Project 5 – Individual Project
Single Cell Analysis of Pancreatic Cells
SHREEN KATYAN
ENG BF528 | Spring 2022

INTRODUCTION

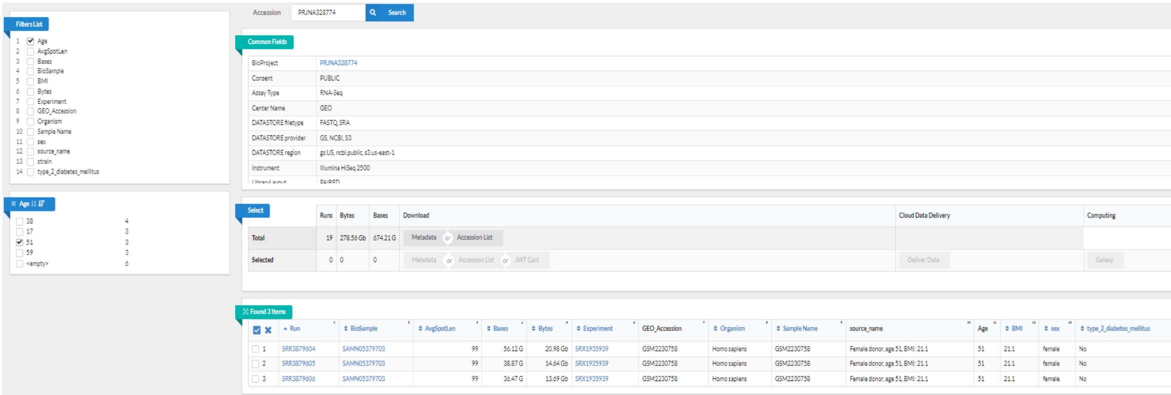
These days there is an increase in the depth studies of diverse cell types owing to the ongoing advancements in RNA-sequencing and single-cell technologies.Baron used a droplet-based single-cell RNA-seq technique to assess the transcriptomes of both human and mouse pancreatic cells in their 2016 paper, A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell population Structure.This study is focused on search for treatments for well-known pancreatic diseases, such as Type I and II Diabetes Mellitus.For our analysis, I re-examined the data from one such individual donor in the study and evaluating the cellular makeup of the corresponding samples in order to corroborate the initial findings.

Being the Programmer before for this project, I decided to be the Data Curator and Analyst this time.

Data curator —

Locate the sample metadata

For our study,out of 13 samples , the SRR files associated with the 51 year old female donor were used.I searched for the SRA (short read archive) selector link on the GEO accession page.The datasets and corresponding metadata are made available via the Gene expression Omnibus under the Accession Number GSE84133.For supplementary data,I even referred authors Baron et al.'s Supplementary Table 1 details to retrieve the ages,BMIs and sexes of the 4 human donors from whom the samples are obtained.It helped in determining the sample ID for our donor.The donor's SRA accession number was SRP07832, which comprised the runs SRR3879604, SRR3879605, and SRR3879606 that were used for further analysis.



Filters List

1 ☑ Age
2 □ AvgSpotLen
3 □ Bases
4 □ BioSample
5 □ BMI
6 □ Bytes
7 □ Experiment
8 □ GEO_Accession
9 □ Organism
10 □ Sample Name
11 □ sex
12 □ source_name
13 □ strain
14 □ type_2_diabetes_mellitus

Age

□ 38	4
□ 17	3
☑ 51	3
□ 59	3
□ <empty>	6

Common Fields

BioProject	PRJNA328774
Consent	PUBLIC
Assay Type	RNA-Seq
Center Name	GEO
DATASTORE filetype	FASTQ, SRA
DATASTORE provider	GS, NCBI, S3
DATASTORE region	gs.US, ncbi.public, s3.us-east-1
Instrument	Illumina HiSeq 2500
LibraryLayout	PAIRED

Select

	Runs	Bytes	Bases	Download	Cloud Data Delivery	Computing
Total	19	278.56 Gb	674.21 G	Metadata or Accession List		
Selected	0	0	0	Metadata or Accession List or JWT Cart	Deliver Data	Galaxy

Found 3 Items

☑ ✖	Run	BioSample	AvgSpotLen	Bases	Bytes	Experiment	GEO_Accession	Organism	Sample Name	source_name	Age	BMI	sex	type_2_diabetes_mellitus
□ 1	SRR3879604	SAMN05379703	99	56.12 G	20.98 Gb	SRX1935939	GSM2230758	Homo sapiens	GSM2230758	Female donor, age 51, BMI: 21.1	51	21.1	female	No
□ 2	SRR3879605	SAMN05379703	99	38.87 G	14.64 Gb	SRX1935939	GSM2230758	Homo sapiens	GSM2230758	Female donor, age 51, BMI: 21.1	51	21.1	female	No
□ 3	SRR3879606	SAMN05379703	99	36.47 G	13.69 Gb	SRX1935939	GSM2230758	Homo sapiens	GSM2230758	Female donor, age 51, BMI: 21.1	51	21.1	female	No

Figure1. represents Short archive files used for our study(51 years old donor)

The raw 1 read barcodes couldn't match the InDrop Barcode schemes due to the noise present in the protocol.Thus the barcode provided were added to the 25 bases count(19 bases with 6 UMI bases).The barcodes with very low frequency (exactly 1 read per se) would be filtered out with remaining frequent barcodes that would be informative were further processed.

Count Number of Reads by Barcodes

The AWK command on the command line was used which iterates through each line of the 3 FASTQ.gz files.The advantage of using this method is that it would count the number of reads by barcodes without decompressing the fastq files.Using these 3 fastq files and the whitelisted barcodes which we created previously via filtering ,I generated the UMI matrix. On examining the plot values, merging runs, deleting non-length 19 barcodes, and further filtering by the cumulative read count inflection point were considered to be the most acceptable and were used for subsequent processing.

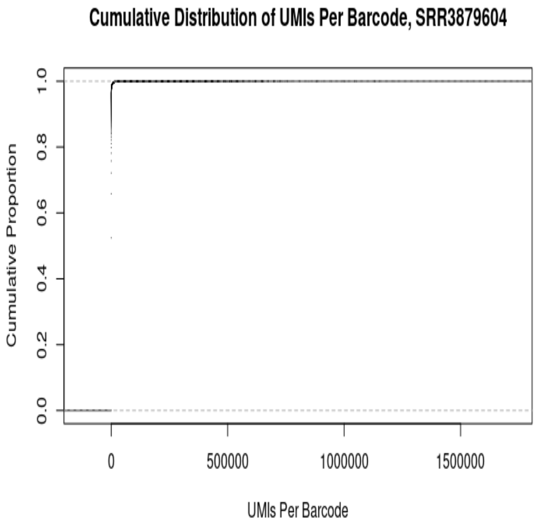


Figure2.a Cumulative distribution plot per barcode for the sample SRR3879604

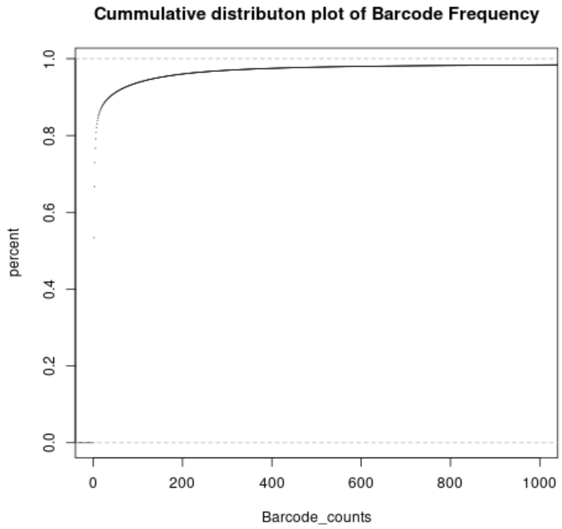


Figure2.b Cumulative distribution plot per barcode for the sample SRR3879605

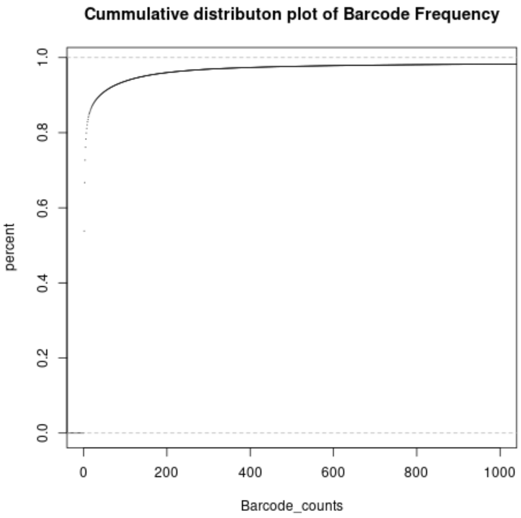


Figure2.c Cumulative distribution plot per barcode for the sample SRR3879606

Whitelist Informative Barcodes

In order to retain the reads which were informative after the differential gene expression ,a whitelist of barcodes which is stored in my directory /projectnb/bf528/students/shreenk. We downloaded gencode.v40.transcripts.fa.gz ie the reference transcriptome and the annotation file(gencode.v40.annotation.gtf) and the genome file(GRch38.p13.genome.fa.gz).We custom barcodes and UMI reads when quantifying the reads —end 5 —barcodeLength 19 —umiLength 6.The —tgMap option takes in a transcript to gene map file of each transcript included in the reference to the corresponding gene, and salmon collapses from the transcript to the gene level.

This is primarily achieved by using using the gene in current reference human transcriptome available ((GRCh38.p13) transcript sequences (ENSGXXX)) which are mapped to the transcript ID (ENSTXX).I derived UMI matrix of the number of reads in a cell originating from the corresponding gene.The Summary statistics of salmon and alevin Output was plotted for further analysis with the help of the library type ISR (inward, stranded, reverse strand). Alevin is a tool within the Salmon software that allows for the quantification and analysis of single-cell sequencing data .

Reporting_Source	**Output Statistics**	**Argument**
Salmon_quant.log	43.4911%	Mapping Score
Salmon_quant.log	245, 900	Index targets
Alevin_output	4189039	Barcodes whitelisted
Alevin_output	4251176	Unique barcodes

Table 1. Summary statistics of salmon and alevin Output

Salmon Alevin requires the raw FASTQ files as input , the previously generated whitelist of barcodes, an index file containing an index of the reference transcriptome which is generated using the “salmon index”, and a transcript-to-gene map, which matches each transcript identifier with a corresponding gene name.

Analyst—

Identify Marker Genes

The Seurat package feature FindAllFeatures command was used to identify Marker genes for each cluster.The positive markers were reported in our study .The criteria for filtering the genes was that genes should that pass a 0.25 log2 fold change threshold.This corresponds to a feature being detected at a minimum of 25% coverage in either of the cell groups in order to be considered .The top ten marker genes for each of the twelve clusters, by average log2FoldChange, were then exported for further analysis.In order to identify cell type of each cluster,we either refer to Baron et al paper or we look in the Human Protein Atlas website.Using the above-generated marker genes, clusters were assigned their corresponding cell types. UMAP projection was used to visualize them after performing the Differential analysis.

Visualizing clustered genes

Umap is non linear,graph based data visualization method for dimensional reduction which is used in single cell RNA sequencing.We identified Marker genes for each of the eleven clusters.In order to rename the labels of each cluster, we used “RenameIdents” function .The DimPlot function was used to plot UMap and the reduction parameter was set to “umap”.

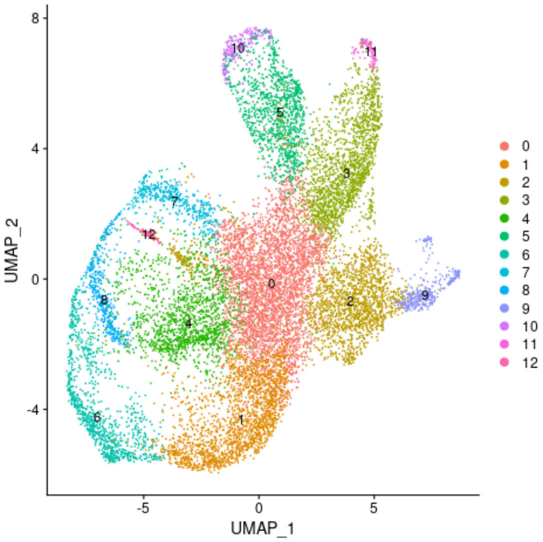


Figure 3. representing UMap plot with color legend according to the cell clusters

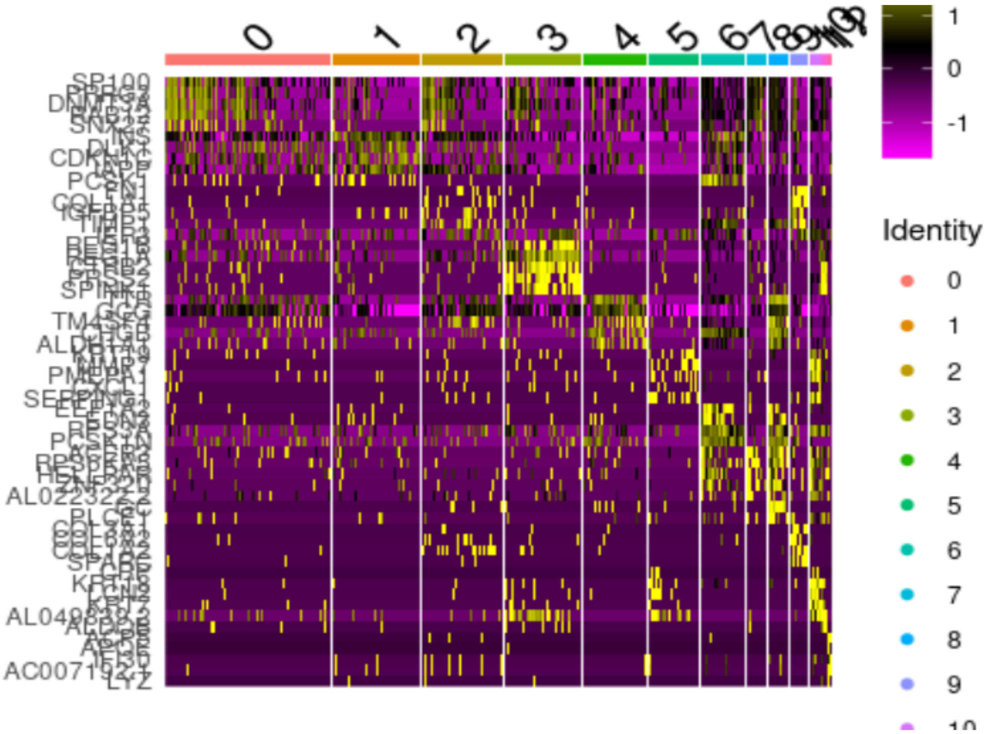
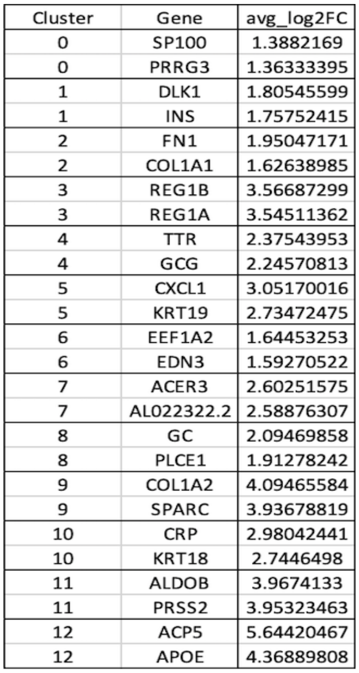


Figure 4. -Clustered heatmap of log normalized UMI counts, showing top ten differentially expressed genes in each of the eleven clusters.The highest expression level of gene is indicated by yellow, whereas the lowest is depicted by purple.

Visualizing top marker genes per cluster and identifying novel marker genes

The heatmap was generated using the top 5 marker genes, by their most significant average log2FoldChange, were exported for each of these clusters. I used the top 2 highest average log2 fold change genes to plot in our study.In order to classify a gene as a novel marker gene,check if the gene shows higher expression probability in one specific gene cluster and wasn't previously identified as a marker gene for that specific cluster.The novel marker genes are as follows-



Cluster	Gene	avg_log2FC
0	SP100	1.3882169
0	PRRG3	1.36333395
1	DLK1	1.80545599
1	INS	1.75752415
2	FN1	1.95047171
2	COL1A1	1.62638985
3	REG1B	3.56687299
3	REG1A	3.54511362
4	TTR	2.37543953
4	GCG	2.24570813
5	CXCL1	3.05170016
5	KRT19	2.73472475
6	EEF1A2	1.64453253
6	EDN3	1.59270522
7	ACER3	2.60251575
7	AL022322.2	2.58876307
8	GC	2.09469858
8	PLCE1	1.91278242
9	COL1A2	4.09465584
9	SPARC	3.93678819
10	CRP	2.98042441
10	KRT18	2.7446498
11	ALDOB	3.9674133
11	PRSS2	3.95323463
12	ACP5	5.64420467
12	APOE	4.36889808

Figure 5.- represents Top 2 highest average log2FC marker genes from each of these clusters.

A violin plot is also generated in addition in order to show the expression probability across the clusters.They determined if the genes from the paper used in our study matched with a particular cluster in our dataset as observed in figure 6.

Results

After the differential expression analysis,we extracted the cell type of each cluster and the marker genes which were used to identify the cell type of each cluster.The largest cells according to size are the delta cells(cluster 0),beta cells (cluster 1 and 6) and alpha (cluster 4).We couldn't identify the cell type of cluster 7.As we observe in the heatmap that cluster 0 and 1 looks similar which could mean that these two clusters belong to the same cell type.

Genes such as SP100, PRRG3, and INS seem to be expressed in all cell types. On the contrary, the expression probability distribution seems to be more specific for genes such as

APOE, ACP5, CRP, ALDOB, and GC. Therefore, these genes could be novel marker genes that help identify cell types in the future.

Cluster	Cell Type	Marker Gene
0	Delta	SST
1 & 6	Beta	INS
2	Gamma	GCG
3	Acinar	PRSS1
4	Alpha	TTR
5	Ductal	KRT19
9	Endothelial	SPARC
10 & 12	Macrophage	CRP, LCN2, ACP5
11	Exocrine Glandular	CELA3B, DUUOXA2

Table 2. The marker genes used to label the cell types of each cluster.

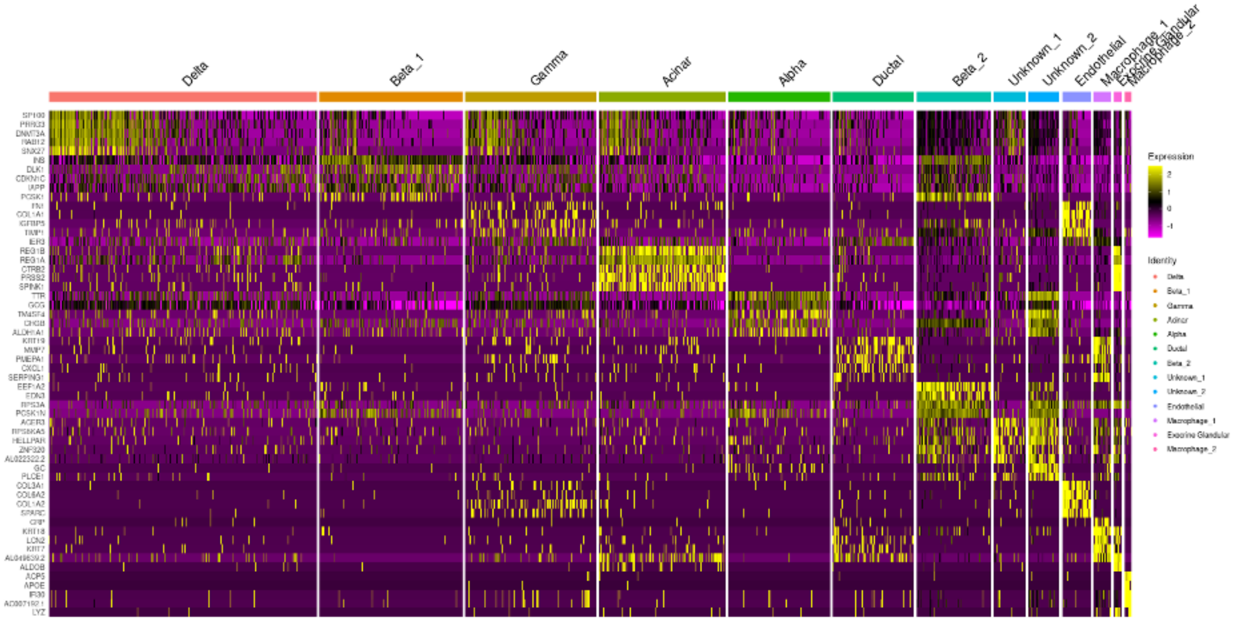


Figure 6. Top 5 marker genes in each cluster obtained by average log2 fold change value.

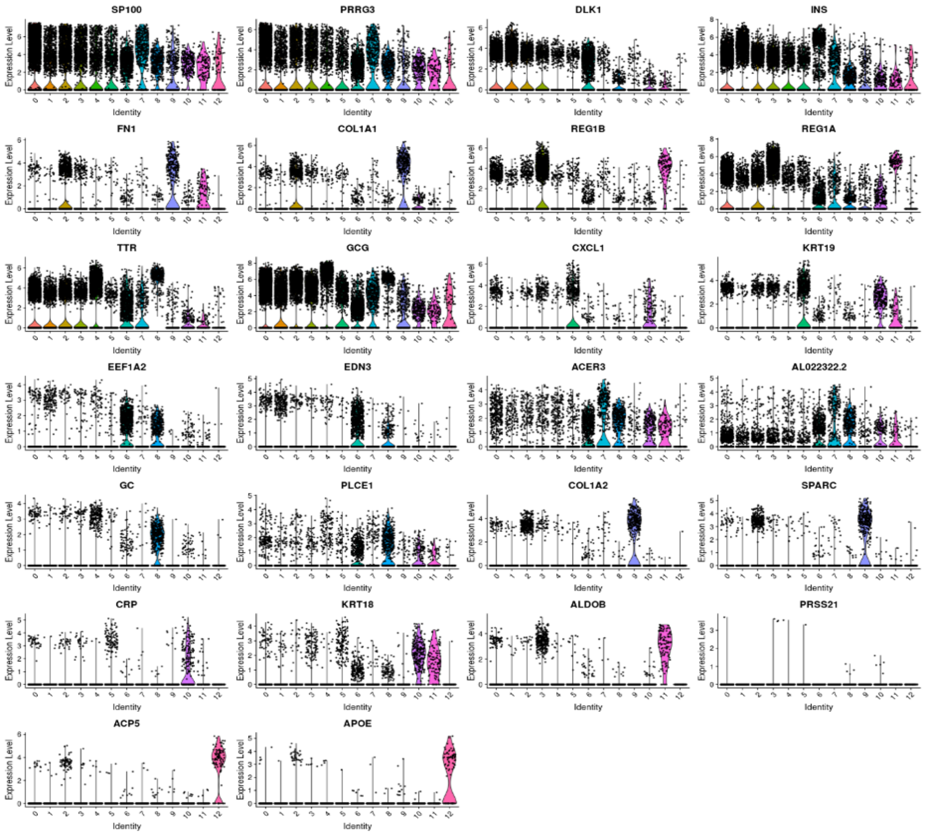


Figure 7. represents the violin plot of the top 2 highest average log2FC marker genes

Discussion-

The marker genes from our cluster marker genes to those reported in the paper, such as GCG, INS, KRT19 from cell types such as alpha, beta, gamma, and delta seem to overlap . However, marker genes such as VWF, RGS5 etc that represent cell types such as stellate are absent in our marker genes.. Cell types such as alpha, beta, gamma, delta, acinar, ductal, and endothelial cell types were identified, but I did not find the epsilon, vascular, cytotoxic T cell types. The gamma and delta cell types ought to have the highest expression level compared to other cell types, whereas our heatmap showed the acinar cell type to have the highest expression level which was a huge difference as compared to the paper. I also identified only a few of novel marker genes as most genes show active expression in many different cell types.

My UMAP looks quite different from Baron et al.'s original tSNE plot, where they were able to identify fourteen distinct cell types using the novel genes. I was able to identify only eight of these fourteen cell types.The original goal of this study was to apply more advanced single-cell analysis tools to the data obtained by Baron et al. in order to replicate the results of their study.I was able to differentiate between eight different cell types using the original marker genes provided, as well as identify additional novel genes to be evaluated for their efficacy. Some of the cell types, such as the pancreatic alpha and beta cells, appear to have clear, distinct "bands" in the heatmap along the x-axis, indicating the enrichment of their respective marker genes.The merging of the PPY and SST genes, indicative of the pancreatic delta and gamma cells, within the pancreatic delta cell cluster. I was able to reproduce similar results compared to Baron et al with some slight variations due to the possibility that we chose different parameters and methods to find gene markers.