Single Cell RNA-Seq Analysis of Pancreatic Cells

Data Curator: Teresa Rice
Programmer: Arushi Shrivastava
Analyst: Maha Naim

INTRODUCTION

The pancreas is an essential organ that belongs to your digestive system and endocrine system. It is responsible for producing a cocktail of enzymes, and a handful of hormones the digestive system utilizes. A malfunctioning pancreas can induce several disease states such as type 1 (T1D) and type 2 diabetes mellitus (T2D), pancreatitis, and cancer [1]. Characterizing the distinct cell types of the pancreas and their respective transcriptomes help illuminate its complex nature. The article, *A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure,* approaches this task using single-cell sequencing methodology. Their RNA-seq gene expression profile data revealed 15 cell-type clusters with distinct expression profiles. Additionally, researchers were able to identify distinct differences in gene regulation for maturation and ER stress levels. My team will look at a subset of their RNA-seq data and perform similar analyses. We aim to replicate a portion of their steps to produce data describing the cell clusters and highly expressed genes.

DATA

The study we are trying to replicate utilized over 12,000 individual pancreatic cells obtained from four human donors and two mouse strains for a total of 13 sequencing libraries [1]. Single-cell transcriptomics of these cells were determined using illumina's inDrop, a high-throughput single-cell labeling method similar to Drop-seq that uses hydrogel microspheres to introduce the oligonucleotides [3].

We will only be looking at the cells from the 51-year-old human female donor, otherwise referred to as donor 2 throughout the paper, for a total of 3 sequencing libraries. Libraries were gathered from the Gene Expression Omnibus repository, accession number GSE84133. After examining the human samples, it was determined that the GSM2230758 human pancreatic islets, (sample 2) was our female donor [7]. The SRA Run Selector on the GEO repository was used to identify the following files as the runs for the 51-year-old female: SRR3879604, SRR3879605, and SRR3879606.

SRR files corresponding to the donor were provided by Boston University on their Shared Computing Cluster (SCC). Each SRR number had three files associated with it including forward and reverse FASTQ files, and a preprocessed file of read 1. Here the barcode and UMI for each transcript were in a more manageable format. Human reference GRCh38.p13 transcript sequences were downloaded from Gencode.

METHODS

The preprocessed SRR file described above was used as input to count the number of reads by barcode. Our team developed a **python** script that would read the pre-processed gzipped FASTQ SRR file, identify lines with barcode identifiers, and add them to a dictionary. Once the script was run on all three files, the unique number of barcodes, total number of repeat barcodes, and total number of barcodes per sample were recorded (Table 1). It was assumed there would be some barcodes shared between files so a second **python** script was developed in order to find duplicate barcodes and combine the counts, producing a single dictionary for barcodes and their respective counts. To understand how the reads were distributed

among the barcodes, a cumulative distribution plot was produced with R package **ggplot2**, utilizing the **ecdf** function (Figures 1&2).

Barcode Count Metadata			
Sequence File/Run	**Unique Barcodes**	**Repeat Barcodes**	**Total Barcodes**
SRR3879604	2,018,986	562,207,073	564,226,059
SRR3879605	2,031,816	390,485,663	392,517,479
SRR3879606	1,911,977	366,182,446	368,094,423

Table 1. Barcode count metadata from each sequence file for indicating unique count of barcodes, total count of repeat barcodes, and total barcodes.

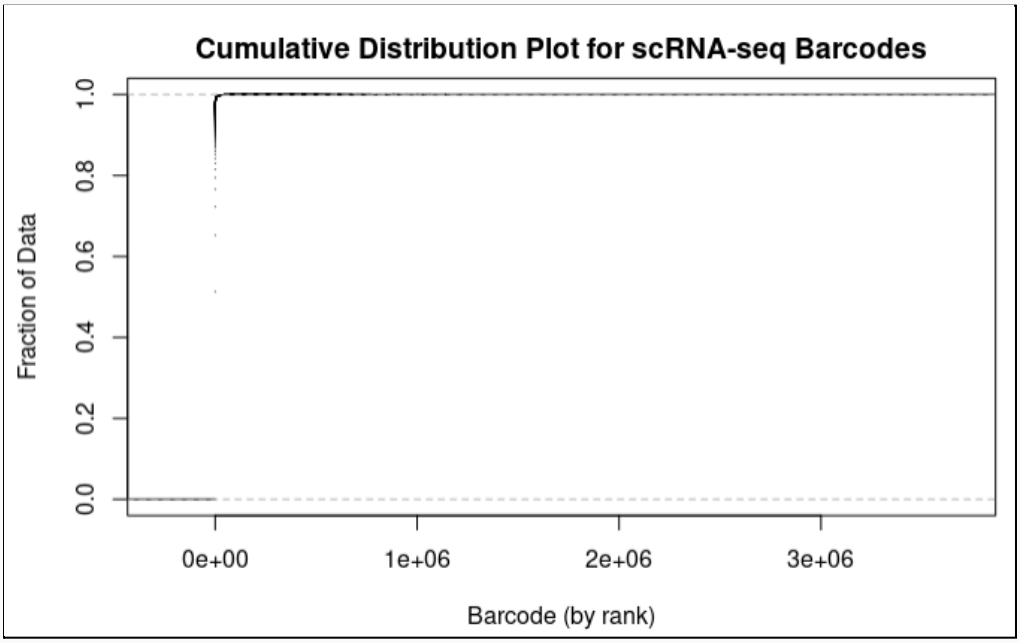


Figure 1. Cumulative distribution of all barcode counts displays the proportions of barcodes to help understand how the reads are distributed.

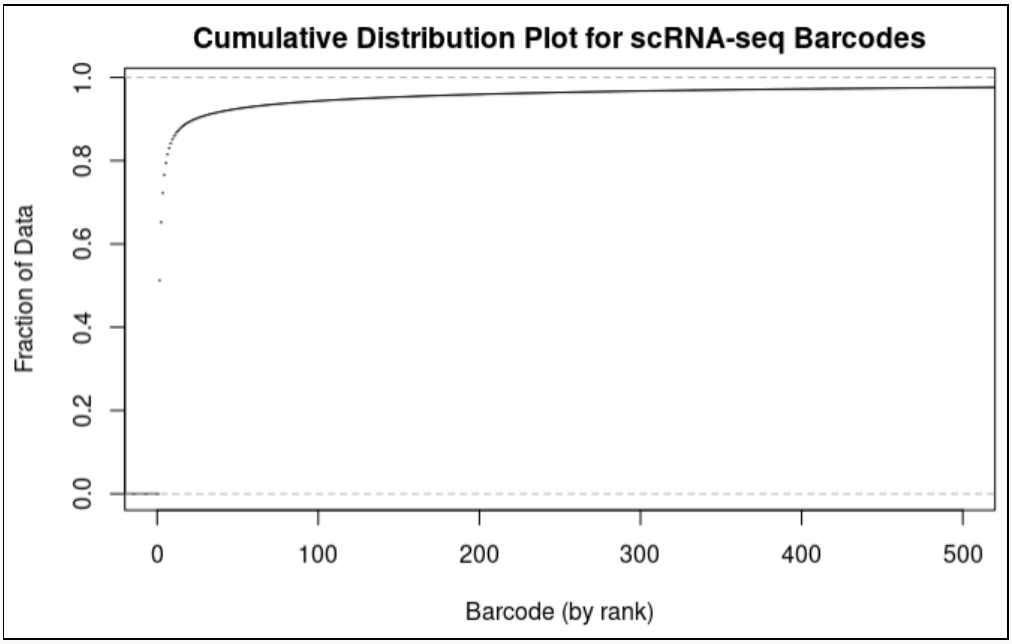


Figure 2. Zoom-in of Cumulative distribution reveals most variability in barcode counts occurs in the first 50-100 barcodes.

Barcodes with a higher frequency were added to a whitelist. To determine a read count threshold for the whitelist a zoomed-in portion of the Cumulative Distribution Plot is useful (Figure 2). This plot reveals most of the variation in barcode read frequency is within approximately the first 50-100 barcodes. Three tables were produced using thresholds counts of 10, 100, 1000. The table with the 10 count threshold was not stringent enough as 10 counts for a read is too low and would create too much noise. The 1000 count threshold was much more stringent, but may have excluded informative reads. After re-examining the cumulative distribution, the 100 threshold was deemed most appropriate as it would include the most informative reads without excess.

Salmon is a software that can be used to quantify RNA-seq data with its many tools. **Alevin** is the tool we used from **salmon** in order to analyze the single-cell data and generate a cell-by-gene count matrix. The **salmon** software version available on the SCC was not up to date, instead **salmon (version 1.4.0) [5]** was downloaded from the Combine-lab's github page. In order to use **salmon** on our data we needed to make a "genemap" file that converted the transcript ID into the gene ID (ENST to ENSG) using transcript files from Gencode. Another essential input for **alevin,** a salmon index, was also generated utilizing the transcript files from Gencode. The preprocessed forward FASTQ file, reverse FASTQ read file, genemap file, salmon index, and whitelist previously described were input for the **salmon alevin** analysis.

Generating the barcodes dictionaries and running salmon were two time-consuming steps that each used large zipped files. In order to minimize how long the barcode counting step would take different tools were tested. Ultimately, instead of reading the entire preprocessed FASTQ files into memory, a combination of **io.TextIOWrapper** and **io.BufferedReader** was utilized to read the lines from the gzipped files into memory one at a time. Since the barcodes needed to be gathered from each SRR preprocessed FASTQ file, the script was broken into 3 pieces instead of one in order to run the **python**

scripts in parallel, reducing the time by approximately ⅓. However, despite these modifications, this step and running **salmon alevin** took a total of nearly 3 hours when run back to back. However, once completed, a few of the output files include a UMI counts matrix for further analysis and a summary of the mapping statistics (Table 2).

Processing the UMI counts matrix

Seurat is the most popular R package (version 4.0.1) for single-cell RNA-seq data analysis. Initially, it was popular as a clustering tool, but at present it is capable of performing QC, analysis, and exploration of scRNA-seq data [2]. Before we can utilize the **salmon alevin** UMI counts matrix with **Seurat** filtering, we need to modify the format of the genes.

Cells and Gene Filtering

Each row of the UMI matrix consists of Ensembl gene identifiers (ENSG ID). As these identifiers are not very useful the **biomaRt** package in R was used in order to convert ENSG IDs to gene symbols. Once the gene symbol was generated, the **Seurat** object was initialized with non-normalized data containing the genes expressed in at least three or more cells. The initial dataset consists of 60,232 genes. After filtering for only the cells that contain at least 200 detected genes, 56,120 genes remain. This preliminary filter will remove cells that would otherwise cause technical or biological errors.

Standard pre-processing with filtered data

Once the basic filtration was applied, the Seurat object was ready for some pre-processing steps. These steps include selection and filtration of cells based on QC metrics, data normalization, scaling, and the detection of highly variable genes. Then the cells were also filtered based on if the percentage of mitochondrial genes was less than 0.05 and if cells that contain unique gene counts more than 2,500 or less than 200. After applying all the filters 25,380 genes remained.

After having removed unwanted cells from the dataset, the data was normalized on a log normalized scale using **NormalizeData()** function from the **Seurat** [2] to conduct gene expression measurements for each cell by total expression using a scale factor of 10,000 as the parameter.

Detection of variable genes for downstream analysis was performed using the **FindVariableGene()** function in **Seurat** [2]. This function assists in estimating the relationship between variability and average expression. Since the data may contain variation from technical noise, batch effects, or biological noise caused due to variation in the cell stages, scaling the data can potentially help in removing the unwanted sources of variation. Data scaling was performed using **Seurat's ScaleData** function that regresses out the total number of RNA molecules detected within a cell (nCounts_RNA) and mitochondrial percentage.

Linear dimensional reduction was performed on the scaled data using the **RunPCA()** and **VariableFeature()** function from **Seurat**. The **JackStraw** function was used to identify significant principal components with strong enrichment of low p-value genes. To visualize the results, the **JackStrawPlot** function was implemented (Figure 4).

Graph-based clustering is the default clustering method in **Seurat** for identifying the cell clusters [2]. In PhenoGraph, the KNN(k-nearest neighbor graph)graphs were generated based on euclidean distance in

PCA space, and the edge weights among cells were refined based on the shared overlap. To cluster the cells, the default Louvain algorithm was utilized, and two functions from Seurat were applied: **FindNeighbors()** function and **FindClusters()**.

Clustering According to Marker Genes

Based on the generated clusters, cell subtypes were inferred according to identified biomarkers using the **Seurat package (version 4.0.1)** in R [2]. Using the **FindAllMarkers** function, marker genes were identified for every cluster. This function implements the Wilcoxon Rank Sum Test, from which positive and negative markers are identified for each cluster automatically [2]. A minimum threshold of 0.25 was set during the process of marker gene identification. This measure permits features detected at a minimum of 25% in either of the two groups of cells to be considered meaningful. These identified marker genes were then used to assign clusters their respective cell type.

To all clusters that were not assigned a cell type by the list of genes referenced in the study, an additional list of cell-type specific gene expression markers from PanglaoDB was applied [4]. PanglaoDB is a curated database of human mouse single cell RNA-sequencing samples [4]. It contains information regarding gene expression markers which are specific to cell types and biological organs. This information was downloaded as a CSV and then loaded into R as a dataframe. The data frame was filtered for the *Homo sapiens* species and pancreas organ. The top ten genes from the cluster were then searched against this curated gene list to determine the appropriate cell type assignments. Feature plots, a UMAP projection plot, and clustered heat maps were then produced using **Seurat** to depict the resulting defined clusters.

RESULTS

After reformatting the **alevin** matrix with **Seurat**, the **Seurat** object underwent various filtration stages to remove any unwanted genes and cells that may cause fluctuation in the downstream analysis as a result of technical or biological interference. Detailed information about changes in number of genes and cells after filtering can be found in Table 2. This shows that about 34,852 genes and 221,682 cells were removed from the original dataset after applying all the filters.

Filtering Stage	Genes	Cells
Original UMI matrix	60232	227390
Converting ENSG to gene symbol	56120	227390
Filtering out low-quality cells	25380	9713
Filter out low variance genes	25380	5708

Table 2. Number of genes and cells remaining after each filter step. Filter stages are listed in chronological order: the starting UMI matrix, conversion of ENSG IDs to gene symbols, filtering out low-quality cells, and filtering out low variance genes. Gene and cell count reduced dramatically after the final stage.

Using the filtered dataset, QC metrics were visualized based on the genes and cell counts, along with the percentage of mitochondrial gene composition. This was generated using the **VlnPlot()** function from the Seurat package. Figure 3 demonstrates the highest value in the nCount_RNA along with a high percentage of mitochondrial genes.

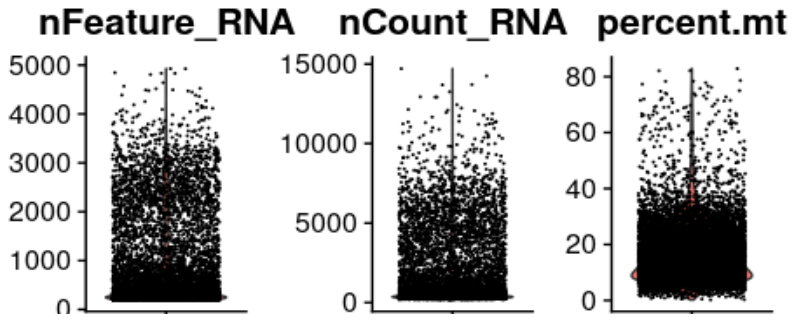
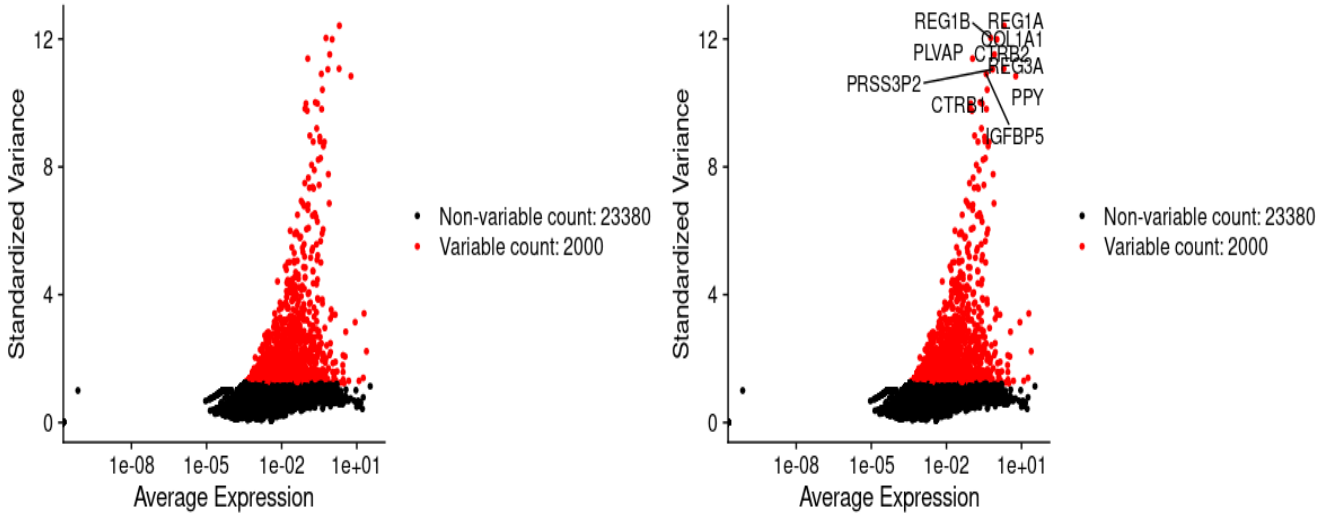


Figure 3. Visualization of QC metrics. The violin plot was generated using the filtered dataset on the cells with feature counts of 2,500 or less than 200, and have >5% mitochondrial counts.

To detect the variable genes that we would like to carry forward for downstream analysis, specific parameters were set used to identify the 2,000 variable genes. This helped in the identification of top 10 most highly variable genes among the entire gene set. The top ten genes can be found in Table 3 and the graphical representation is shown in Figure 4.

No.	1	2	3	4	5	6	7	8	9	10
Genes	REG1A	REG1B	COL1A1	CTRB2	PLVAP	REG3A	PRSS3P2	IGFBP5	PPY	CTRB1

Table 3. Top 10 most highly variable genes obtained after feature selection.



(A) Variable Feature **(B) Variable Feature with top 10 genes**

Figure 4: The scatter plot representing the relationship between the Average Expression and the Standardized Variance. The red colored points are the points of interest for the downstream analysis of variable count of 2,000. **(A)** This scatter plot represents those without the top 10 most variable genes and **(B)** shows the top 10 highly variable genes referred to in Table 3.

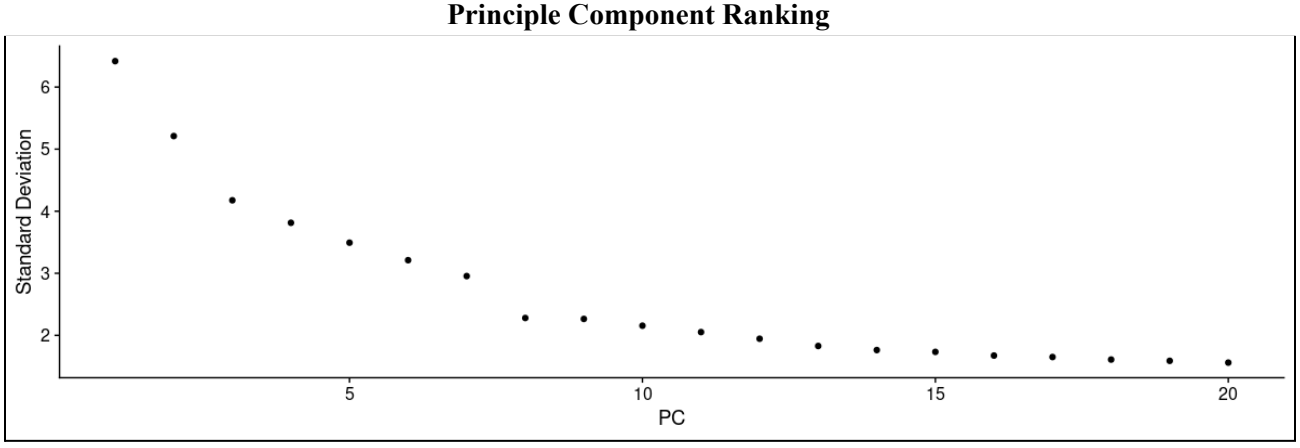


Figure 5: Elbow plot which represents the PCs based on the percentage of variance explained above. After a PC of 7, the PCs got more stable. However, before that, a drastic fall is observed in the relationship between standard deviation and PCs. This indicates that PC 7 can be used as a PC cut-off.

The cell clusters were visualized based on the UMAP dimension reduction method shown in Figure 10. The **RunUMAP** function in **Seurat** was used to obtain the clusters with dim(parameter) upto 10. Moreover, the colors were automatically selected for each cell type. This technique can be very helpful in analyzing multiple cell clusters without any concerns of color selection.

As observed in Figure 10, Group 0 is highly dense and represented in orange coloring. This suggests that group 0 consists of the maximum number of cells where each dot represents individual cells. Contrastingly, group 11 represents the lowest number of cells among all the clusters. Identification of these clusters types will be performed in the further steps. Moreover, the graphical representation and tabular representation of the number of cells found in each cell type can be found in Figure 6 and Table 4, respectively.

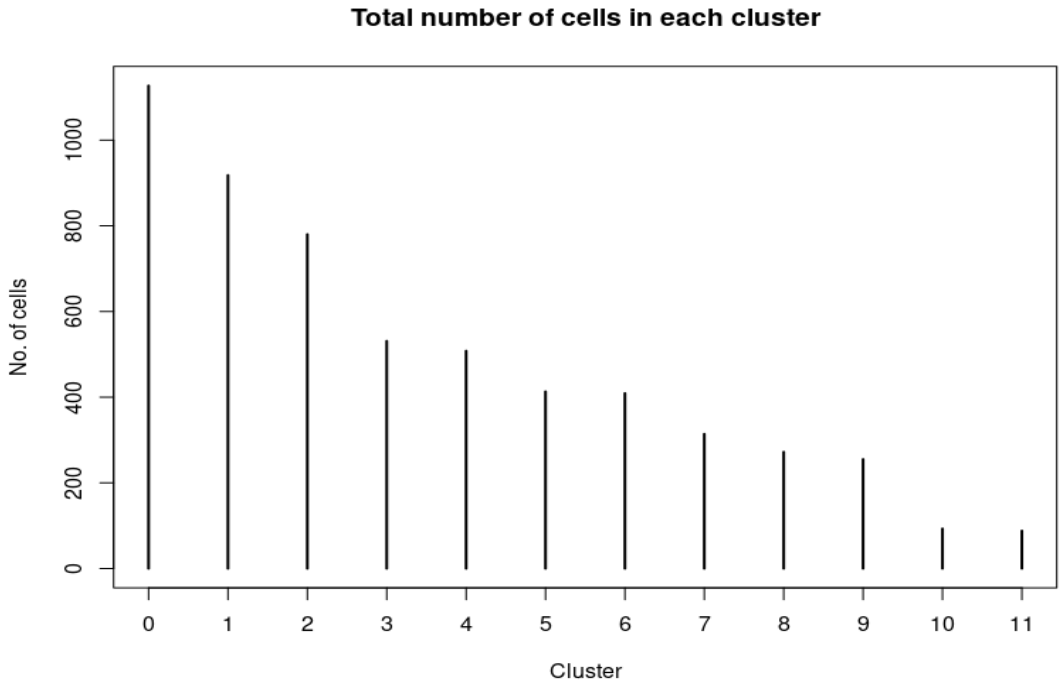


Figure 6: Graphical representation of number of cells in respective clusters. The x-axis represents the cluster count numbered from 0 to 11. A total of 12 clusters were found in our dataset. The y-axis represents the number of cells.

Cluster No.	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Cell Count	**1127**	**918**	**780**	**531**	**508**	**413**	**409**	**314**	**272**	**255**	**93**	**88**

Table 4: Tabular representation of the number of cells in each cluster.

In addition to the UMI matrix salmon alevin produces there is a log file produced from the analysis itself. The log file produced from our salmon alevin contained 14 metadata statistics, 8 of the most useful are listed in Table 5. Total reads consist of every read alevin read as input, the total number of used reads consists of reads that remain after removal of noisy barcodes, noisy UMI, and reads with at least one nucleotide N. It is reassuring to see the used reads counts is comparable to the total reads, this means our data was decent quality. The mapping rate is the number of mapped reads over the total reads. It was surprising to see the number of cellular barcodes output by **alevin** is about 5% of the total number of barcodes observed by **alevin**. The Deduplicated UMIs is the total number of UMIs present in the experiment post UMI deduplication across all cells.

Alevin Log File Summary	
Total Reads	1,324,837,961
Used Reads	1,304,472,600
Mapping Rate	43.14829593
Total Number of Cellular Barcodes(CB)	4,251,176
Number of CB in the Output Matrix	249,735
Deduplicated UMIs	25,654,431
Mean UMIs Per Cell	102
Mean Number of genes >0 Per Cell	67

Table 5. Log file produced by Salmon Alevin analysis containing various metadata.

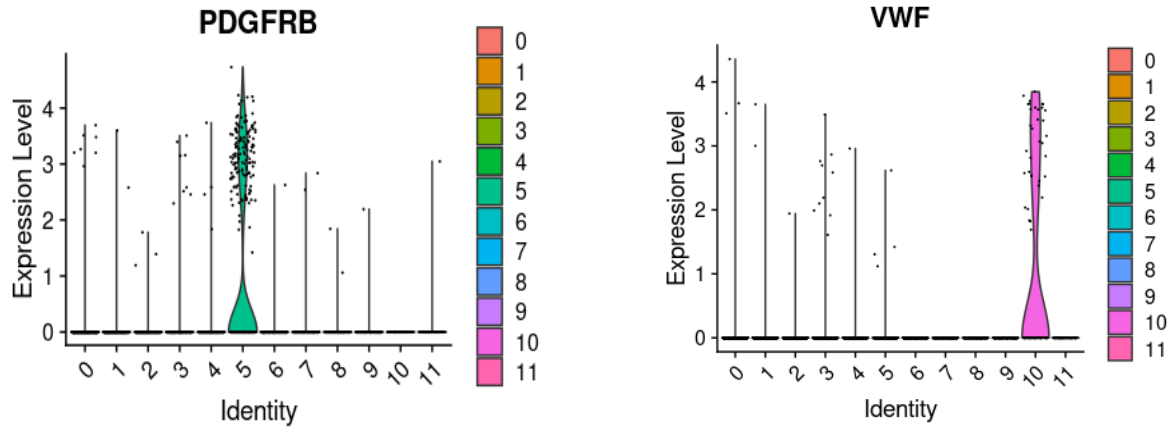
Cell Type Identification

A total of twelve clusters were produced from which select marker genes were identified using the methods described in the *Cell Type Clustering* subsection above. 1050 biomarkers were selected as marker genes. To better visualize the top marker genes per cluster, a heatmap was generated as shown in Figure 7 below.

Figure 7: Clustered heatmap showing the top ten differentially expressed genes in each of the twelve clusters. Highly expressed bands are indicated by the yellow bands.

By generating the heat map prior to cell typing the clusters, the decision-making process behind determining the cell types can be guided by the genes which were shown to be highly expressed in each respective cluster. Upon first glance, one can infer that each of the cell subtypes appear to be distinctly expressed given the relatively clear separation between the highly expressed bands indicated in yellow coloring. This observation allows for some reassurance that the methodology up to this point was applied appropriately. Additionally, pairs of clusters are observed which display several shared up-regulated and down-regulated genes between one another. These include clusters 1 and 2, clusters 3 and 7, and clusters 6 and 9. Despite their organization into different cell subpopulations, one can hypothesize that these pairs of clusters may belong to similar, if not identical, cell types. This may be confirmed in further analyses and gene set enrichment.

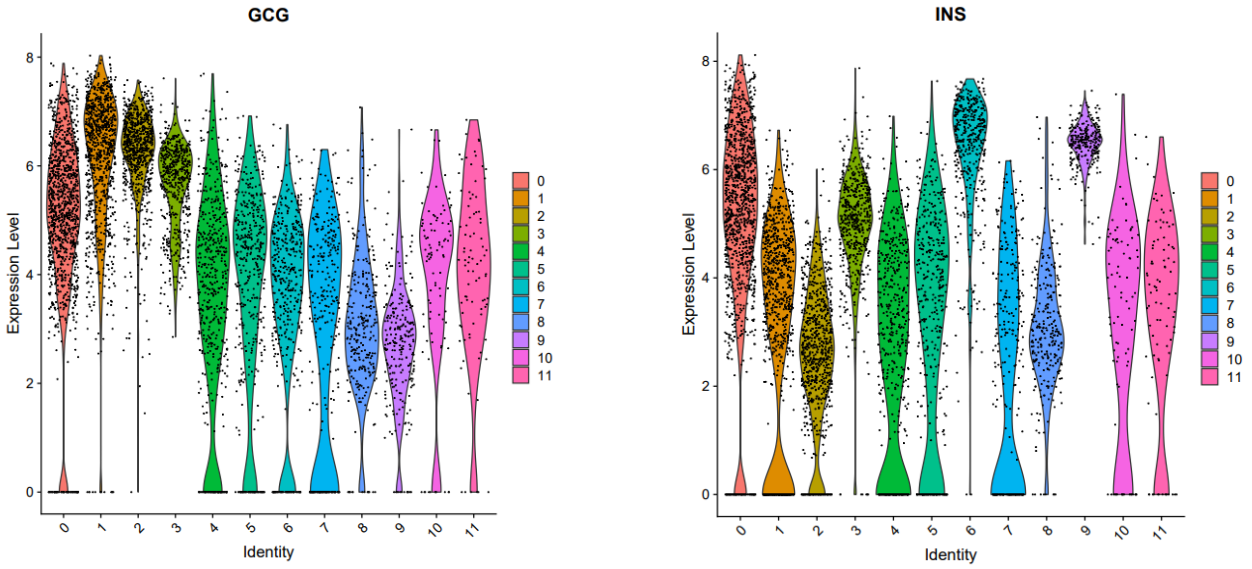
The marker genes chosen to indicate clusters of a particular cell type were guided by those from the paper [1]. Violin plots were generated to indicate these clusters as shown in Figure 17. Some clusters were able to be uniquely identified by a single gene. For example, in Figure 8A, PDGFRB is differentially expressed only in cluster 5, and thus, allowed this cluster to be assigned the corresponding gamma cell type stellate cell type. In Figure 8B, VWF is uniquely expressed only in cluster 10 and consequently assigns it a vascular cell type. Since these genes were only significantly expressed in one of the twelve clusters, they were able to be assigned with minimal difficulty. A full table of the top three marker genes for each cluster can be found in Table 6.



(A) **(B)**

Figure 8. Violin plots depicting the UMI counts per cluster for (A) the PDGFRB gene and (B) the VWF gene. PDGFRB is highly expressed in cluster 5, making its cell type assignment correspond to the stellate population. Similarly, VWF is highly expressed in cluster 10, making vascular cells its cell type assignment.

Contrastingly, some of the marker genes used by the paper were less informative than PDGFRB. For example, in the paper, the GCG gene was the marker gene for alpha cells and the INS gene was the marker gene for beta cells. However, both genes were highly differentially expressed in cluster 0 as seen in Figure 9A and 9B. Additionally, GCG was highly enriched in clusters 1 and 2, while INS was highly enriched in clusters 6 and 9.



(A) **(B)**

Figure 9: Violin plots depicting UMI counts for the GCG **(A)** and INS **(B)** genes by cluster. Aside from other clusters where they were highly expressed, both genes share a high expression level for Cluster 0.

In order to deduce the correspondence of the appropriate genes to their respective clusters, feature plots were generated using the **FeaturePlot** function in the **Seurat** package [1] as shown in Figure 16. Cells were uniquely colored in a resulting UMAP projection according to gene enrichment for a specific cell type. This process visualized clusters where certain genes were enriched in different positions across the UMAP projection as shown in Figure 10.

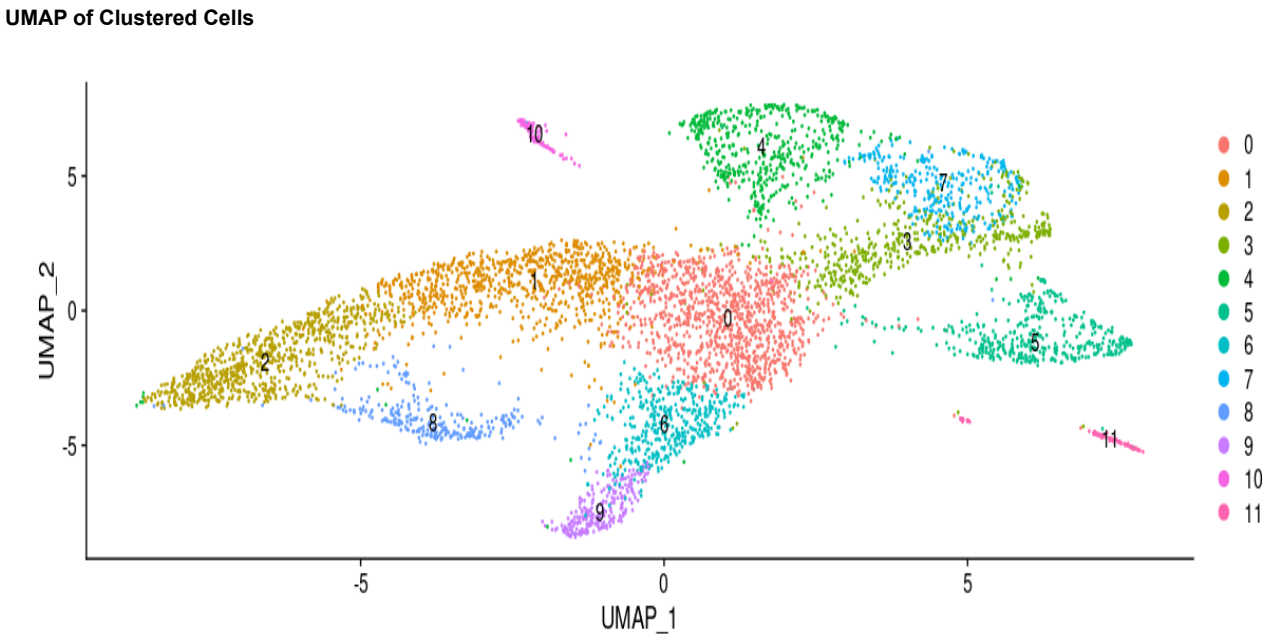


Figure 10. UMAP projection of clustered cells colored by cell population. Visualization of the intensity of the clusters allows for additional understanding of the density of every cell present in the sample.

In Figure 11, one can observe the correspondence between the enrichment of genes GCG and INS to their respective clusters marked in Figure 10. When examining Figure 11A with the information provided by Figure 10, GCG appears highly expressed in areas which correlate to clusters 1 and 2. This finding is supplemented by the information offered by the heat map in Figure 7. Thus, it is reasonable to infer that clusters 1 and 2 belong to the cell type that is most markedly known for enrichment of the GCG gene: pancreatic alpha cells. Similarly, INS is largely localized in the areas of the projection which correspond to clusters 6 and 9 in Figure 11B. Given that this information is also supplemented by the shared differentially expressed genes between these clusters in Figure 7, it is likely that these clusters belong to the same cell subpopulation and can be labeled as pancreatic beta cells. This rationale was applied on all clusters which genes shared enrichment for in order to determine cell types.

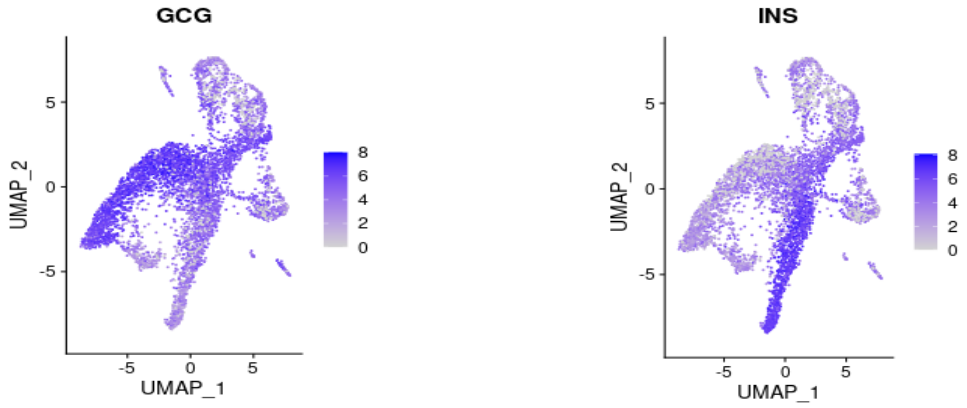
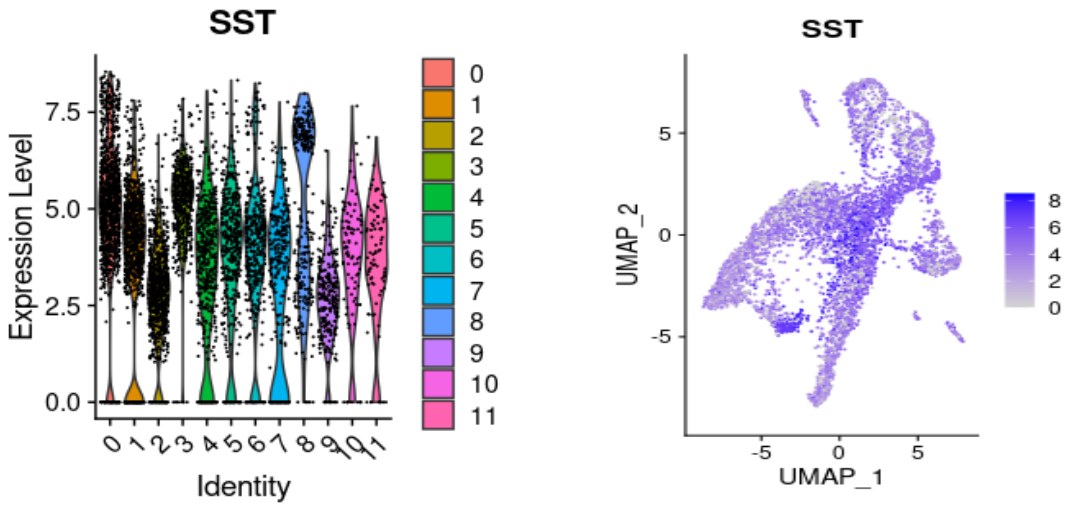


Figure 11: UMAP projection of clustered cells colored according to (A) GCG gene enrichment and (B) INS gene enrichment.

For clusters which had enrichment of several genes even after assigning cell types to all other clusters, a curated gene list from PanglaoDB was applied. Specifically, this strategy was applied towards cluster 0 as it displayed enrichment in several of the feature plots and violin plots for other cell types which were already assigned. For example, though the alpha and beta cell type cluster assignments were rationalized according to the process described above, cluster 0 still shows enrichment of the marker genes for these cell types: GCG and INS. Of the top ten marker genes of cluster 0, the SST gene was found to be indicative of the pancreatic delta cell type. The remainder of marker genes were not clearly labeled as a particular cell type in the curated PangalaoDB gene list. To supplement the decision of assigning cluster 0 the delta cell type, violin and feature plots were generated for SST which are shown in Figure 12A and 12B, respectively. Collectively, SST appears the most enriched in clusters 0 and 8. However, cluster 8 was ruled out given that it expressed enrichment for PPY which is indicative of the gamma cell type.



(A) **(B)**

Figure 12. Violin plot **(A)** and feature plot **(B)** of the SST gene which, together, indicate the highest level of enrichment in clusters 0 and 8.

Certain cell types which were identified in the study were not obtained in our analysis including epsilon cells indicated by the GHRL marker gene, mast cells which are indicated by the TPSAB1, KIT, and CPA3 marker genes, and cytotoxic T-cells which are indicated by the marker genes CD8, CD3, and TRAC. These cells were not present in our sample given the absence of differential expression of these marker genes in our dataset.

Each of the nine identified cell types were confirmed with violin and feature plots where the twelve clusters were organized into a total of nine different cell types, as shown in Figure 13. To confirm the identification of these clusters, a repeat heat map was generated, shown in Figure 14. Our findings are reassured after viewing relatively clear and distinct bands grouped according to cell type and highly expressed genes observed across each of the cell types. The heatmap also allows for identification of novel marker genes for each cell type identified. In this analysis, novel marker genes are those that appear to be differentially expressed in each cell type, yet not identified by the study as being indicative of a particular cell subpopulation. This is seen in Figure 14 in the form a heatmap.

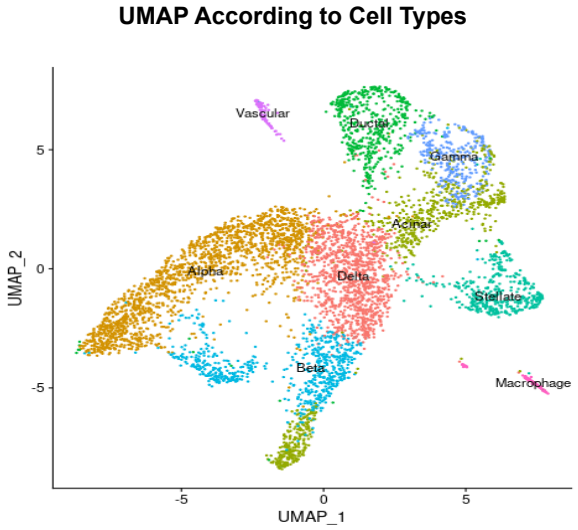


Figure 13: UMAP projection plot of log-normalized count matrix as colored by cell type. Nine total cell types were identified when using the marker genes outlined in the study [1].

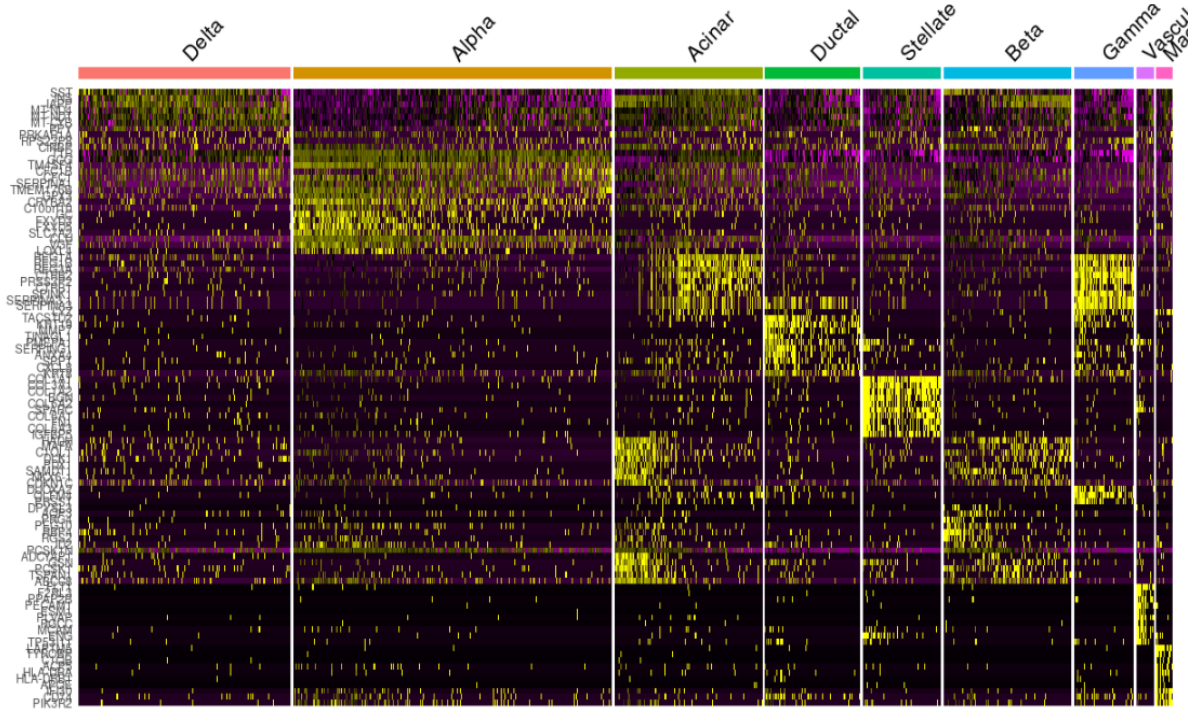


Figure 14: Heat map of the ten most significantly expressed genes for each cell type cluster. A total of nine cell types were identified. Significantly enriched genes are indicated by yellow bands.

DISCUSSION

UMAP Projection Analysis

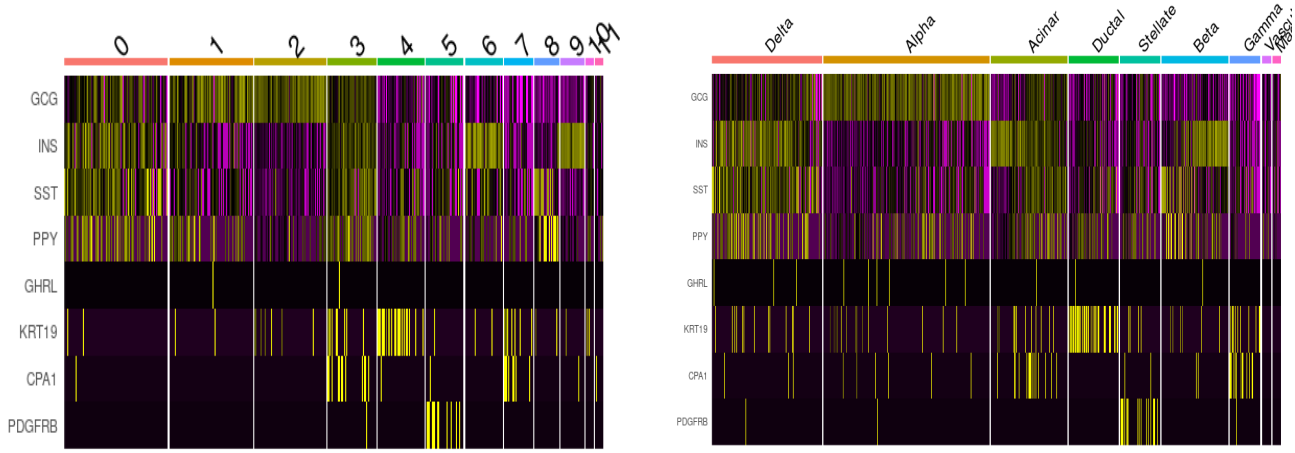
To visualize the clustered cells, a UMAP projection was generated. While the study chose to display these findings in the form of a tSNE, we chose a UMAP given that it places greater emphasis on both the clustering of the cells and the distances between the cell clusters [6]. Contrastingly, tSNE does not preserve the global data structure, making the distances between clusters variable in nature, thus resulting in the loss of meaningful information to assist with cell type assignments [6]. Consequently, using a UMAP to visualize the data deemed more appropriate in this analysis.

The paper's projection is shown in Figure 1D and depicts 14 unique, well-separated cell clusters [1, Figure 1D]. It is important to remember, however, that since a tSNE plot was used, these distances offering to the separation among clusters are arbitrary. Direct comparison between our UMAP plot and their tSNE plot is not ideal given that the global structure of the study's dimensionality reduction cannot be accurately inferred. Nonetheless, it is noted that the paper was able to distinguish between two subpopulations of stellate cells: quiescent and activated. No clusters indicating the vascular cell type are observed in the paper's plot, however, a cluster associated with an endothelial cell type was identified

which is not shown in our plot. It is worth noting that the authors make mention of the "vascular" cell type and "endothelial" cell type interchangeably in the paper, so this may have contributed to the difference between our cell type projection plots. Additional cell types were also identified by the study which will be discussed further in the cell identification analysis.

Heatmap Comparison

As a part of our analysis, we attempted to reproduce Figure 1B from the paper [1]. To do so, the same marker genes used in the study were used to generate two new heatmaps, one grouped by cluster and one grouped by assigned cell type, shown in Figures 15A and 15B, respectively. During the process of generating these heat maps, however, several of the genes which the study noted notable enrichment of were not present in our analysis. These included the following: TPSAB1, KIT, CPA3, CD3, CD8, CD163, CD58, IgG, VWF, PECAM1, and CD34. No identifiable transcriptional patterns or banding within the heatmap were present to associate with these genes.



(A) **(B)**

Figure 15: Heatmaps of significant marker genes according to the referenced study, grouped according to cluster **(A)** and assigned cell type **(B)**.

Several of the cell types depict distinct, clear banding patterns such as among the ductal and stellate cell types. Other cell subpopulations, however, do not show the same. It appears a number of cell type clusters are dependent on shared genes as indicated by clusters sharing high expression of the first four genes: GCG, INS, SST, and PPY. As a result, clusters with cell types that are reliant on the up-regulation of these genes can be subject to conflicting assignments, such as with alpha and beta cells. This was observed in earlier analyses as well, where alpha and beta cell type assignment required further deduction using feature plots.

These sources of variation between our heatmap and the paper's may be due to the limited cell sample population our dataset permitted. The study drew data from a much larger cell population than our analysis did, which only looked at a single donor. In Figure 1B of the paper, a significant portion of the heatmap results are for the third donor which is indicated in orange coloring of the Sample bar, while our donor's sample, donor 2, makes up a minimal portion of the heatmap findings [1, Figure 1B]. Thus, it is

reasonable to expect some differences in the translation of our findings to that of the paper's. Additional down-stream analysis or expanding the sample size may be necessary to best replicate the study's findings in this regard.

Cell Type Identification Analysis

In our analysis, we identified nine different pancreatic cell types, while the study was able to identify fourteen. The five additional cell types which they identified include: quiescent stellate cells, cytotoxic T-cells, pancreatic epsilon cells, mast cells, and Schwann cells.

Though our analysis was able to identify stellate cells, we were unable to distinguish between the quiescent and activated subtypes. This may be because we did not have marker genes associated with these cell types to identify their respective subpopulations. It is likely that the transcriptional profiles of these two cell types are closely related, and thus, require more information to detect a noticeable difference between the two since they share the same overall cell type. This finding may also be limited by the cell population we worked with since we only used one donor's dataset.

The inability for our analysis to detect the cytotoxic T-cell and epsilon cell groups is expected given that these cell types are reported to constitute only about 0.1% of the mixture of cell types each [1]. Since the cell size of our analysis was limited to less than 5000, it is statistically improbable that these cells were present in our dataset for identification and clustering. It is also unlikely that our specific donor had a significant measure of these cells as the paper mentions that epsilon cells in particular are commonly believed to be absent in adults [1].

Interestingly, there is some conflicting data between the paper's reported figures and their supplementary data. In Figure 1G of the paper, the cell types and marker genes indicated do not correspond appropriately to those that were listed in the supplementary dataset [1]. A total of 15 cell populations are identified in the figure, while only twelve are listed in the supplementary data. Because of this difference, a compiled list of both sets of marker genes was used in our analysis in order to best identify the cell populations of the clusters.

CONCLUSION

Overall, the findings from the study were largely reproducible. Though some additional cell types were not identified in our data sample, the nine identified cell types largely reflect that of our donor's when compared to those reported in the paper. Discrepancies in our results are attributed to working with a subset of the data from the study and variations in the chosen threshold values and marker genes when conducting cell type assignments for each of the clusters. Each step in the pre-processing workflow assigns certain arbitrary parameters, such as when selecting the type of clustering method or threshold values. This may also constitute the difference to the overall results when examining the number of cells and number of genes per cluster. Such discrepancies may be retained during downstream analysis, thus potentially leading to more differences in our overall findings. However, if a stringent threshold is set in the filtering stages, this should still allow for the selection of the appropriate genes. Therefore, threshold selection is a crucial step in the quality control process in order to ensure replicability. Repeating this analysis with the complete dataset from the study would likely enhance the reproducibility of their results.

Additionally, confirming our findings with gene set enrichment analysis would allow for greater confidence in our cluster assignments and grant more room for biological interpretation of our findings.

REFERENCES

1. Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, 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). https://doi.org/10.1016/j.cels.2016.08.011
2. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., … Satija, R. (2020). Integrated analysis of multimodal single-cell data. https://doi.org/10.1101/2020.10.12.335331
3. Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., & Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell, 161(5), 1187–1201. https://doi.org/10.1016/j.cell.2015.04.044
4. Franzén, O., Gan, L.-M., & Björkegren, J. L. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database, 2019. https://doi.org/10.1093/database/baz046
5. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods.
6. McInnes L, Healy J et al. UMAP: Uniform Approximation and Projection for Dimension Reduction. Journal of Open Source Software. 2018;3(29):861. doi: 10.21105/joss.00861
7. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1;30(1):207-10

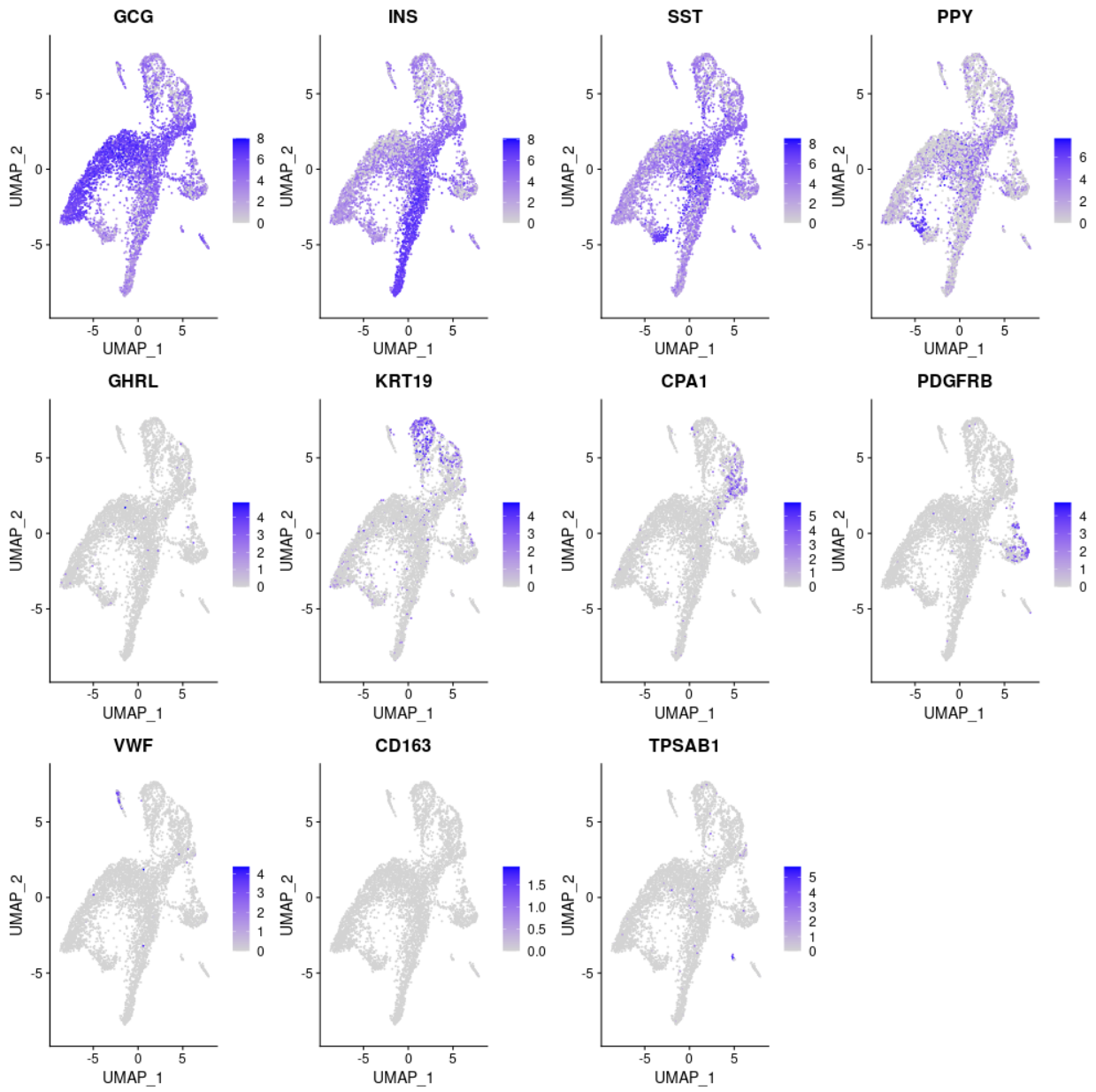


Figure 16. Feature plots of each of the marker genes. Darker coloring indicates greater density of the gene in the respective location of the plot. Inferences regarding each marker gene's association with a specific cluster can be made using these plots.

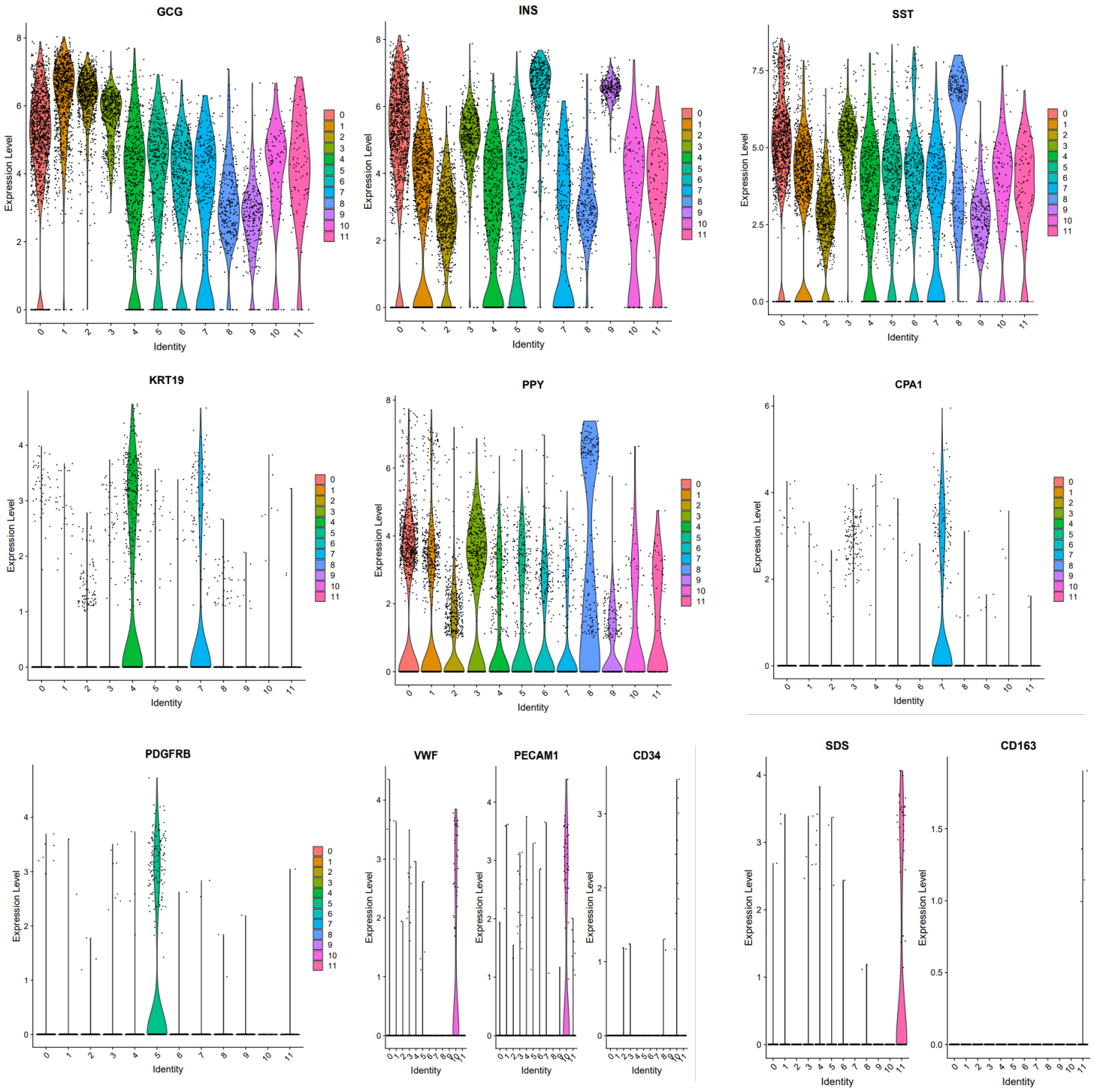


Figure 17. Violin plots depicting the UMI matrix for each of the marker genes. High concentration of dots indicate prevalence of the genes in their appropriate clusters.

Gene	Cluster Number	P-Value (nominal)	Average Log2 Fold Change
SST	0	1.987E-122	1.623
INS	0	1.6944E-112	1.258
IAPP	0	5.5425E-102	1.217
TTR	1	1.2606E-190	1.313
GCG	1	9.0452E-161	1.458
TM4SF4	1	8.39234E-56	1.056
VGF	2	2.8725E-227	1.484
LOXL4	2	2.093E-188	1.385
PTPRS	2	2.3454E-188	1.161
REG1A	3	3.4896E-252	1.956
PRSS3P2	3	5.7974E-145	1.253
CTRB1	3	5.439E-132	1.791
TACSTD2	4	2.3737E-261	2.781
KRT19	4	3.4365E-197	2.739
SERPINA5	4	1.0755E-190	2.283
COL1A1	5	0	5.152
COL3A1	5	0	4.613
COL1A2	5	0	4.414
INS	6	1.2155E-144	2.138
IAPP	6	8.60103E-96	2.088
MAFA	6	2.99354E-64	1.713
CTRB2	7	1.2363E-247	4.099
IL32	7	1.098E-228	3.405
CPA1	7	2.0968E-147	3.133
DPYSL3	8	2.656E-182	1.078
AQP3	8	8.2916E-139	2.232
PRG4	8	1.8863E-118	1.146
MAFA	9	1.6095E-248	1.833
SAMD11	9	1.8327E-234	1.395
NKX6-1	9	1.9532E-216	1.484
FLT1	10	0	3.921
F2RL3	10	0	3.310
VWF	10	7.2193E-287	3.157
TYROBP	11	0	3.860
C1QB	11	0	3.640
SDS	11	9.302E-255	3.188

Table 6: Table listing the top three marker genes for each cluster. Nominal p-values and average log2 fold changes are reported.