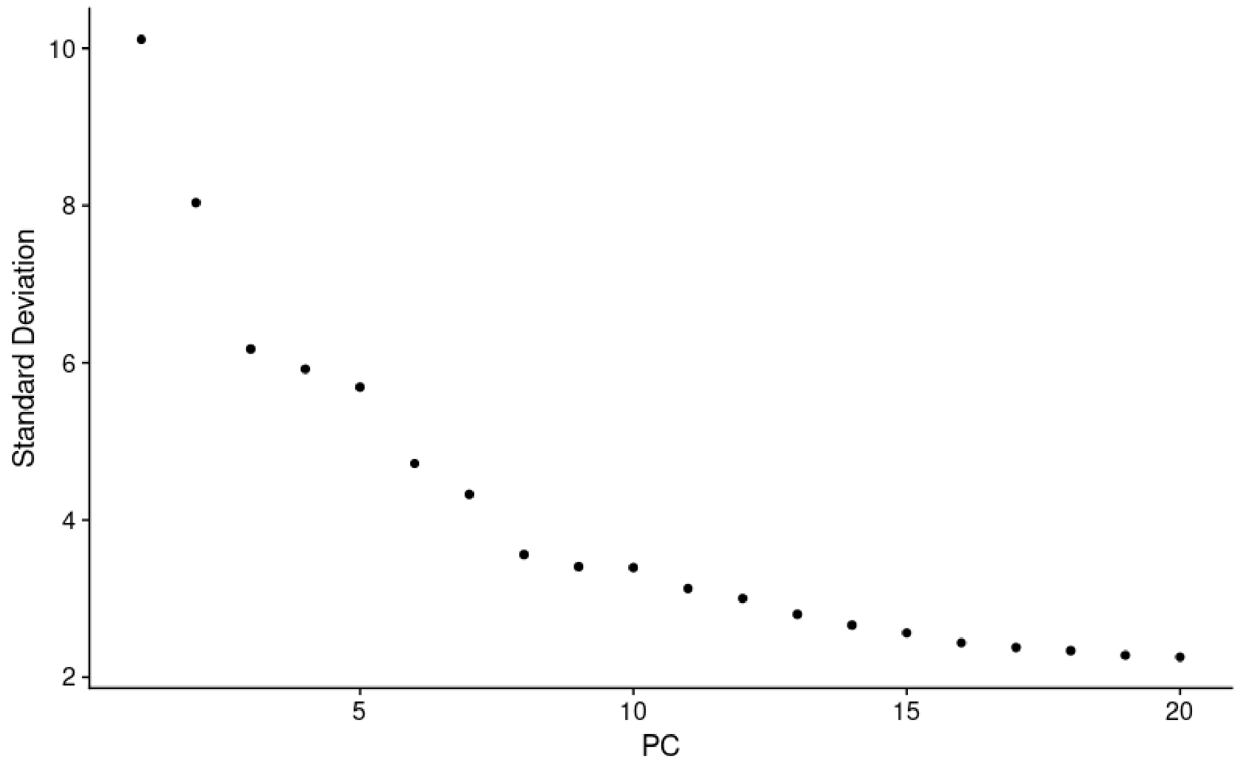

Figure 5 Elbow plot showing the standard deviation of top 20 principal components.

Figure 6a and 6b are two UMAP dimension reduction plots using 19 and 15 as dimension cut-off respectively. It can be seen that both plots identified 12 clusters with clear separation. There are no big differences between the two plots except some positions change. The number of cells identified in each cluster was exactly the same for both cut-off choices. In this case, dimension 15 was chosen in the end to be used for later analysis since fewer dimensions can have better performance efficiency. Figure 7 is a pie chart displaying the number of cells in each cluster and each cluster's percentage using 15 as the dimension cut-off. 12 cell clusters of 1806 cells is quite different from the paper's 15 clusters from 12000 cells. Analysis methods difference (Seurat vs custom analysis methods by the paper author) definitely could be a big reason. This could also be due to the fact that only human cells from only the 51 year old female donor were included in the data curator part. The strict cut-off choice in the data curator part, which is the reason for good dataset quality and low cell numbers before filtration, could contribute to the difference as well.

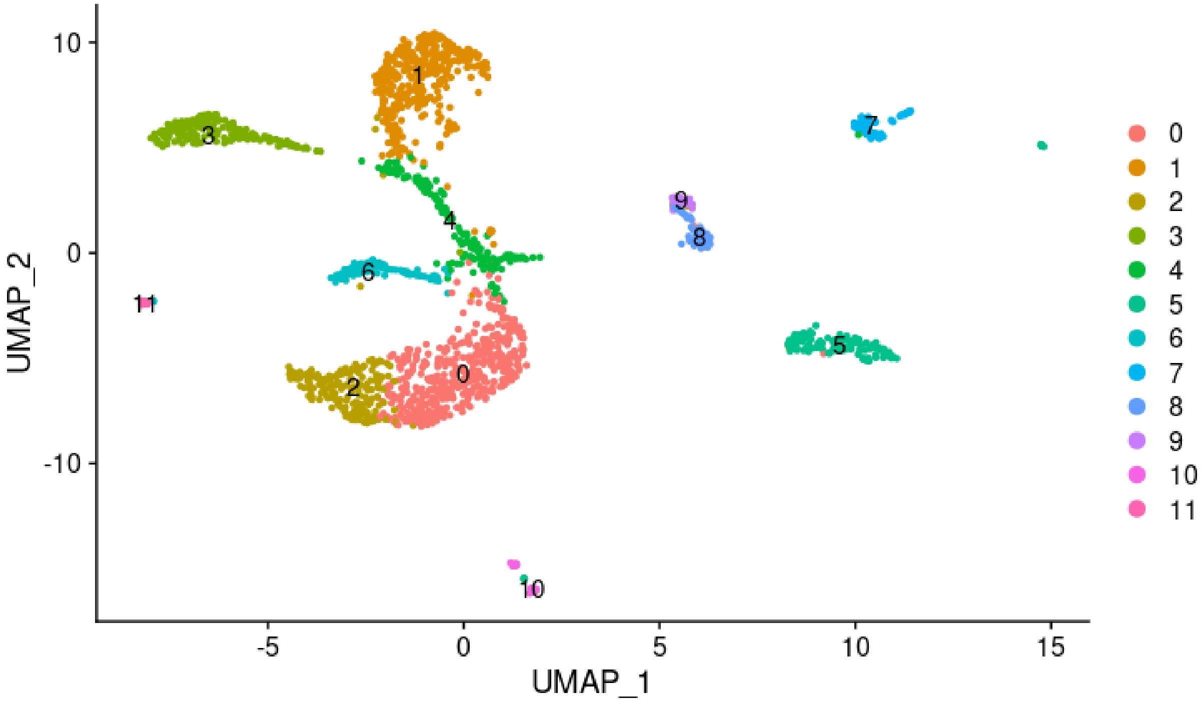

Figure 6.a UMAP dimension reduction plot using 19 dimensions as a cut off

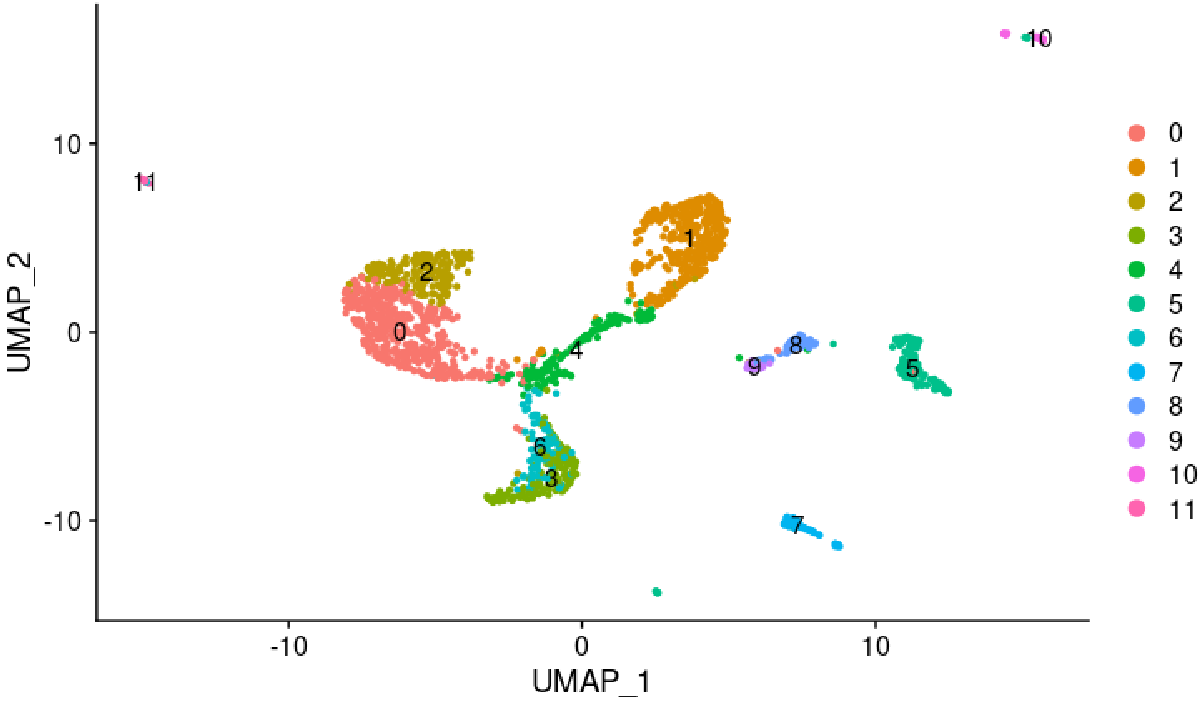

Figure 6.b UMAP dimension reduction plot using 15 dimensions as a cut off

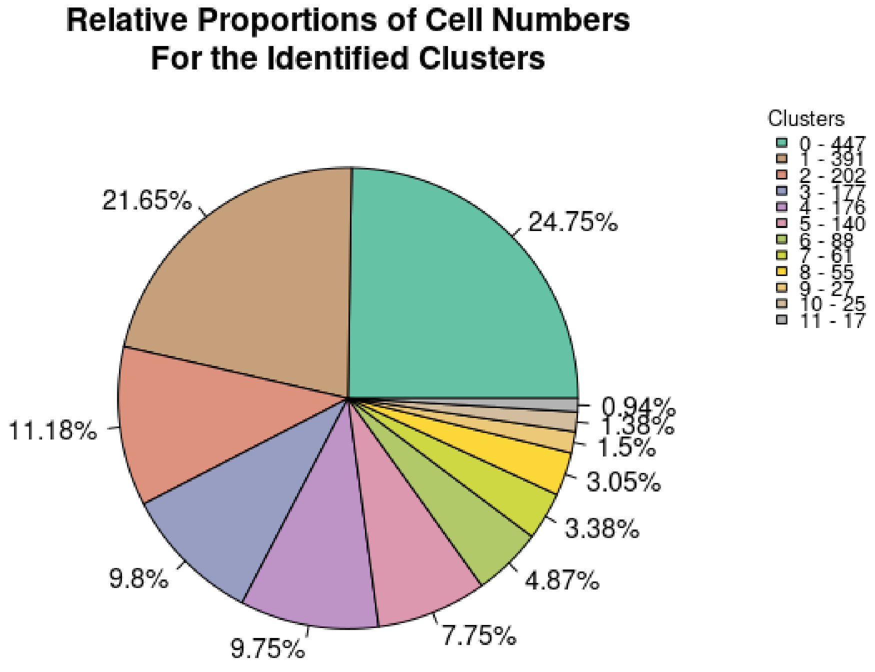

Figure 7 Pie chart showing the the relative proportions of cell number in each cluster

Compared to the paper's plot, the respective clusters for each cell type are crammed together, as seen in Figure 8 below. More marker genes are represented in this plot that could be causing this, as well as the names of the cell types found possibly being different. The cluster marker genes found had many differences or disagreements with what was seen in the paper. The differences are possibly due to the varied filtering methods. Altering the minimum fold change threshold or p-value threshold to filter the marker genes could make the result more like what was seen in the paper. All cluster could be labeled using the marker genes.

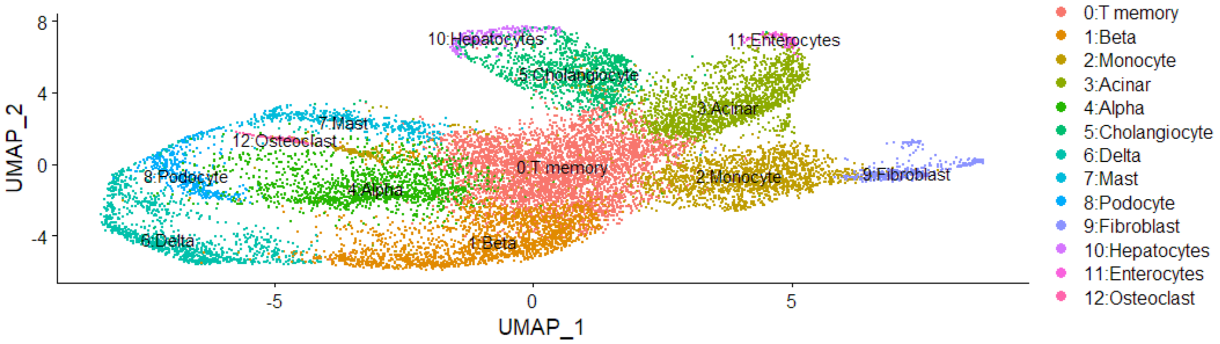

Figure 8. UMAP projection method to visualize clusters.

The heatmap seen below in figure 9 contained the top expressed genes in each cluster, with the gene names on the vertical axis and each cluster labeled on the horizontal axis. Expression levels might seem a little random in each of the clusters, but the genes expected to have expression in each cluster are still seen to have at least some expression. Marker genes found in each cluster were most likely filtered differently from the paper as mentioned before, so the heatmap has a lot of different genes shown. This could also be the cause of more randomness in expression.

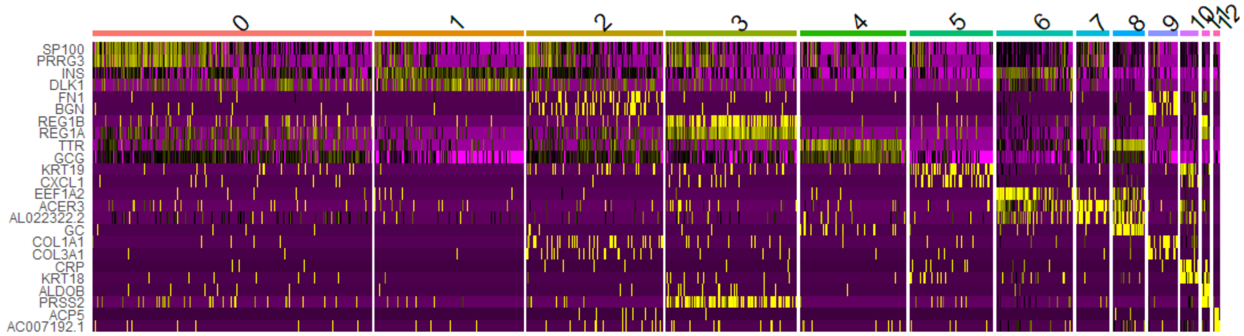

Figure 9. Heatmap of top 2 genes per cluster.

Table 2 below shows the novel genes found in each cluster that are discriminative of cell type as the main marker genes. The genes found have functions that are related to the functions or characteristics of each cell type.

Cluster	**Cell Type**	**Novel Genes**
0	**T-Memory cell**	**HLA-C**
1	**Beta**	**NK6-1**
2	**Monocyte**	**NFKBIA.2**
3	**Acinar**	**SOX4**
4	**Alpha**	**MAFB**
5	**Cholangiocyte**	**CFTR**
6	**Delta**	**PRG4**
7	**Mast**	**CCL5**
8	**Podocyte**	**TUBB**
9	**Fibroblast**	**IL6**

10	**Hepatocyte**	**CYP3A5**
11	**Enterocyte**	**CCL2**
12	**Osteoclast**	**TGFBI**

Table 2. Marker genes found for each cluster.

		Unfiltered		Filtered (Via p_val_adj < 0.05 and avg_log2FC > 0)	
Cluster #	Label	Marked Genes	Top 3 Biological Processes (BP)	Marked Genes	Top 3 Biological Processes (BP)
0	T memory	252	SRP-dependent cotranslational protein targeting to membrane, cytoplasmic translation, translational initiation	36	mitochondrial respiratory chain complex I assembly, mitochondrial electron transport, NADH to ubiquinone, electron transport coupled proton transport
1	Beta	423	Acetylation, neutrophil degranulation, translational initiation	98	protein localization to secretory granule, insulin secretion, virion assembly

2	Monocyte	356	SRP-dependent cotranslational protein targeting to membrane, cytoplasmic translation, viral transcription	55	extracellular matrix organization, collagen fibril organization, endodermal cell differentiation
3	Acinar	529	SRP-dependent cotranslational protein targeting to membrane, cytoplasmic translation, viral transcription	115	SRP-dependent cotranslational protein targeting to membrane, cytoplasmic translation, viral transcription
4	Alpha	327	SRP-dependent cotranslational protein targeting to membrane, cytoplasmic translation, translational initiation	51	insulin processing, cellular protein metabolic process, negative regulation of endopeptidase activity
5	Cholangiocyte	508	neutrophil degranulation, negative regulation of apoptotic process, response to drug	161	cell adhesion, negative regulation of apoptotic process, neutrophil degranulation
6	Delta	1900	SRP-dependent cotranslational protein targeting to membrane, translational initiation, cytoplasmic translation	1430	SRP-dependent cotranslational protein targeting to membrane, translational initiation, cytoplasmic translation
7	Mast	1418	SRP-dependent cotranslational protein targeting to membrane, translational initiation, cytoplasmic translation	593	electron transport coupled proton transport, ATP synthesis coupled

					electron transport, aerobic respiration
8	Podocyte	2542	SRP-dependent cotranslational protein targeting to membrane, translational initiation, cytoplasmic translation	2021	SRP-dependent cotranslational protein targeting to membrane, translational initiation, cytoplasmic translation
9	Fibroblast	1343	translational initiation, SRP-dependent cotranslational protein targeting to membrane, cytoplasmic translation	1043	translational initiation, SRP-dependent cotranslational protein targeting to membrane, cytoplasmic translation
10	Hepatocytes	2921	translational initiation, SRP-dependent cotranslational protein targeting to membrane, neutrophil degranulation	2362	translational initiation, SRP-dependent cotranslational protein targeting to membrane, cytoplasmic translation
11	Enterocytes	2214	SRP-dependent cotranslational protein targeting to membrane, translational initiation, cytoplasmic translation	1273	SRP-dependent cotranslational protein targeting to membrane, translational initiation, cytoplasmic translation
12	Osteoclast	898	neutrophil degranulation, antigen processing and presentation of	422	neutrophil degranulation,

			exogenous peptide antigen via MHC class II, membrane organization		immune response, cytokine-mediated signaling pathway

Table 3. Gene Set Enrichment Analysis on Marker Genes. The table summarizes the cluster value, cell type, number of marked genes before and after filtering, and the top 3 GO Biological Process (BP) terms for each cluster. The marked genes were filtered via two criterias: p_val_adj < 0.05 and avg_log2FC > 0.

Cell Type	Generic Function of Cell Type
T Memory	These are long-term antigen specific T cells that will remain within the body even after an infection has been eliminated.
Beta	These are cells that regulate insulin levels within the body. Insulin is a hormone that is responsible for controlling the glucose level within blood.
Monocyte	These are immune cells made within the bone marrow. These cells travel via the bloodstream to become a dendritic cell or a macrophage.
Acinar	These cells store, synthesize, and release digestive enzymes.
Alpha	These are endocrine cells that release glucagon so that glucose levels within the bloodstream will increase.
Cholangiocyte	These are epithelial cells that have an important function in bile composition as well as solute

	transport.
Delta	These cells are responsible for the production of somatostatin cells.
Mast	These are white blood cells that play a role in inflammation.
Podocyte	These are epithelial cells that are seen on the surfaces of glomerular capillaries.
Fibroblast	These cells are responsible for the development of connective tissue.
Hepatocytes	These cells are located in the liver and are key for detoxification, metabolism, and even protein synthesis.
Enterocytes	These are cells located in the nucleus. They are key in absorbing molecules from the lumen and transporting them to tissue via the blood vessels.
Osteoclast	These are cells that will spark bone remodeling by degrading previous bones. They also play a mediation role in bone loss.

Table 4. Summarization of the Cell Types and their corresponding functions.

Discussion

To begin, it is necessary to compare the cell type results that were produced in the Baron et al paper to the cell type results that were produced by the group. After thorough comparison, unfortunately only 5 out of the 13 clusters were similar between the paper and the group's results. In order to determine the biological functions of the gene clusters, the DAVID bioinformatics tool was utilized. On a deeper level, of the 5 clusters that were similar between the paper and the results, only 2 were accurately confirmed via the gene set enrichment analysis on the marker genes. Beta cells are cells that are responsible for the regulation of insulin within the body. Insulin is a type of hormone that is key for regulating glucose concentration within the blood (Table 3). Based on the enrichment analysis it was reassured that cluster 1 fell under the umbrella of Beta cell type because the top three Gene Ontology (GO) biological process terms correlated to insulin secretion. On the other hand, Alpha cells are a type of endocrine cells that regulate glucose levels by secreting glucagon. The enrichment analysis that was performed via DAVID reassures the fact that cluster 4 pertains to Alpha cells because the top three GO biological process terms correlated to insulin processing and metabolic processing.

Moreover, there were 11 cell type labels that were not reassured via the DAVID gene set enrichment analysis on the marker genes. The genes corresponding to cluster 0 were deemed to fall under the T memory cells type, however, the GO terms do not support this classification. Memory T cells are antigen specific cells that will remain in the body even after an infection has been eradicated so that they can fight future similar infections. The expectation was to find GO terms that deal with immunological processes, however, the group's analysis yielded GO terms that deal with the electron transport chain and the mitochondria itself. Monocytes are immune cells produced within the bone marrow. In theory, the group expected to see GO terms that pertain to immunological pathways, similar to memory T cells, instead the terms dealt with the extracellular matrix as well as endodermal cell differentiation. The clusters pertaining to Acinar, Delta, Podocyte, and Enterocyte cells all yielded very similar GO terms: SRP-dependent cotranslational protein targeting to membrane, translation initiation, and cytoplasmic translation. Unfortunately, these GO terms do not match the functionalities of these cell types. Acinar cells are responsible for the storage, synthesis, and secretion of digestive enzymes, while Delta cells are key for the formation of somatostatin cells. Podocyte cells are epithelial cells on the surface glomerular capillaries and Enterocyte cells are located within the nucleus and are key in the absorption of molecules from the lumen and the transportation of molecules to tissue via blood vessels.

Onwards, the genes under the assumed Fibroblast and Hepatocyte cell types generated identical GO terms that include: translation initiation, SRP-dependent cotranslational protein targeting to membrane, cytoplasmic translation. Fibroblast cells are critical in the development of connective tissue, while Hepatocyte cells, located in the liver, are responsible for metabolism, detoxification, and protein synthesis. Based on the functions of these two cell types, it is clear that the GO terms produced via DAVID do not match the functionalities of these cell types. The final three clusters deal with Cholangiocyte, Mast, and Osteoclast cells. Cholangiocyte cells are epithelial cells with roles in solute transportation and bile formation, Mast cells are white blood cells with an integral role in inflammation, and Osteoclast cells mediate bone loss and degrade previous bones so they can be remodeled. After analyzing the corresponding GO terms to the functions of the corresponding cells it can be assumed that the cell label terms may not be the most accurate representation of their corresponding clusters.

Overall, there were clear discrepancies between the determined cell type and GO terms generated. After analyzing the genes, it is clear that a single gene may be found in multiple cell types, which can lead to errors when classifying the gene into a single cell type. For finding novel genes however, there still seemed to be genes found for each cluster that can be confidently concluded as discriminative of cell type. Due to this requiring intensive research, there likely were many other genes in each cluster not mentioned in the table that could also be considered novel marker genes.

Furthermore, Single cell RNA-Sequencing is a relatively new technology so there are no clear cut efficient algorithms to classify a gene into a particular cell type, although that is something being worked on. Future advancements in single cell RNA-Sequencing will lead to more accurate and precise classifications of genes. Finally, having more in-depth knowledge of the cell types, may make the classifications of these genes much easier.

Reference

Baron, Maayan et al. “A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure.” *Cell systems* vol. 3,4 (2016): 346-360.e4. doi:10.1016/j.cels.2016.08.011

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

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets--update.

Fiona Cunningham et al. Ensembl 2022, *Nucleic Acids Research*, Volume 50, Issue D1, 7 January 2022, Pages D988–D995, https://doi.org/10.1093/nar/gkab1049

Frankish et al. GENCODE 2021, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D916–D923, https://doi.org/10.1093/nar/gkaa1087

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.

Hao and Hao et al. Integrated analysis of multimodal single-cell data. Cell (2021) [Seurat V4]

{ author = {Yuhan Hao and Stephanie Hao and Erica Andersen-Nissen and William M. Mauck III and Shiwei Zheng and Andrew Butler and Maddie J. Lee and Aaron J. Wilk and Charlotte Darby and Michael Zagar and Paul Hoffman and Marlon Stoeckius and Efthymia Papalexi and Eleni P. Mimitou and Jaison Jain and Avi Srivastava and Tim Stuart and Lamar B. Fleming and Bertrand Yeung and Angela J. Rogers and Juliana M. McElrath and Catherine A. Blish and Raphael Gottardo and Peter Smibert and Rahul Satija},

title = {Integrated analysis of multimodal single-cell data},

journal = {Cell},

year = {2021},

doi = {10.1016/j.cell.2021.04.048},

url = {https://doi.org/10.1016/j.cell.2021.04.048},}

Seurat - guided Clustering tutorial - satija lab. (n.d.). Retrieved May 2, 2022, from https://satijalab.org/seurat/archive/v3.1/pbmc3k_tutorial.html

Database, G. C. H. G. (n.d.). *Genecards®: The Human Gene Database*. GeneCards. Retrieved May 1, 2022, from https://www.genecards.org/

PanglaoDB - a single cell sequencing resource for gene expression data. (n.d.). Retrieved May 2, 2022, from https://panglaodb.se/

Uniform manifold approximation and projection for dimension reduction¶. UMAP. (n.d.). Retrieved May 1, 2022, from https://umap-learn.readthedocs.io/en/latest/