BF528 Project 4: Single Cell RNA-Seq Analysis of Pancreatic Cells

Authors: Vrinda Jethalia (Data Curator), Manasa Rapuru (Programmer), Pragya Rawat (Analyst), Pooja Savla (Biologist)

Date: 28th April, 2023

Introduction:

Single-cell RNA (scRNA) sequencing is a powerful method that allows for the analysis of transcriptomic information in small populations of cells, providing insights into cell population heterogeneity (Tang et al., 2019; Zhang et al., 2021). It enables the identification of differential gene expression, epigenetic changes, and specific cell markers, contributing to our understanding of disease mechanisms and treatment strategies across various fields, including oncology (Tang et al., 2019; Zhang et al., 2021). Tumor development, characterized by the accumulation of somatic mutations, is influenced by cell heterogeneity, which can be effectively detected through scRNA sequencing by examining transcription patterns and identifying cell types within cancer (Zhang et al., 2021).

In a study conducted by Baron et al., they utilized a droplet-based scRNA sequencing method called inDrop to analyze the transcriptomes of 12,000 pancreatic cells from two mouse strains and four human donors (Zilionis R. et al., 2016). By encapsulating mRNA from lysed cells within droplets containing barcoded hydrogel beads, they achieved barcoding of mRNA through reverse transcription. The study revealed distinct expression profiles of subpopulations of ductal cells and identified disease-associated differential expression patterns and gene regulation heterogeneity in B-cells within the human pancreas, particularly regarding functional maturation and levels of ER stress (Baron et al., Year).

In our own study, we aimed to replicate the findings of Baron et al. using sequencing data obtained from a 51-year-old female donor. Our research involved processing the barcode reads, generating a matrix of unique molecular identifier (UMI) counts, performing quality control on the UMI matrix, and analyzing it to identify patterns of cell clustering and marker genes for each cluster. Additionally, we conducted gene set enrichment analysis on the identified marker genes to determine their biological significance.

Understanding the transcriptomes of individual pancreatic cells is crucial for unraveling the pathology of various diseases, including diabetes (specifically type 1 and type 2 diabetes mellitus), pancreatitis, and cancer (Yanai et al., 2017). Although the function of the mammalian pancreas relies on complex interactions among distinct cell types, gene expression profiles have primarily been described for bulk mixtures. Therefore, the authors of this paper aimed to

overcome this limitation by determining the transcriptomes of over 12,000 individual pancreatic cells from four human donors and two mouse strains using a droplet-based, single-cell RNA-seq method.

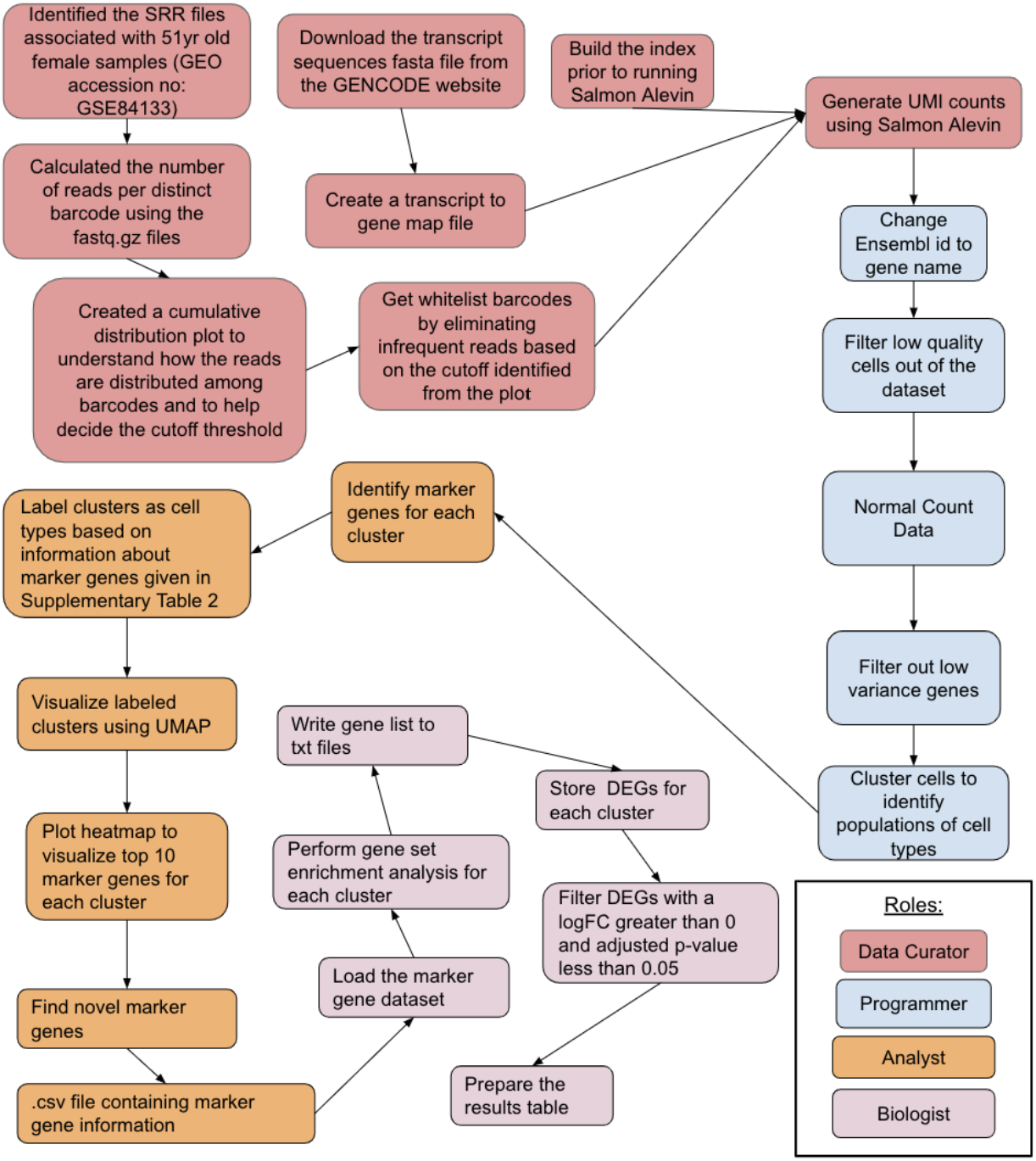

Figure 1: The work flow diagram describes the steps that each group member took to collect, process and analyze the data.

Data:
In the paper, inDrop was used to determine single cell transcriptomics in over 12,000 cells from two mouse strains and four human donors. The researchers isolated and sequenced about 10,000 human pancreatic cells from four cadaveric donors and about 2000 mouse pancreatic cells from two mouse strains. In our project, we only worked with the samples of one human donor who was a 51 year old female. The samples were located from GSE84133 which is the GEO accession page of this research. Three samples are associated with the 51 year old female human donor which were SRR3879604, SRR3879605 and SRR3879606.

The FASTQ files for the three samples were used to count the number of reads per distinct barcode. SRR3879604_1_bc.fastq, SRR3879605_1_bc.fastq and SRR3879606_1_bc.fastq files had the barcode and UMI information. As it is known, in a FASTQ file, 4 lines represent one read. For example,

@SRR3879605.7 bc1=AGAGGTGCTAC bc2=GATTGGGA umi=GAGCGG
AGAGGTGCTACGATTGGGAGAGCGG
+
BBBBBFFFFIIFFFFFFFFFFFFFF

indicates a single sequence read from a sequencing experiment. The "@SRR3879605.7" indicates the identifier of the sequencing read. "bc1=AGAGGTGCTAC" and "bc2=GATTGGGA" represent the two cell barcodes used in the experiment, where "bc1" is the barcode of 11 nucleotides (AGAGGTGCTAC) and "bc2" is the barcode of 8 nucleotides (GATTGGGA). "umi=GAGCGG" indicates the unique molecular identifier (UMI) sequence of 6 nucleotides (GAGCGG) used to distinguish between duplicated cDNA molecules. The second line "AGAGGTGCTACGATTGGGAGAGCGG" represents the nucleotide sequence of the read, where the first 11 bases (AGAGGTGCTAC) correspond to the first cell barcode (used to identify which cell a particular RNA molecule came from), the following 8 bases (GATTGGGA) correspond to the second cell barcode (used to distinguish between RNA molecules that originated from the same cell but are different copies of the same transcript), and the last 6 bases (GAGCGG) correspond to the UMI sequence (Svensson et al., 2018).
The third line "+\n" indicates the start of the quality scores, where each symbol represents the quality score of the corresponding base in the nucleotide sequence. The quality scores provide information on the reliability of the base calls and are important for downstream analyses such as read filtering and trimming (Malhotra et al., 2022).

The count of reads per unique barcode for each sample was consolidated to one file in R so that we could capture the count of a particular barcode across the 3 samples. The next step was to filter out the low count/infrequent barcodes and hence, a cutoff threshold was to be determined. There is no one-size-fits all threshold for filtering out barcodes, but a common approach is to

plot the distribution of counts per barcode using a histogram or cumulative distribution plot. In our project, we plotted a cumulative distribution plot in R and visually inspected the plot to identify the “elbow” or “inflection point” of the distribution (to remove the low count tail of the distribution).

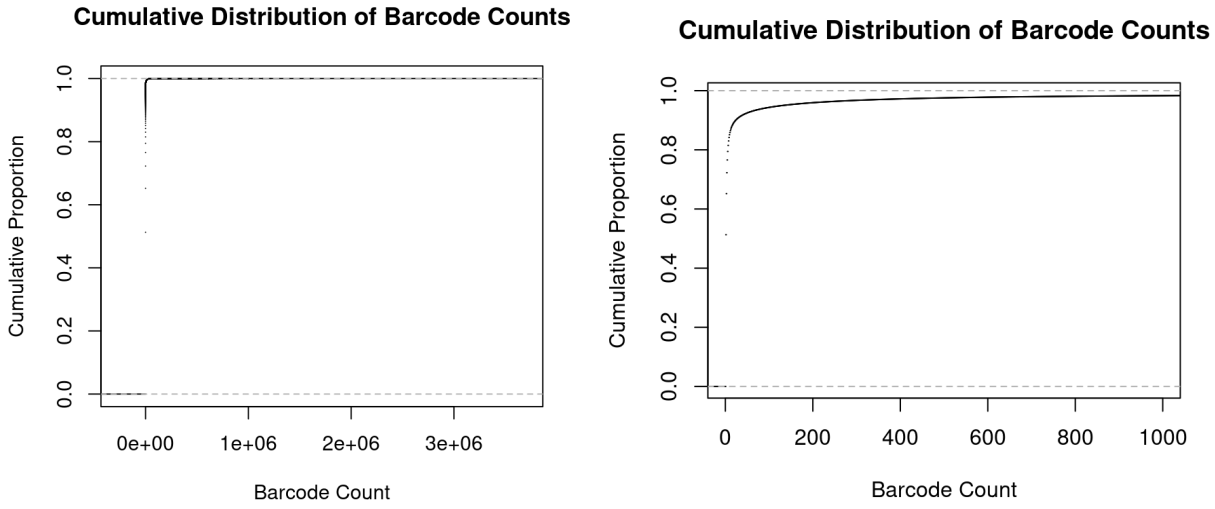

Figure 2: Two plots that showcase the cumulative distribution of barcode counts. The figure on the left is the overall distribution whereas the figure on the right is a zoomed in distribution with a shorter range on the x-axis. The purpose of the zoomed in figure was to identify the “inflection point” on the graph.

From figure 2, it can be seen that the plot on the left has a very large range on the x-axis and it is difficult to accurately determine the value at the inflection point. Therefore, another plot with the same data (figure 2, right hand side figure) was created where the x-axis range was lowered to the 1000 counts. It can be seen that the “dip” is roughly around 180 and therefore, we set it as our threshold. This implies that barcodes with counts less than 180 were filtered out and the remaining barcodes were the “whitelist” barcodes.

The next step was to obtain the UMI counts matrix for the single cell sequencing data which was achieved using the Salmon Alevin software. To prepare the mapping file necessary for running this tool, the human transcript sequences fasta file was downloaded from GENCODE and a mapping file was created by extracting the headers of the fasta file which resulted in a two column file of transcript and gene ids. This allows salmon to collapse transcriptome data to gene level. A file containing the salmon index of the reference transcriptome was also created using the “salmon index” command prior to running the alevin. Finally, alevin was set to run with custom parameters for a barcode length of 19 bases, a UMI length of 6 bases, and an end length of 5 bases for analyzing the paired FASTQ files using the whitelist barcode counts read.

There were multiple files associated with Alevin's output. Firstly, a log file that records the software's actions and output, as well as any error messages or warnings encountered during the analysis. The log file can be used for troubleshooting issues with the software or for verifying the completeness and correctness of the analysis results. Some of the information contained in the log include the parameters specified for the analysis, the number of reads processed etc. Second, a featureDump file that contains information on the features (genes or transcripts) and their quantification values in a single-cell RNA sequencing experiment. The file can be further processed and analyzed using various tools and methods to gain insights into gene expression patterns and identify differentially expressed genes between different cell types or conditions. The other output files are the quants_mat.gz file which is the compressed count matrix and is used in the next step of the project; the quants_mat_cols file that holds the column header (Gene-ids) of the matrix; the quants_mat_rows that holds the row index (CB-ids) of the matrix and the quants_tier_mat which is a file with tier categorization of the matrix that is based on confidence in uniquely mapping to the reference.

4251176 total number of unique cellular barcodes were identified in the sequencing data. We had a total of 189587 white-listed barcodes that passed quality control and were included in the downstream analysis. 2.14% reads were identified as having noisy cellular barcodes and were excluded from further analyses. 237516 barcodes were used for downstream analyses, excluding those that did not pass quality control. 16045 barcodes were skipped because of no mapped reads.

The broad research question was to determine the transcriptome of a large number of pancreatic cells for which a low percentage of discarded reads due to noisy cellular barcodes is desirable as seen in our results. Therefore, selecting 180 as the threshold based on the plot (figure 2) seems justified because a high-quality dataset with minimal noise is necessary for accurate identification of subpopulations and for robust differential gene expression analysis. However, other quality control metrics are essential as seen in the next steps to identify accurate clusters.

Methods:

Processing Count Matrix

The UMI count matrix that was generated in the previous steps needed to be processed to filter out low-quality cells, filter out low variance genes, and identify clusters of cell type subpopulations. Seurat is a R package that is used for the analysis of heterogeneity in single cell transcriptome data. The rows in the matrix are organized by gene and cell are the columns. Each value in the table represents the number of molecules found for each feature. The salmon alevin counts file generated in the data section was loaded into R using the tximport library. Initial data set consisted of 662296 genes and 17342 cells. Before the file can be further processed, the ensembl id names that were used to identify the genes, had to be changed to the actual gene

name. This was accomplished using the biomart library. First the "." and following digits that indicate version on the ensembl id number were removed so the id can be mapped with the gene name. A table with the ensemble ids and the corresponding gene names was created and then it was merged with the ensembl ids that existed in the row names. The ensembl id row names were then replaced with the corresponding gene names as the row names in the counts table. Next a Seurat object was created using the count matrix so that the data along with corresponding plots can be contained together. Next quality control was done on the dataset to filter out low quality cells and genes with low variance.

Filter Low Quality Cells:

Low quality cells were filtered for by looking for cells that had very few gene count. After viewing a distribution of features and count data to set thresholds as to where the bulk of the data lied, the bulk was found to be within 5000 for the number of features with a minimum of 200 (Satija Lab, 2020). After making the select the data set consisted of 21,675 genes and 12,485 cells.

Normalize Count Data:

Once the low quality cells were filtered out the remaining data was normalized using log normalization method with a scale factor of 1000 (Satija Lab, 2020).

Filter Low Variance Genes:

This step was achieved by selecting features that had high cell-to-cell variation for these genes as good indicators of biological signal because they represent heterogeneous features.This was done using the FindVariableFeatures function (Satija Lab, 2020). This function models the mean-variance relationship that is present in the data. The default of 2000 features per dataset was used. Next the data was scaled before any dimensional reduction analysis was performed. Scaling serves to shift the mean expression of a gene across samples to 0 and scale the variance to 1.The number of cells and genes remained the same.

Identify Clusters of Cell Type Populations

Lastly, the cells in the data were clustered using the FindNeighbors and FindClusters functions. And then a UMAP of the data was generated to graph the cell clusters. 14 clusters were identified (Fig3).

Gene Set Enrichment Analysis using DAVID

Within the biologist's domain, the objective involved utilizing the marker genes acquired through an impartial clustering procedure to validate the cell type and function attributed to the distinct clusters. To accomplish this, a gene set enrichment analysis was executed on the marker genes within each cluster using the bioinformatics tool DAVID. The marker genes underwent filtration based on two criteria: a p_val_adj value less than 0.05 and an avg_log2FC value greater than 0.

Subsequently, the resulting Gene Ontology (GO) terms were compared to the cell type labels associated with each cluster.

Results:

Following data processing and quality control, dimensionality reduction and clustering was visualized using the Uniform Manifold Approximation and Projection (UMAP) technique. Formation of 14 clear clusters suggested the presence of distinct cell types (Fig3). From figure 4 we can determine that the cells from the human pancreas are organized into distinct clusters, suggesting the presence of different cell types and populations in the tissue; further corroborating our findings. Data was clustered using a graph based clustering approach (Satija Lab, 2020) with the Seurat package in R.

To characterize the identities of the distinct cell types using genes, differential gene expression was conducted using the FindAllMarkers function. The FindAllMarkers function compares every cluster compared to all the remaining cells (Satija Lab, 2020). 28 markers having recommended thresholds such as a positive differential gene expression with a minimum log2 fold change value of 0.25, and a minimum detection frequency of 25% in all the cells were identified. The parameters were chosen in order to ensure selectivity while identifying distinct cell types. For instance, having a stringent PCT ensured that the marker genes were highly expressed in the clusters they were identified in, and in turn were more informative in providing characterizing information on the cell type.

Clusters were labeled based on the markers highly enriched in a particular cluster. For instance, cluster 0 was annotated as a pancreatic Delta cell since it was enriched in the expression of Somatostatin (SST). These annotations were based on previous literature including Supplementary Table 2 provided in Baron et al. (2016) as well as the CELLxGENE suite of tools (Chan-Zuckerberg Initiative, 2021). It is important to note that while there were 14 total cell clusters, only 10 were distinct; the mesenchymal, ductal and beta cell types were found to have more than one resulting clusters (Fig3). Moreover, all clusters did not have 10 marker genes that passed the recommended threshold minimums. In fact, just like the findings of Baron et al. (2016), this study also used 1 or 2 markers to identify clusters.

For a more robust method of cell type identification, novel markers were identified by using less stringent thresholds for the FindAllMarkers function. 9311 markers having both positive and negative differential expression, a minimum with a minimum log2 fold change value of 0.25, and a minimum detection frequency of 25% in all the cells were identified. Novel markers reported are statistically significant at the 0.5% level (adjusted p-value <0.005). Resulting marker genes identified using recommended thresholds as well as novel markers were visualized using a clustered heatmap (Fig4, Fig5), similar to the methods described in the paper.

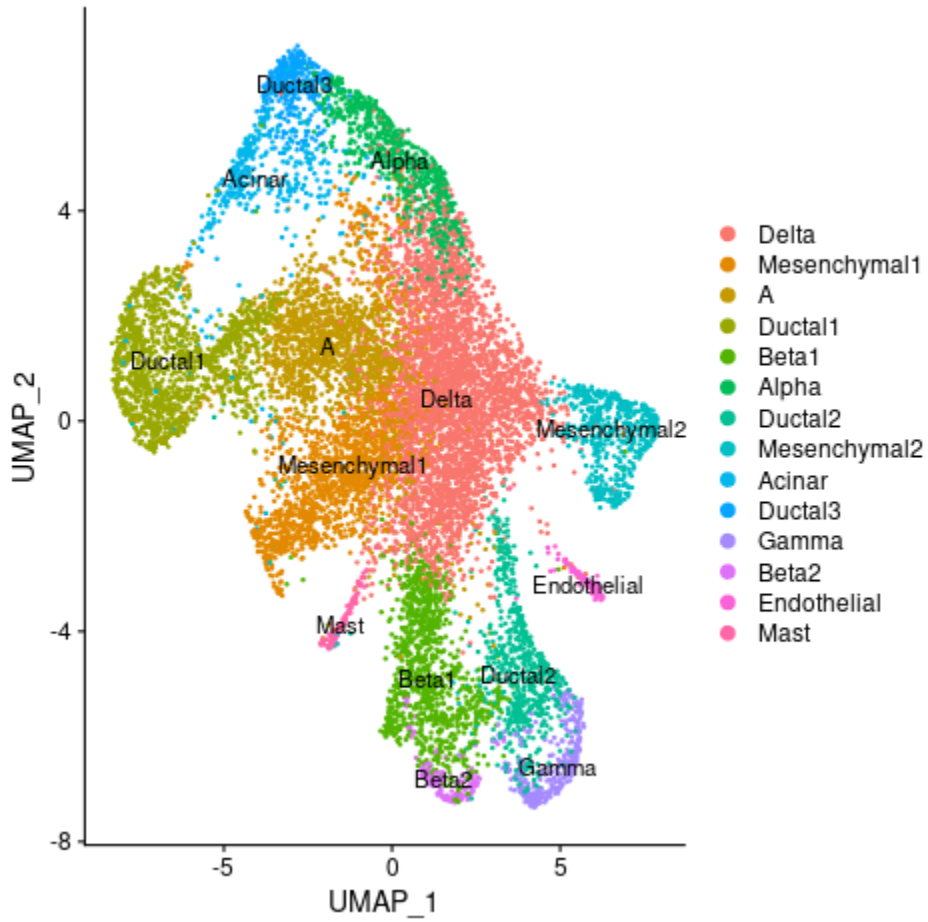

Figure 3: Umap showing clustering of human pancreatic cell types

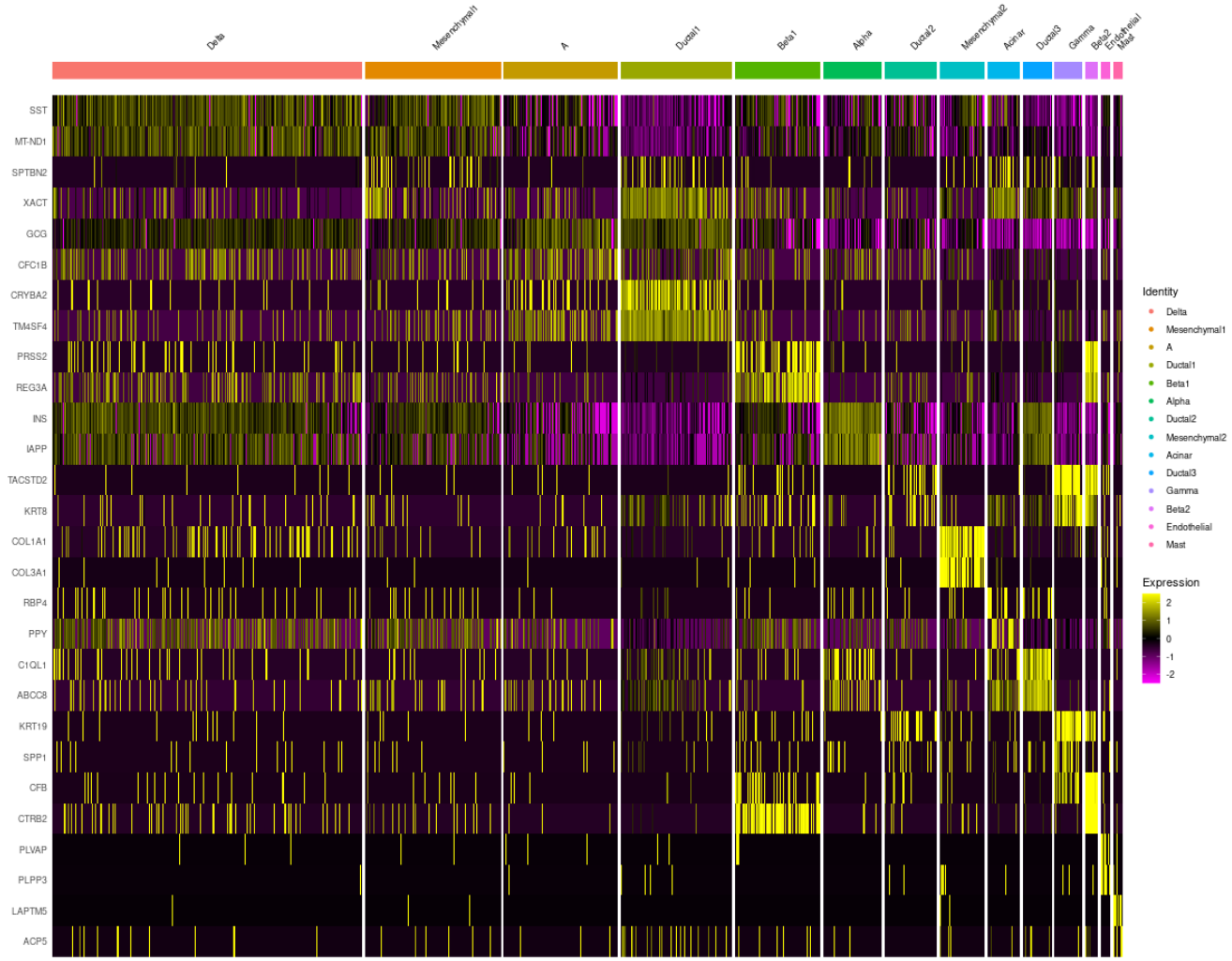

Figure 4: Distribution of markers in each cell type. Markers were determined using following threshold parameters: only.pos = TRUE, min.pct = 0.25, logfc.threshold = 0.25.

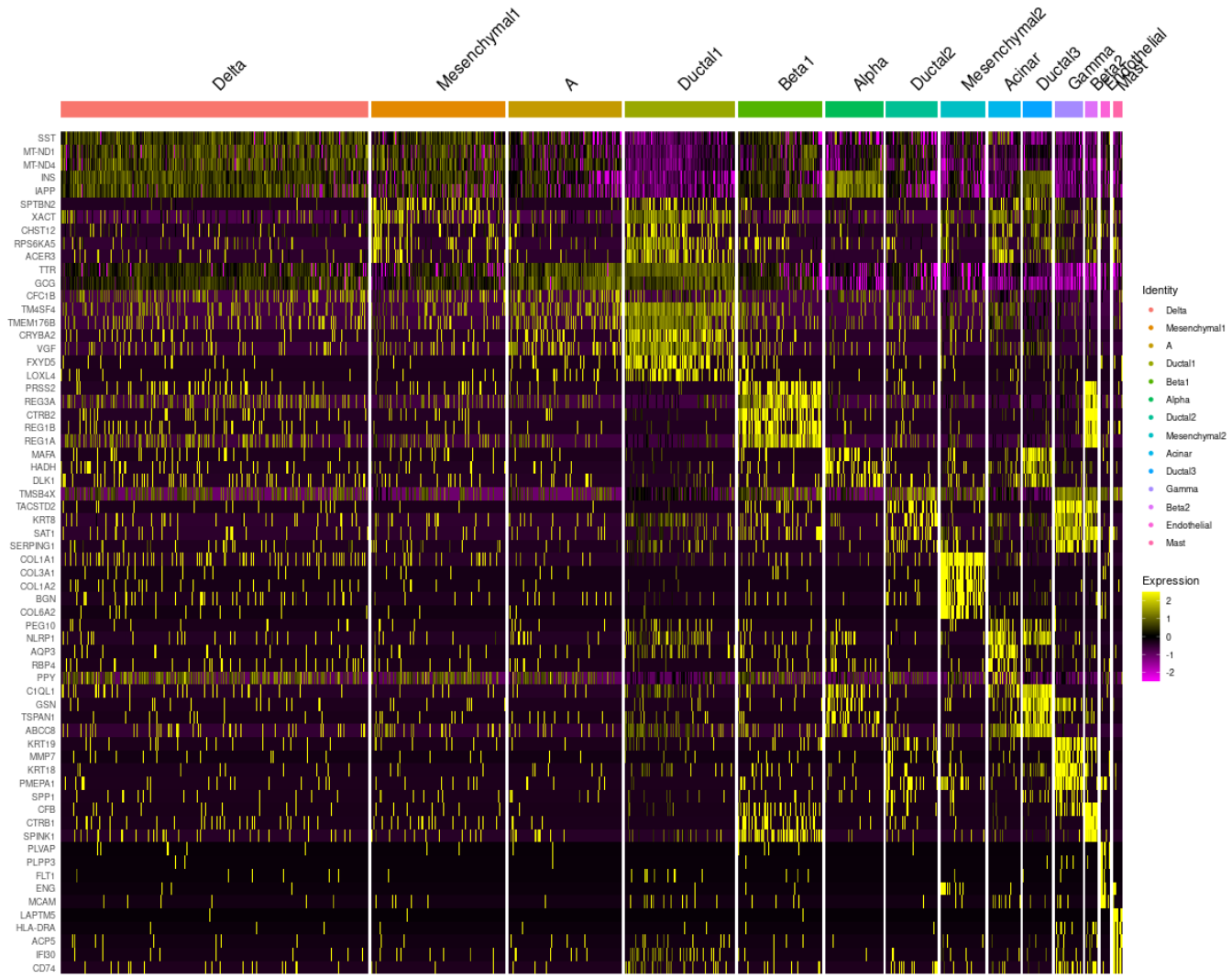

Figure 5: Distribution of novel markers in each cell type. Markers were determined using following threshold parameters: only.pos = FALSE, min.pct = 0.25, logfc.threshold = 0.25 & p_adj <0.005.

		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	Delta	202	SRP-dependent	41	mitochondrial

			cotranslational protein targeting to membrane, cytoplasmic translation, translational initiation		respiratory chain complex I assembly, mitochondrial electron transport, NADH to ubiquinone, electron transport coupled proton transport
1	Mesenchymal1	60	cytoplasmic translation, cytosolic small ribosomal subunit	21	Respiratory chain, Electron transport, Digestion, mitochondrial electron transport, NADH to ubiquinone
2	A (Pancreatic Stellate Cells)	140	Leber hereditary optic neuropathy, mitochondrial ATP synthesis coupled proton transport, Diabetic cardiomyopathy	76	Extracellular Region, Cell Adhesion, Extracellular Matrix
3	Ductal1	411	cytoplasmic translation, cytosolic ribosome, structural constituent of ribosome	57	Leber hereditary optic neuropathy, mitochondrial membrane, Electron transport
4	Beta1	173	Acetylation, neutrophil degranulation, translational initiation	100	protein localization to secretory granule, insulin secretion, virion assembly
5	Alpha	2027	SRP-dependent cotranslational protein targeting to membrane, cytoplasmic translation,	1788	insulin processing, cellular protein metabolic process, negative regulation of endopeptidase activity

			translational initiation		
6	Ductal2	182	cytosolic ribosome, cytoplasmic translation, structural constituent of ribosome, translation	38	platelet aggregation, Host-virus interaction, Regulation of actin cytoskeleton
7	Mesenchymal2	481	Translation, cytoplasmic translation, cell adhesion	383	protein binding, Phosphoprotein, cytosol
8	Acinar	730	SRP-dependent cotranslational protein targeting to membrane, cytoplasmic translation, viral transcription	642	SRP-dependent cotranslational protein targeting to membrane, cytoplasmic translation, viral transcription
9	Ductal3	1576	cytosolic ribosome, cytoplasmic translation, structural constituent of ribosome, translation	1367	platelet aggregation, Host-virus interaction, Regulation of actin cytoskeleton
10	Gamma	1245	extracellular exosome, cytosolic ribosome, cytoplasmic translation	1049	cytosolic ribosome, cytoplasmic translation
11	Beta2	1668	Acetylation, neutrophil degranulation, translational initiation	1385	protein localization to secretory granule, insulin secretion, virion assembly

12	Endothelial	207	Angiogenesis, protein binding, Phosphoprotein, plasma membrane	183	Angiogenesis, protein binding, Phosphoprotein, plasma membrane
13	Mast	209	SRP-dependent cotranslational protein targeting to membrane, translational initiation, cytoplasmic translation	189	electron transport coupled proton transport, ATP synthesis coupled electron transport, aerobic respiration

Table 1. 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
Delta	Contribute to the production of somatostatin cells, which regulate the release of various hormones
Mesenchymal	Supportive cells of connective tissue involved in tissue repair, remodeling, and immune response modulation
A (Pancreatic A Cell)	Secrete glucagon to regulate glucose levels in the blood
Ductal	Epithelial cells forming ducts in various organs, involved in transport and secretion of substances
Beta	Regulate insulin secretion and play a crucial role in glucose homeostasis
Alpha	Secrete glucagon to regulate glucose levels in the blood
Acinar	Responsible for storage, synthesis, and secretion of digestive enzymes in exocrine

	glands
Gamma	Generic designation, more specific information needed to provide a function
Endothelial	Cells lining the interior of blood vessels, involved in vascular functions such as blood flow regulation and exchange of nutrients and waste
Mast	Play a critical role in the inflammatory response and allergic reactions

Table 2. Summarization of the Cell Types and their corresponding functions.

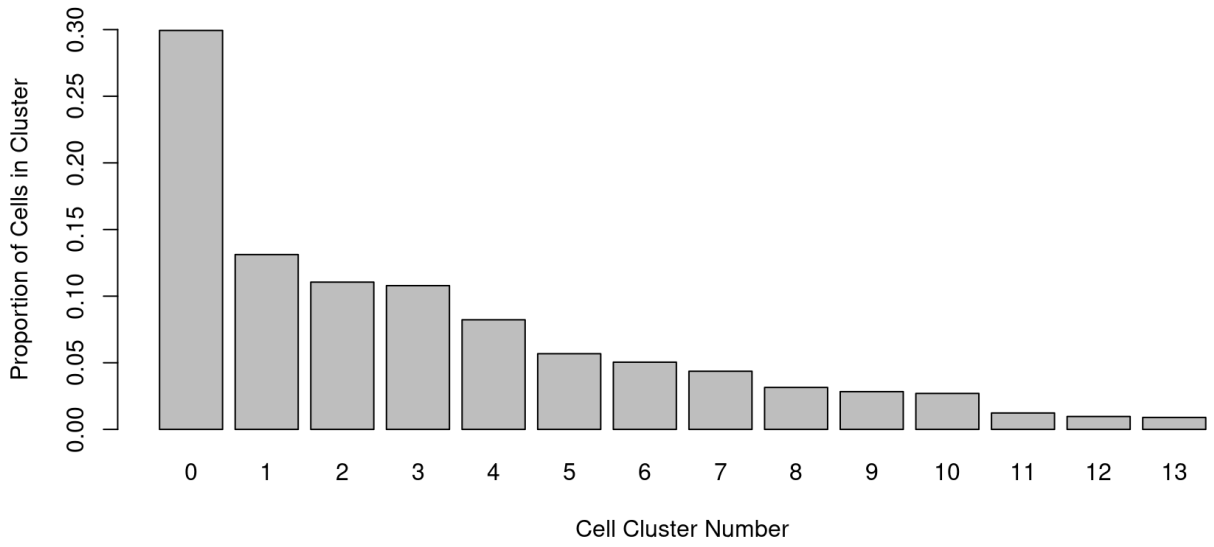

Figure 6: This bar plot illustrates the proportion of cells that were found in each cluster. The figure was generated before clusters were labeled.

Discussion:

Gene set enrichment analysis provided additional evidence regarding the cell types assigned to specific gene clusters. In general, many cell clusters showed enrichment for terms related to the extracellular region and secretion of various products. These findings align with the expectation that pancreatic cells secrete enzymes and hormones into the extracellular space for digestion and signaling purposes. After thorough comparison, unfortunately only 10 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.
Acinar cells, responsible for producing digestive enzymes, exhibited enrichment for genes associated with the extracellular region and ribosomes, reflecting their role in enzyme production. Ductal cells, known to secrete bicarbonate and assist in enzyme transport, displayed enrichment for terms linked to secretory vesicles, potentially involved in enzyme transportation.

The islet category of pancreatic cells comprises alpha, beta, delta, gamma, and epsilon cells, which secrete hormones regulating blood sugar levels. Epsilon cells were not seen as a cluster. We discovered that the alpha and beta clusters were enriched in terms related to hormone metabolic processes, as expected for islet cells. Interestingly, the alpha and delta clusters showed enrichment for terms associated with ATP production and mitochondria. This suggests that these cells may provide energy for other islet cells to carry out hormone production and other vital functions. Resolving the islet cell types proved challenging as they perform similar functions.

The final two cell labels, endothelial cells and mast cells, exhibited distinct functional annotation analysis evidence supporting their respective assignments. Endothelial cells, implicated in angiogenesis with pancreatic tumors, displayed enrichment for terms related to blood vessel development and cellular adhesion. Mast cells, an immune cell type, were enriched in immune response-related terms.

The identification of specific cell types, such as Alpha, Mast, Delta, and Beta, posed certain uncertainties. Additionally, replicating the results from Baron et al.'s 2016 study proved challenging. It is important to consider that different Seurat models were employed in obtaining these results, which may lead to inconsistencies in figures, including the number of clusters. Gene set enrichment analysis offered some evidence supporting the assignment of clusters as pancreatic stellate cells and mast cells. However, distinguishing other cell clusters becomes more intricate due to their general function of secreting substances into the extracellular space. Given the similarity in basic cellular functions, these clusters present greater difficulty in differentiation through functional clustering analysis.

It is worth noting that single-cell RNA sequencing is a relatively new technology, and there are currently no established and efficient algorithms for classifying genes into specific cell types. Future advancements in this field are expected to improve the accuracy and precision of gene classifications. Additionally, having a more comprehensive understanding of cell types will contribute to better and more accurate gene classifications in the future.

In conclusion, while the analysis presented some discrepancies between the determined cell type labels and the associated GO terms, the identification of novel marker genes within each cluster still holds value. In fact, this analysis creates a resource for the discovery of novel cell type-specific transcription factors, signaling receptors, and medically relevant genes as does the original study Barton et al. (2016) that inspired our analysis. As the field of single-cell RNA

sequencing advances, improvements in classification algorithms and a deeper understanding of cell types will lead to more accurate gene classifications and enhance our knowledge of cellular functions.

Works Cited

Baron M, Veres A, Wolock SL, et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst*. 2016;3(4):346-360.e4. doi:10.1016/j.cels.2016.08.011

BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma and Wolfgang Huber, Bioinformatics 21, 3439-3440 (2005).

Chan Zuckerberg Initiative. (2021). Cellxgene. Retrieved from https://cellxgene.cziscience.com/

Charlotte Soneson, Michael I. Love, Mark D. Robinson (2015): Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* http://dx.doi.org/10.12688/f1000research.7563.1

Huang, D.W., Sherman, B.T., Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 37, 1-13 (2009).

Huang, D.W., Sherman, B.T., Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc*. 4, 44-57 (2009).

Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Steffen Durinck, Paul T. Spellman, Ewan Birney and Wolfgang Huber, Nature Protocols 4, 1184-1191 (2009).

Malhotra, A.; Das, S.; Rai, S.N. Analysis of Single-Cell RNA-Sequencing Data: A Step-by-Step Guide. *BioMedInformatics* 2022, *2*, 43-61. https://doi.org/10.3390/biomedinformatics2010003

Satija Lab. Seurat - Guided Clustering Tutorial, 2020, https://satijalab.org/seurat/archive/v3.1/pbmc3k_tutorial.html

Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc. 2018;13(4):599-604. doi:10.1038/nprot.2017.149.

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

Srivastava A, Love M (2023). *eds: eds: Low-level reader for Alevin EDS format*. R package version 1.2.0, https://github.com/mikelove/eds.

Tang, X., Huang, Y., Lei, J. *et al.* The single-cell sequencing: new developments and medical applications. *Cell Biosci* 9, 53 (2019). https://doi.org/10.1186/s13578-019-0314-y

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Zhang, Y., Wang, D., Peng, M. *et al.* Single‑cell RNA sequencing in cancer research. *J Exp Clin Cancer Res* 40, 81 (2021). https://doi.org/10.1186/s13046-021-01874-1

Zilionis, R., Nainys, J., Veres, A. *et al.* Single-cell barcoding and sequencing using droplet microfluidics. *Nat Protoc* 12, 44–73 (2017). https://doi.org/10.1038/nprot.2016.154