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

Group 1

TA : Kritika Karri

Divya Venkatraman (Data Curator)
Garima Lohani (Biologist)
Marlene Tejeda (Analyst)
Xudong Han (Programmer)

INTRODUCTION:

Most organs consist of a variety of cell types with interdependent functions. To understand organ function and disease, genome-wide information on each cell type is crucial [1]. The pancreas is a vertebrate-specific organ and made from two distinct components: the exocrine pancreas and the endocrine islets, which are composed of a diverse set of cell types [2]. The cell types within the pancreas can be divided into two classes: endocrine, responsible for the regulation of blood sugar homeostasis, and exocrine cells, responsible for producing and secreting large amounts of digestive enzymes [3]. To study heterogeneity and classify previously known cell types as well as rare and novel cell type subpopulations within these pancreatic cell types, the single-cell resolution is essential. In this project, we used raw single-cell sequencing data from a 51-year-old female donor in a former experiment [4] and performed further analyses to recognize the cell type, identify the marker genes and develop the in-depth biological functions for these cell types.

DATA:

The reference paper implemented a droplet-based, single-cell RNA-seq method to determine the transcriptomes of over 12,000 individual pancreatic cells from four human donors and two mouse strains [4]. For our project, we used the sample of one of the four human donors. The sample was of a 51 year old female donor. There were 3 sequencing libraries corresponding to the donor which had the following accessions; SRR3879604, SRR3879605 and SRR3879606. The libraries contained paired end reads that were used to produce the Unique Molecular Identifier (UMI) counts matrix.

The raw read 1 barcodes were preprocessed to have a specific format for the reads. The reads in the processed files were given as 19 barcode bases followed by 6 UMI bases. Each unique barcode represents a cell and each unique UMI represents a molecule. The number of reads per barcode was determined by counting the frequency of each barcode within each file using awk. The average number of reads per barcode in each of the raw read 1 files was approximately 30,000 reads. Sequencing depth of approximately 50,000 genes can be informative for gene expression in heterogeneous cells in a biological sample. [6]. The threshold used was the average number of reads found which was approximately 30,000 for each library to filter out barcodes that were too infrequent to be informative. The unique barcodes that passed the threshold number of reads were written to a file to be used as a whitelist in further processing.

We used salmon alevin to generate the UMI counts matrix that was used in further analysis. Salmon alevin has algorithms for quantification and analysis of 3' tagged-end single-cell sequencing data. Given the transcriptome and the raw read files, alevin generates a cell-by-gene count matrix.[7] Alevin requires an index for the transcriptome as input to the command. We generated the index for the current human reference transcriptome GRCh38 available on the GENCODE website using salmon. [8][9] In order to map the transcripts to genes, we generated a transcript to gene mapping file that was fed to alevin to produce the cell-by-gene count matrix. The transcript to gene map file was created using the GTF file of gene annotation on the primary assembly sequence regions available on GENCODE.

Since our barcodes and UMI had a custom format, we had to pass the following additional parameters to the alevin command to produce the counts matrix ; --end 5 --barcodeLength 19 --umiLength 6. We ran salmon alevin separately for each library and combined the cell-by-gene count matrices into one UMI counts matrix using python. The mapping rate reported for each library SRR3879604, SRR3879605 and SRR3879606 was 33.0681%, 28.0257% , 28.4249% respectively. Table 1 shows the statistics of the alevin output.

Library	**Number of Cells**	**Number of UMI**
SRR3879604	1831	4,565,371
SRR3879605	1770	4,316,517
SRR3879606	1845	4,069,502

Table 1: Number of cells and total number of UMI given by alevin for each library

METHODS:

Cell and Gene filtering:

We used an R package called Seurat to process our UMI matrix and filter out the cells with low quality. First of all, the total number of reads, the number of expressed RNA, and the percentage of mitochondrial genes for each cell were summarized. Cells with expressed RNA ranging from 200 to 2500 and with at most 5% mitochondrial genes were selected. After removing unwanted cells, we employed a global-scaling normalization method "LogNormalize" that normalizes the feature expression measurements for each cell by the total expression. The top 2000 high variance genes were kept in our analysis.

Cell clustering:

After filtering out the low-quality cells and low-variance genes, we performed the following clustering analysis. Firstly, all the genes in each cell were scaled to have means in 0 and variances in 1. Next, we applied a PCA analysis to make a linear dimensionality reduction for all the gene data. The JackStraw procedure and Elbow plot method were used to determine the dimensionality of the dataset [5]. After choosing the dimensionality, the K-nearest neighbor was applied to cluster these cells. At the end of the clustering analysis, we also performed a non-linear dimensional reduction analysis (UMAP method) with the selected dimensionality and visualized these clusters.

Identification of marker genes in each cluster:

After performing cell clustering, the marker genes for each cluster were identified using Seurat. The function FindAllMarkers was used with three parameters to limit testing genes that are only positives, which have a min fraction of 0.25, and an x-fold change difference of 0.25. The top gene of each cluster was used as a marker gene for that cluster.

Labeling of clusters as a cell type based on marker genes and visualization of clustered cells:
The cluster cell type based on the marker genes were identified by using, PANGIODB, which is a database that finds cell types where a certain gene is expressed. Afterward, the clustered cells were visualized using UMAP.

Visualize the top marker genes per cluster:
Each top marker was visualized using a feature plot that shows the expression of the gene in the cluster that it is highly expressed. Additionally, a heatmap was also performed with the top 5 genes of each cluster.

Finding novel marker genes:

These were filtered out using very stringent filters. Such as an x-fold change difference of 0.98 and a p_val_adj of less than 0.0000001. The list was then filtered out to exclude the marker gene of each cluster.

Assigning cell clusters to cell type

After the preprocessing and clustering, the given five clusters were assigned to their cell type. The Single-cell RNA-seq analysis pipeline shows the steps the biologist performed in Figure 1. The top genes were filtered by p-adj value. With the help of supplementary table 2 , the gene markers used in the background studies were also used in our analyses based on their expression level in each cluster. Also, Panglodb was used to determine the cell type for some of the analyses. The gene set enrichment analysis was performed using enrichR to determine the functional annotation of each cluster. Further, the literature search was done to verify each of the cell types and its function.

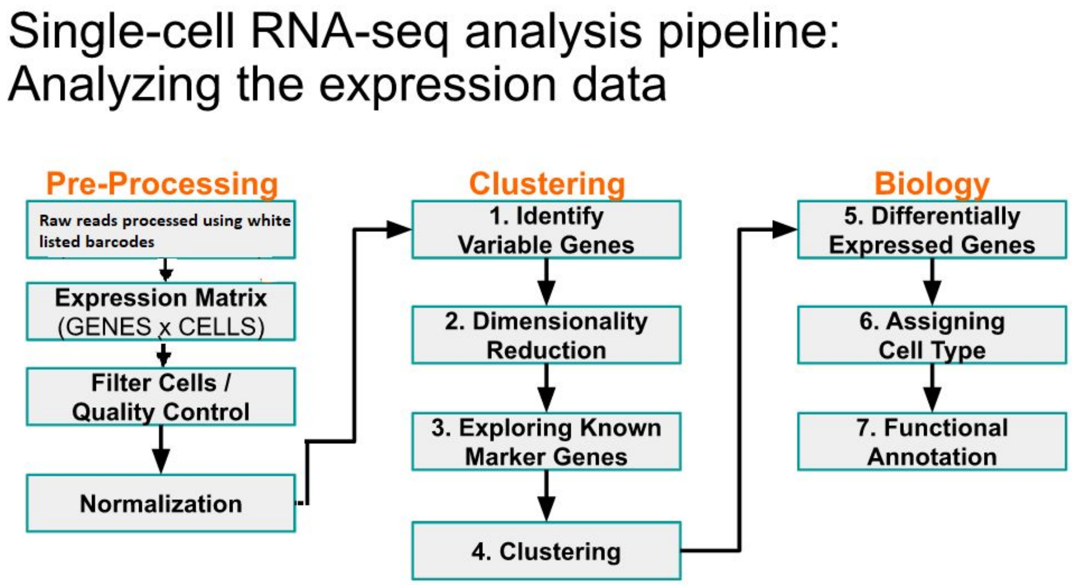

Single-cell RNA-seq analysis pipeline:
Analyzing the expression data
Pre-Processing
Raw reads processed using white listed barcodes
Expression Matrix (GENES x CELLS)
Filter Cells / Quality Control
Normalization
Clustering
1. Identify Variable Genes
2. Dimensionality Reduction
3. Exploring Known Marker Genes
4. Clustering
Biology
5. Differentially Expressed Genes
6. Assigning Cell Type
7. Functional Annotation

Fig1: Schematic diagram of Single-cell RNA seq analysis.(10)

RESULTS:

Quality analysis for the UMI matrix

The UMI matrix shows 5441 cells with 60233 genes in ENSG_ID totally. We converted these ENSGs into Gene Symbol names and merged the duplicated genes to get 55188 distinct genes. Then we filtered out the genes with an abnormal amount of non-expressed genes and a high proportion of mitochondrial genes (see Method). The number of remaining cells is 384, which means most of our cells in the sample dataset are not qualified to be analyzed in the following experiment (Fig1A-B). The reason for that such low quality is primarily caused by the high percentage of mitochondrial genes in most of the cells. If we didn't consider the level of mitochondrial genes, over 3000 cells would be kept. Next, we performed an additional filtering method to keep the top 2000 variance genes for these remaining cells (Fig1C). After the filtering processes, we reduced the dataset to 384 cells with 2000 genes.

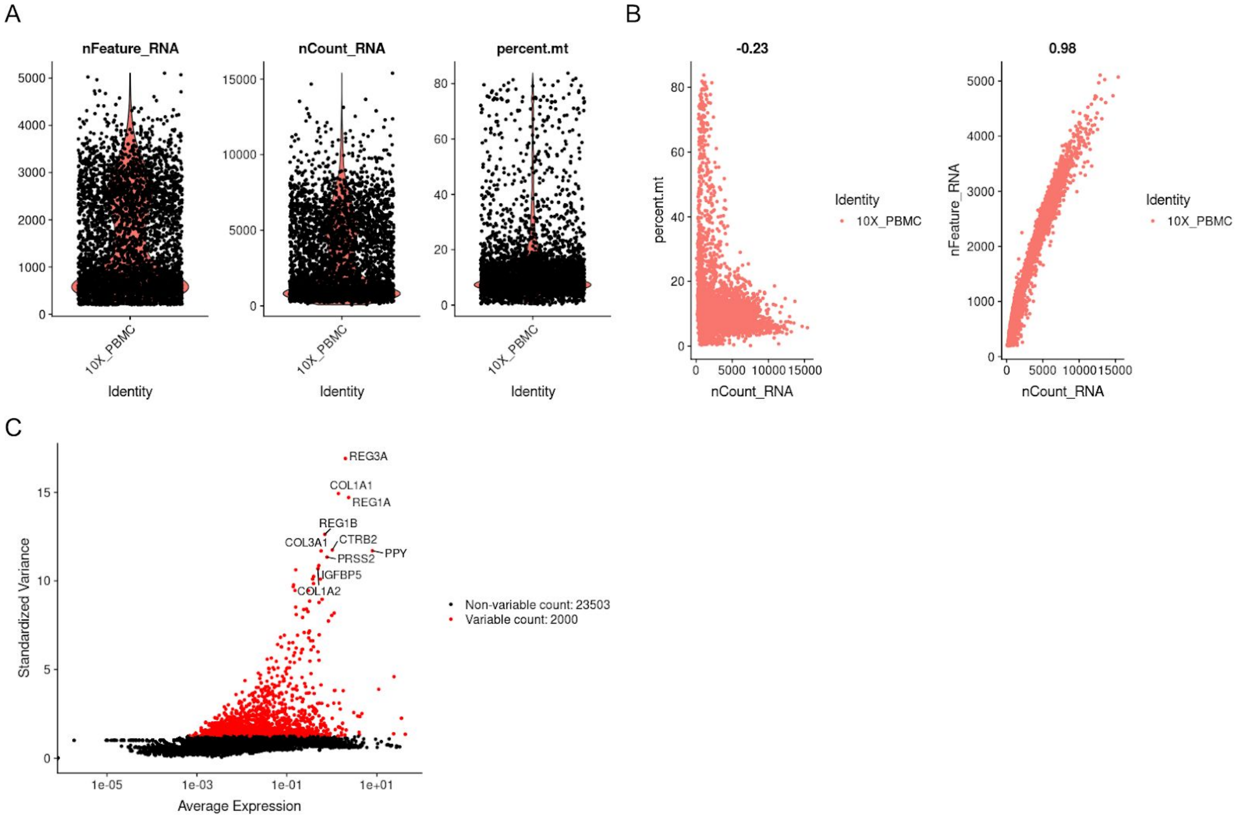
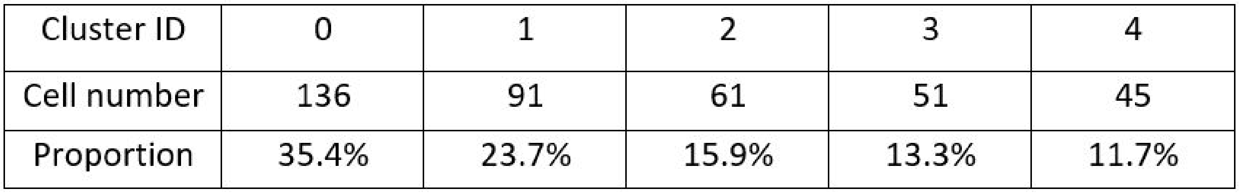


Fig2: The quality analysis for the UMI matrix. A: The violin plots for the number of reads, expressed genes, and percentage of mitochondria genes for each cell. Most of the cells have a proportion of mitochondrial genes more than 5%. B: The scatter plots for the reads-expressed genes and reads-mitochondrial genes relationships. C: The scatter plot for the variances of all the genes. The red dots are the top 2000 high variance genes we kept for the following analyses, the black dots are the genes filtered out.

Dimensional reduction and Cell Clustering

To cluster these 384 cells, we normalized the dataset and performed a linear dimensionality reduction and non-linear dimensional reduction analyses. In the linear dimensional reduction analysis, the 2000 genes had been divided into subsets and assigned into different principal components (PCs). Fig 2A shows the genes in the PC1 and PC2. In the heatmap of PC1 genes (Fig 2B), they have two different patterns significantly, and other PCs such as PC2, PC3, and PC6 also show similar patterns (FigS1). To determine the dimensionality for the reduced dataset, we built the JackStraw plot and the Elbow plot (Fig2C-D). With the combination of these two plots, we found an ‘elbow’ around PC6-7. Consequently, we chose the dimensionality as 7. In the non-linear dimensional reduction analysis, we used the UMAP method and reduced the dataset into 7-dimension. These 384 cells are clustered into 5 groups (Fig2E and Table1).



Cluster ID	0	1	2	3	4
Cell number	136	91	61	51	45
Proportion	35.4%	23.7%	15.9%	13.3%	11.7%

Table2: The number and proportion of cells in each cluster after the clustering. The cluster ID corresponds to the ID in the Fig2E.

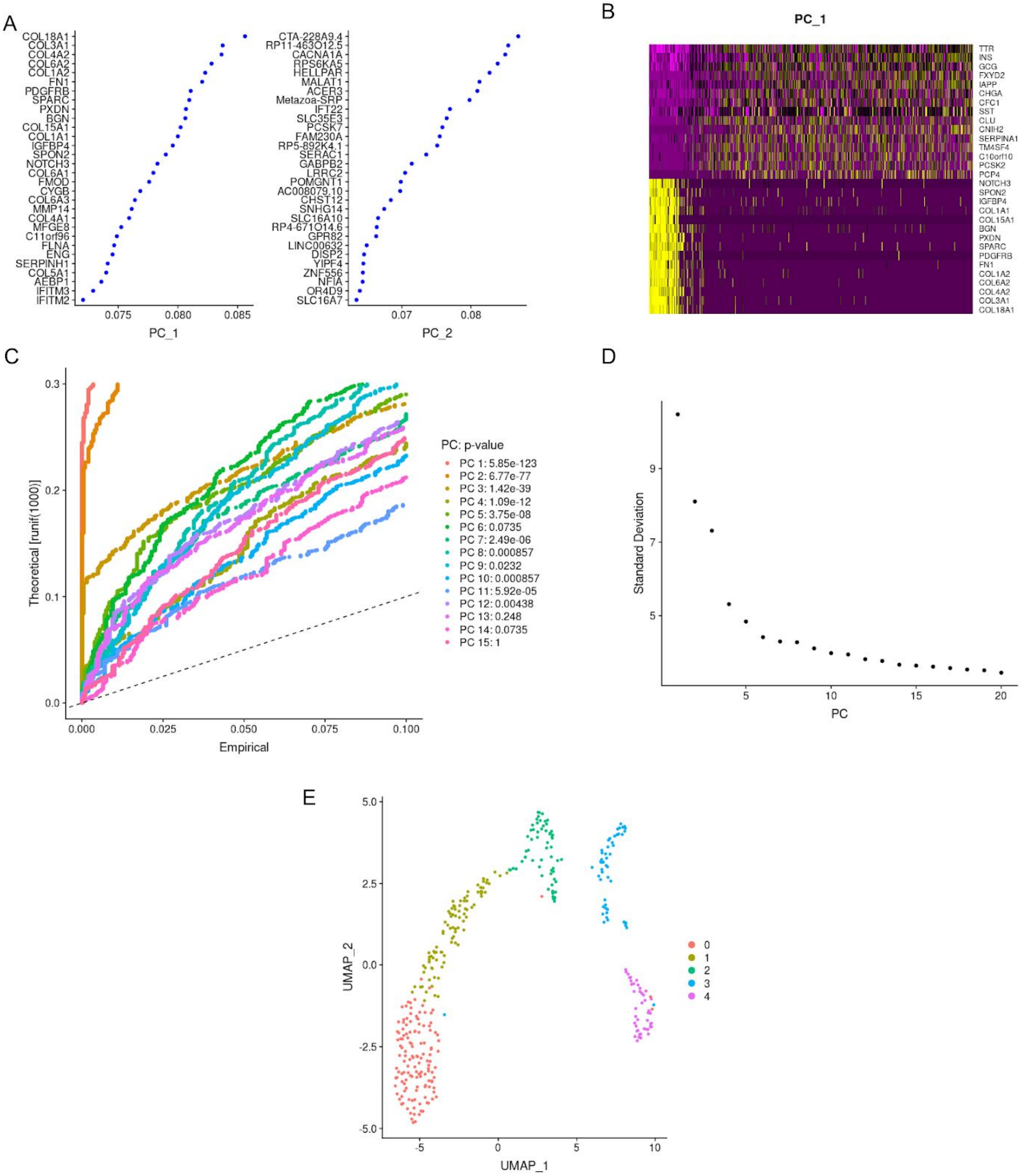
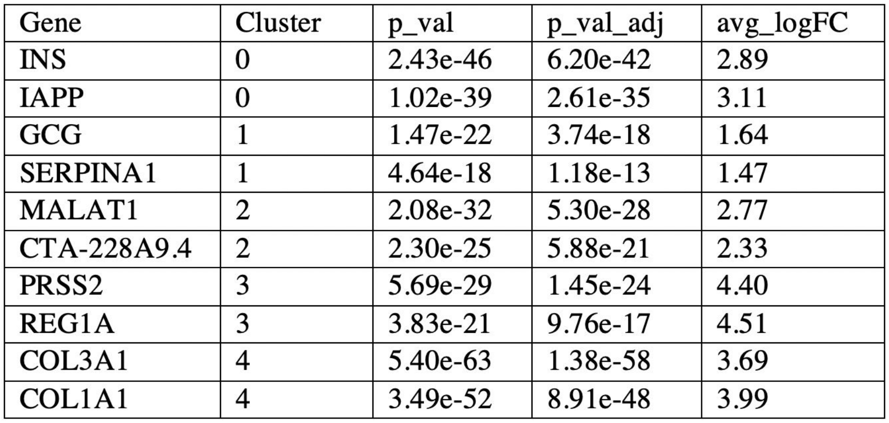


Fig3: The dimensional reduction and cell clustering results for the filtered dataset. A: The scatter plot for gene subsets in PC1 and PC2. Each principal component has a positive gene group and a negative gene group. B: The heatmap for all the genes in PC1. Each column represents a cell. The expression pattern of these genes can be divided into two groups (upregulated or downregulated in these cells). C: The JackStraw plot for each PC. The p-value for each PC is shown in the plot. D: the Elbow plot for each PC. The vertical axis means the standard deviation for every PC. From Fig C and D, we can determine the dimensionality for the reduced dataset is 7. E: The clustering result under the UMAP method. The 384 cells are clustered into 5 groups.

Identification of marker genes in each cluster

A



Gene	Cluster	p_val	p_val_adj	avg_logFC
INS	0	2.43e-46	6.20e-42	2.89
IAPP	0	1.02e-39	2.61e-35	3.11
GCG	1	1.47e-22	3.74e-18	1.64
SERPINA1	1	4.64e-18	1.18e-13	1.47
MALAT1	2	2.08e-32	5.30e-28	2.77
CTA-228A9.4	2	2.30e-25	5.88e-21	2.33
PRSS2	3	5.69e-29	1.45e-24	4.40
REG1A	3	3.83e-21	9.76e-17	4.51
COL3A1	4	5.40e-63	1.38e-58	3.69
COL1A1	4	3.49e-52	8.91e-48	3.99

Fig4: Top differentially expressed genes per cluster. A. Top 2 differentially expressed genes per cluster.

Visualization of clustered cells

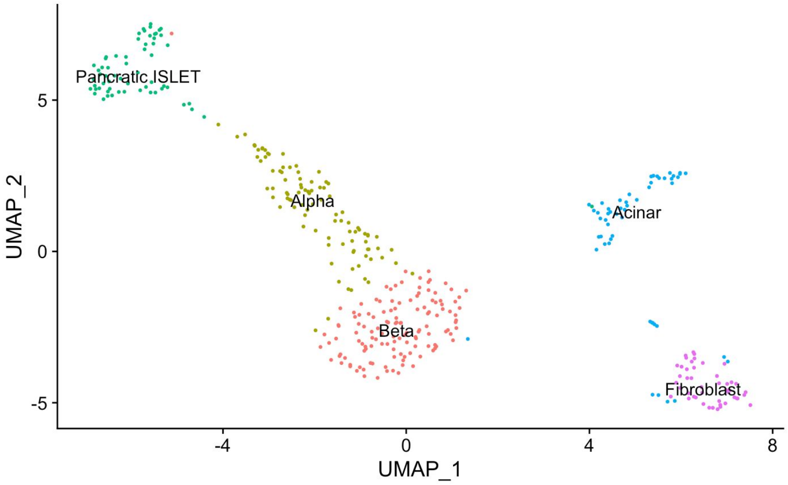


Fig5: With the exception of few outliers in clusters 0, 2, and 3 all cells are cleanly separated by cell type. There are a total of five clusters.

Visualization of the top marker genes per cluster

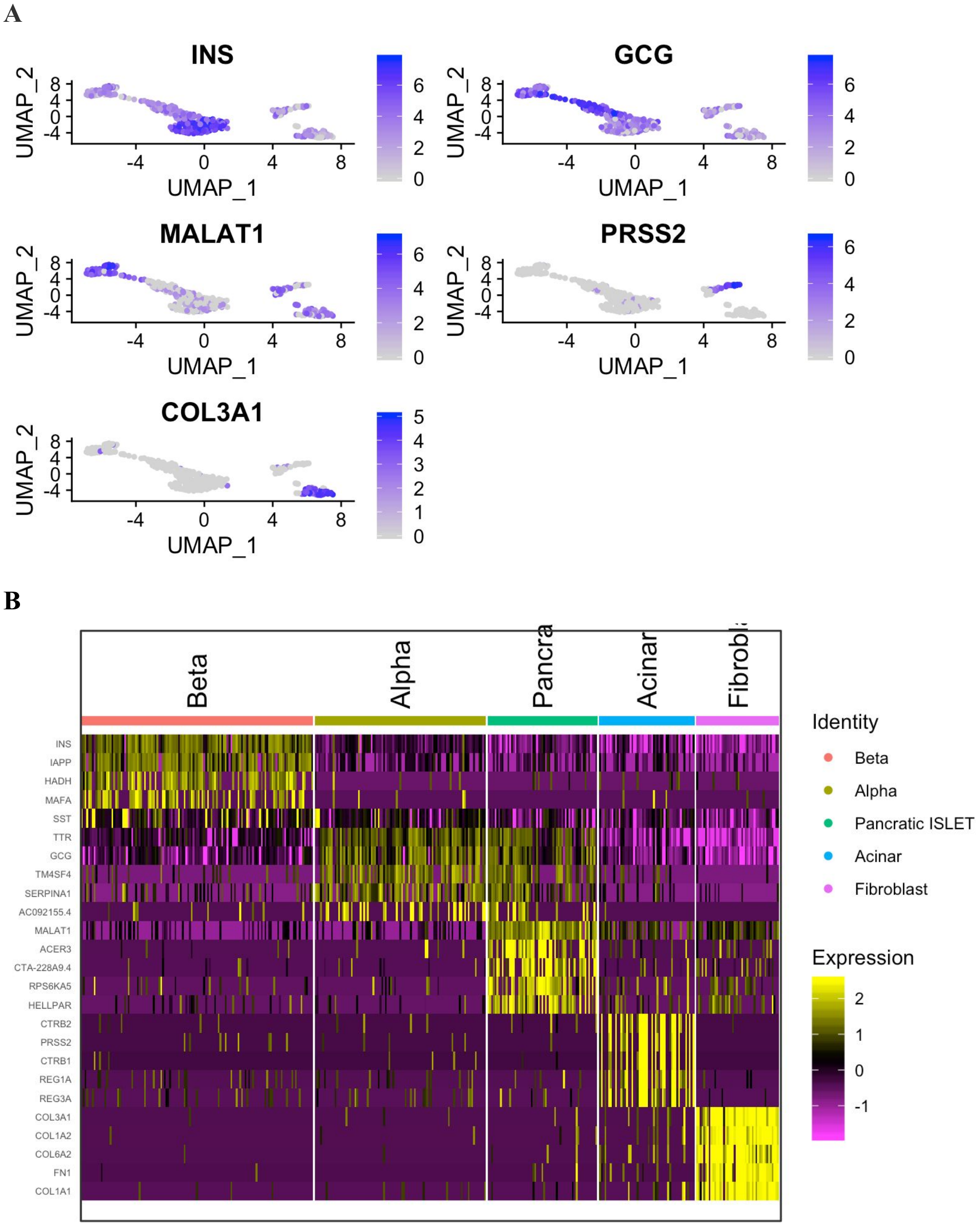


Fig6: A. FeaturePlot of the expression of each top marker. B. Heatmap of the top 5 highly expressed genes in each cluster.

As we expect, FeaturePlot indicates that each of the marker genes is highly expressed in its corresponding cluster. With the exception of INS (cluster 0 marker) and GCG (cluster 1 marker) which are expressed in many clusters, they are highly expressed in their corresponding clusters. Heatmaps further support the UMAP illustration, where we notice that there is a clear separation of the cells as the top five differentially expressed genes for each cluster are clustered together in the heat map. There are a few genes that are highly expressed in various clusters. Such as GCG that although it is the marker for cluster 1 we can see it is highly expressed in cluster 2 which we also observe in the FeaturePlot. This could be novel marker genes that have many functions.

Novel marker genes

A

Gene	Cluster	p_value	avg_logFC	p_val_adj
HADH	Beta	2.36E-37	2.033179	6.02E-33
MAFA	Beta	1.54E-28	1.903205	3.92E-24
DLK1	Beta	1.02E-13	1.597115	2.59E-09
TTR	Alpha	4.79E-24	1.361413	1.22E-19
TM4SF4	Alpha	8.90E-21	1.316566	2.27E-16
CRYBA2	Alpha	1.92E-12	1.114487	4.89E-08
ACER3	Pancreatic ISLET	3.99E-30	2.24027	1.02E-25
HELLPAR	Pancreatic ISLET	2.61E-22	2.154127	6.66E-18
RPS6KA5	Pancreatic ISLET	1.00E-24	2.110948	2.56E-20
REG3A	Acinar	1.99E-19	4.35626	5.08E-15
CTRB2	Acinar	4.80E-29	4.330982	1.22E-24
CTRB1	Acinar	3.01E-27	4.173712	7.68E-23
COL1A2	Fibroblast	6.80E-62	3.466033	1.74E-57
FN1	Fibroblast	3.56E-58	3.453031	9.07E-54
COL6A2	Fibroblast	5.91E-60	3.436518	1.51E-55

B

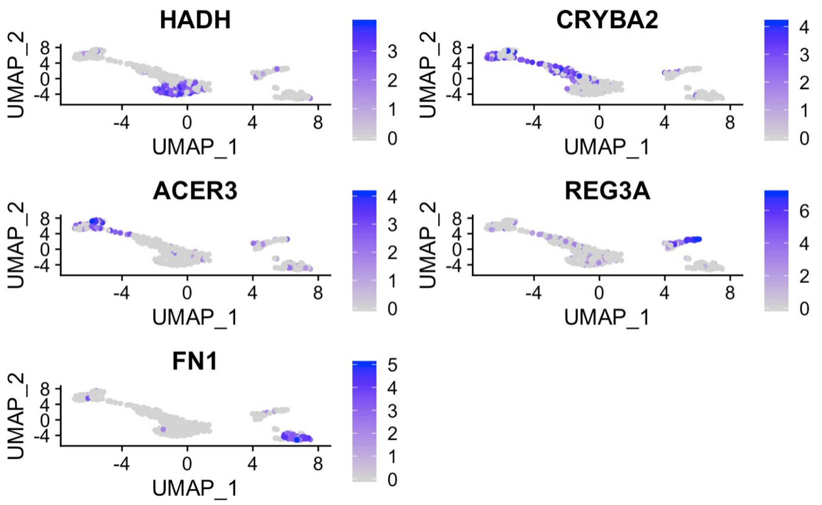


Fig7: A. Table showing novel marker genes in each cluster. B. FeaturePlot of the expression of one novel gene in each cluster. Blue indicates high gene expression which is what we expect for each of these clusters and low expression in clusters where it is not expressed.

The biological analyses were done and the get-set enrichment results were formatted into table3.

Cluster	**Gene marker**	**Cell type**	**EnrichR Terms (pathways)**	**EnrichR score**
0	INS	Beta cells	Gene expression regulation in pancreatic beta cells	2.296e-7
1	GCG	Alpha cells	Glucagon signalling in metabolic regulation	0.3178
2	GCG,DPP4, LOXL4,GLS,MAFB	Alpha Cells	HNF3A pathway	0.8184
3	CPA1	Acinar cells	Pancreatic Secretion	0.00001462
4	PDGFRB	Stellate cells	PDGF genes and receptors	0.002997

Table3: The result of gene set enrichment on the marker genes for each cluster

DISCUSSION:

After the analysis, we found cluster 0 to be beta cells because INS genes appeared as the top marker with p-adj value of 6.20E-42 and fold change value of 2.89. The INS gene is known to provide instructions for producing the hormone insulin.(11) In paper we found they used the INS gene in their analyses for determining beta cells. In the EnrichR, INS could be associated with the pathway gene expression regulation in the pancreatic beta cells. Also, enriched terms such as type 1 diabetes appeared which can be related with the loss of beta cells.(11) So it highly agrees with the function of beta cells which is to secrete insulin and regulate blood glucose level.

We found a high expression of the GCG gene in Cluster 1. GCG is known to lower extracellular glucose in humans .(13) In this paper GCG was also used for analyses corresponding to alpha cell.In Enrichr term, we found GCG gene to be associated with pathway glucagon signalling in metabolic regulation. Moreover, the main function of an alpha cell is to increase glucose level in the blood.(13) So it aligns with the functionality of the alpha cells. Further, we found that in cluster 3 the GCG was not as highly expressed as in cluster 1.There were several other alpha cells gene markers such as DPP4, LOXL4, GLS, MAFB described in PangloDB also present in gene cluster 3. So cluster 3 is most likely to be an alpha cell.

In paper we found CPA1 genes to be used in analyses for acinar cells. In our cluster 3, CPA1 gene marker was significantly expressed with a p-adj value of 1.42E-30 and log fold change value of 3.35. The enriched pathway found in EnrichR associated with this gene was the pancreatic secretion pathway. This is in agreement with the function of acinar cells which are known to secrete digestive enzymes.(12) So cluster 3 strongly suggests to be acinar cells.

We also found stellate cells in cluster 4. In the paper it was reported that they found stellate cells in each of the donors. They used the PDGFRB gene in analysis for stellate cells. It was highly expressed in cluster 4 with p-adj value of 2.11E-59 and fold change value of 2.67.The author in this paper states that activated stellate cells are known to produce large quantities of collagen,fibronectin and extracellular matrix component.(4) In EnrichR , we found enriched pathways such as collagen biosynthesis and modifying enzymes, extracellular matrix organization and PDGF genes and receptors. So it strongly implies this cluster is activated stellate cells as found in paper.

CONCLUSION:

We were able to analyse the raw single-cell sequencing data from a 51-year-old female donor . We were able to successfully process barcodes of the data. Then we also performed cell/gene quantification of the UMI matrix. Further we also applied quality control on UMI matrix. We found five clusters with their marker genes .And finally identified the cell type of each cluster to give biological significance. Overall, our results were synonymous with the paper.

REFERENCES:

[1] Muraro, Mauro J., et al. "A single-cell transcriptome atlas of the human pancreas." Cell systems 3.4 (2016): 385-394.

[2] Wollny D, Zhao S, Everlien I, et al. Single-cell analysis uncovers clonal acinar cell heterogeneity in the adult pancreas[J]. Developmental cell, 2016, 39(3): 289-301.

[3]Enge, Martin, et al. "Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns." Cell 171.2 (2017): 321-330.

[4]Baron, Maayan, et al. "A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure." Cell systems 3.4 (2016): 346-360.

[5]Macosko, Basu, et al. "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets." Cell, 2017, 161(5):1202-1214.

[6] Streets, Aaron & Huang, Yanyi. (2014). How deep is enough in single-cell RNA-seq?. Nature biotechnology. 32. 1005-1006. 10.1038/nbt.3039.

[7] Srivastava, A., Malik, L., Smith, T. et al. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. Genome Biol 20, 65 (2019).

[8] Frankish, Adam et al. “GENCODE reference annotation for the human and mouse genomes.” Nucleic acids research vol. 47,D1 (2019): D766-D773.

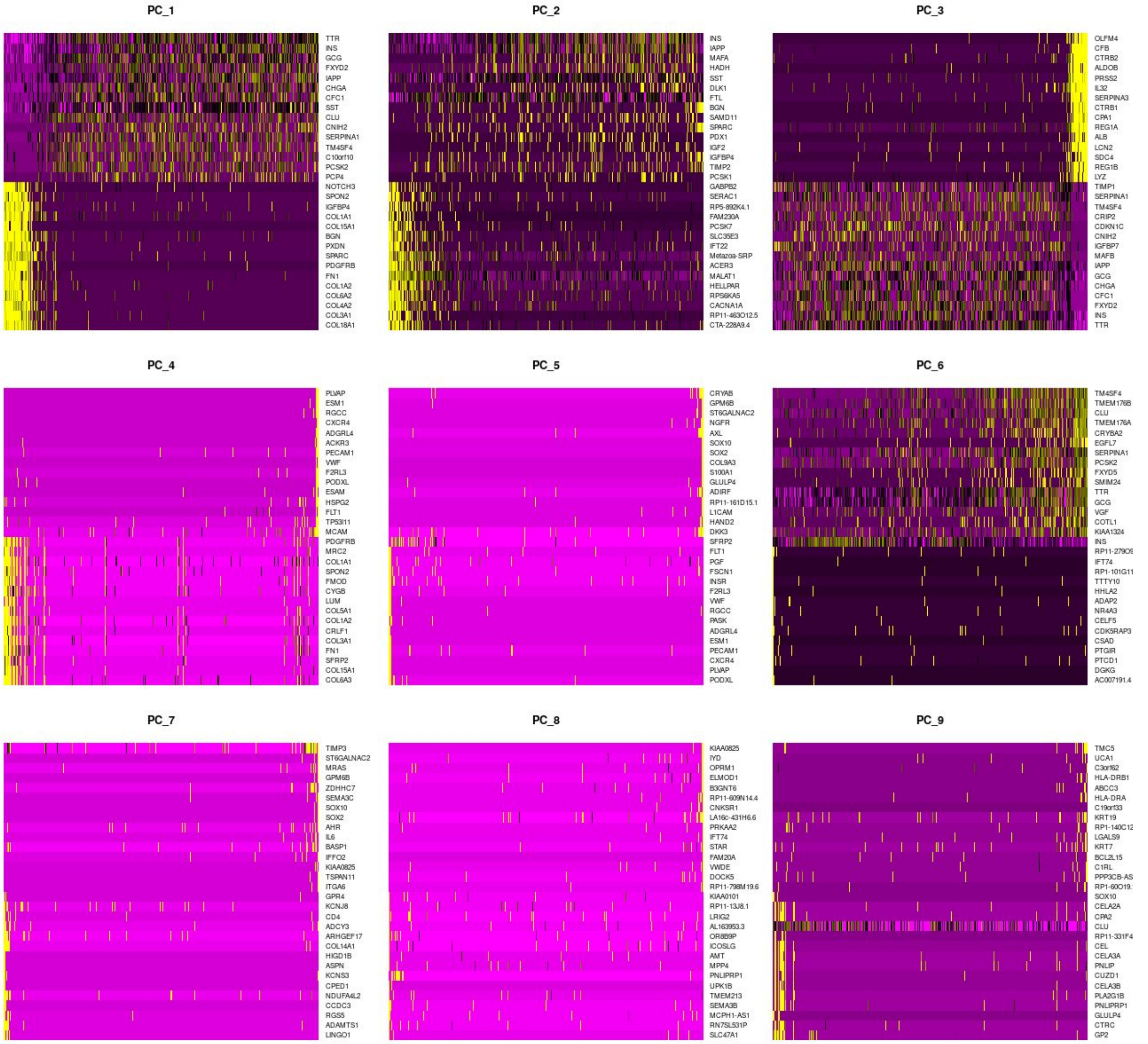
[9] Patro, R., Duggal, G., Love, M. et al. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14, 417–419 (2017)

[10]https://broadinstitute.github.io/2019_scWorkshop/identifying-cell-populations.html#google-slides

[11]Chen C, Cohrs CM, Stertmann J, Bozsak R, Speier S. Human beta cell mass and function in diabetes: Recent advances in knowledge and technologies to understand disease pathogenesis. *Mol Metab*. 2017;6(9):943‐957. Published 2017 Jul 8. doi:10.1016/j.molmet.2017.06.019

[12] Williams JA. Regulation of acinar cell function in the pancreas. *Curr Opin Gastroenterol*. 2010;26(5):478‐483. doi:10.1097/MOG.0b013e32833d11c6

[13]https://www.uniprot.org/uniprot/P01275



FigS1: The heatmaps for the gene subsets in the top 9 PCs.