Single Cell RNA-Seq of Pancreatic Cells

Group: Dachsund
Vamshi Mallepalli (Data Curator) ⧫ Shreen Katyan (Programmer)
⧫ Katherine Tu (Analyst) ⧫ Sana Majid (Biologist)

Introduction

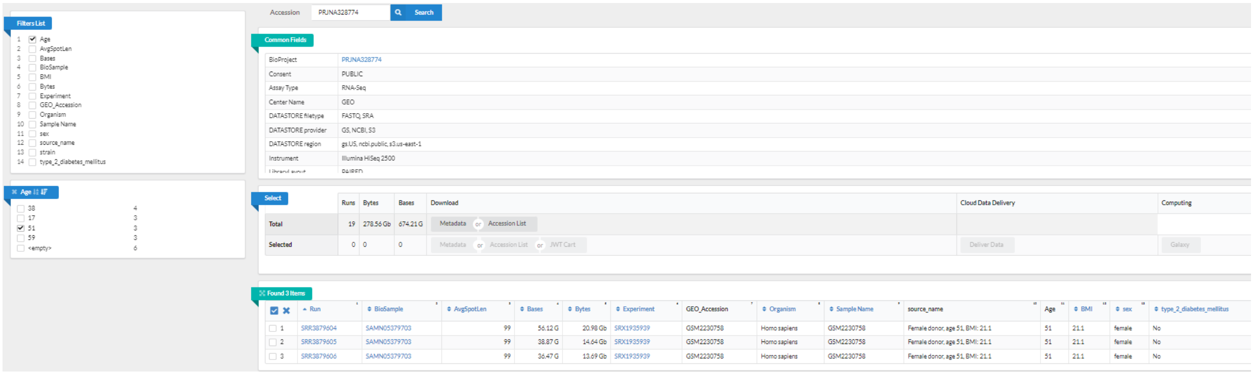
Patterns of gene expression studying tissues in bulk allow researchers to look at gene-regulatory changes in a large number of healthy and diseased samples. The mammalian pancreas is made up of a variety of cell types, each of which supports a plethora of intricate interactions that are critical to the human body's functioning. It becomes possible to elucidate information on the regulation of major dysfunctions that lead to diseases like Type 1 and Type 2 Diabetes Mellitus by obtaining a broader understanding of the expression profiles and transcriptional activity of each of these cell types. Although there have been tremendous advancements in describing the transcriptomes of individual cells using in-situ and RNA sequencing (RNA-seq) methodologies, getting tissues from donors and designing a technology that captures a sufficient number of transcripts remain challenges. Baron et al.[1] 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. They were able to discover population groups of cells with different expression profiles in addition to matching previously recognized cell types. As a result, Baron et al. were able to create a dataset that may be used to uncover and analyze novel, cell-type specific differentially expressed genes analysis.

By evaluating distinct cell types of sequencing data from the pancreas of a 51-year-old female donor, we hope to replicate Baron et al's findings. The most relevant barcodes and UMIs were extracted from the parsed data for barcode[2] extraction. The levels of gene expression in each cell were computed, and the data was grouped using the K-nearest neighbor (KNN) cluster graph approach. The function of linked genes and cell types was deduced using marker genes and cell types.

Keywords: k-nearest neighbor, Gene expression, droplet based single cell transcriptome, barcode UMI.

Data

Only SRR files from a 51-year-old female human donor were processed and analyzed for this project. The metadata from the GEO accession number GSE84133, which contains thirteen samples (sequencing libraries) and four human subjects, was used to determine the SRA accession number. The donor's short read archive (SRA) accession number was SRP07832, which comprised the runs SRR3879604, SRR3879605, and SRR3879606 that were used for further investigation, according to the sample information.



Accession: PRJNA328774

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
Library Layout	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

Figure.1 SRR files from a 51-year-old female human donor

Due to noise in the protocol, raw read 1 barcodes could not be matched to the InDrops barcode scheme, so barcodes were pre-processed and padded to be 19 bases with 6 UMI bases[5]F when a valid barcode pattern was discovered in the read. In downstream analysis, the donor's pre-processed read 1 data, as well as read 2, were further processed and examined.

Methodology

The frequency of readings per distinct barcode was determined to exclude reads with rare barcodes from consideration in order to identify the number of cells actually sequenced and whitelist significant barcodes. Commands like awk and sed were used to extract barcodes from each run, then sort and uniq -c were used to compute the counts of each single individual barcode and initally plotted the results of the each sample. 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 in R[4]. To proceed with our furher analysis. We have downloaded the human reference transcriptome from the Gencode webiste.

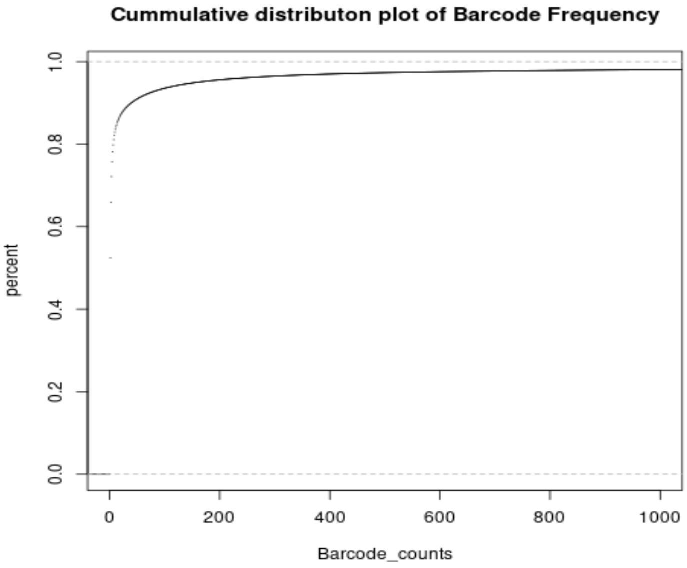


Figure2.a Cumulative distribution plot for the sample SRR3879604

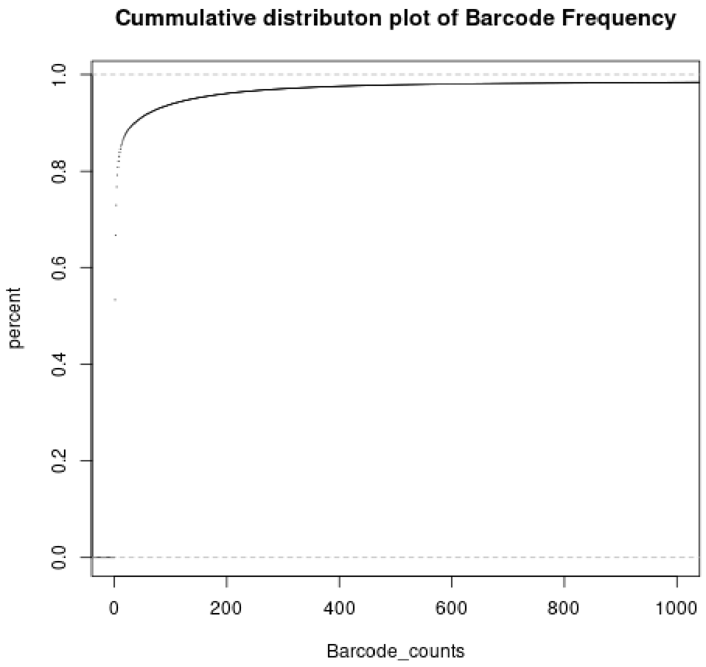


Figure2.b Cumulative distribution plot for the sample SRR3879605

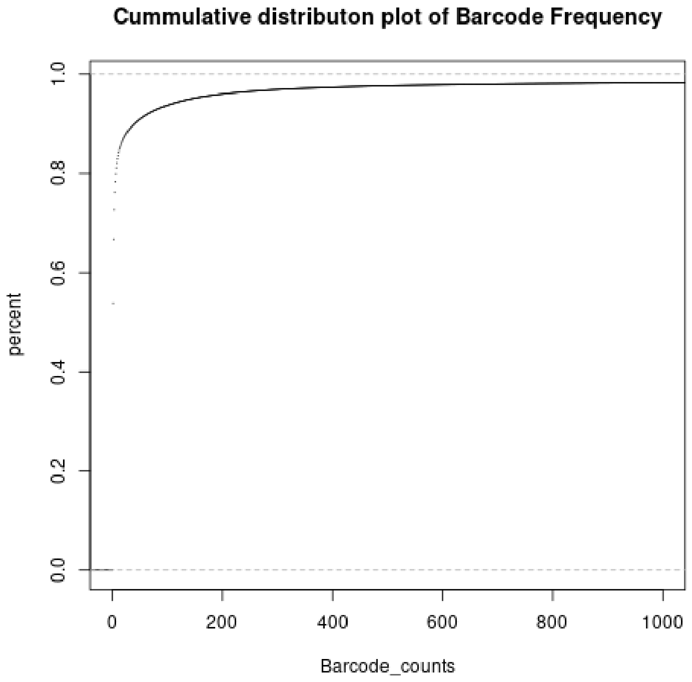


Figure2.c Cumulative distribution plot for the sample SRR3879606

Along with the reference transcriptome (gencode.v40.transcripts.fa.gz) we have downloaded the annotation file(gencode.v40.annotation.gtf) and the genome file(GRch38.p13.genome.fa.gz). A cell-by-gene (UMI) count matrix was built using salmon alevin[6] to quantify the readings. The parameters —end 5 —barcodeLength 19 —umiLength 6 based on the custom barcode and UMI lengths were entered into salmon alevin with the parameters —end 5 —barcodeLength 19 —umiLength 6 based on the whitelist.txt of one distinct barcode per line that passed the filters and the fastq files. The —tgMap option, which takes in a transcript to gene map file of each transcript included in the reference to the corresponding gene, was given to enable salmon to collapse from the transcript to the gene level. This file was developed by mapping the transcript ID (ENSTXX) to gene in the current human reference (GRCh38.p13) transcript sequences (ENSGXXX). Additionally, the transcript sequences were utilized to create a reference transcriptome index using salmon index, which was entered using the -i parameter. A UMI matrix of the number of reads in a cell originating from the corresponding gene, as well as a read mapping summary statistics, were outputted and analyzed in downstream analysis after running the command as a job on the Shared Computing Cluster(SCC) with the library type ISR (inward, stranded, reverse strand).

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

Filtering out Low quality reads

We load the salmon alevin counts file i.e. the UMI Matrix into R using the tximport program. The file was moved to our group directory from the Salmon Alevin folder. Then we created a Seurat object using the count matrix which consists of 15147 samples across 1 assay. Seurat is an R package designed for QC and analysis of single-cell RNA-seq data which we implemented in our study. Seurat helps to understand sources of heterogeneity from single-cell transcriptomic measurements and to integrate diverse types of single-cell data. Here it is implemented to identify the subpopulation cell types of the pancreatic cells.

We implemented a QC workflow where we considered the criteria such as the number of unique genes detected in each cell and the number of molecules detected within a cell. We calculated the percentage of reads that map to the mitochondrial genome with the usage of the PercentageFeatureSet function. Mitochondrial gene attributes to the low-quality or dying cells. This is how we filter out the low-quality cells from our count matrix. We segregate the cells that have unique feature counts over 2,500 or less than 200 and cells that have >5% mitochondrial counts. We are getting the percent mt plot as empty because there are multiple null and missing values in the dataset.

We use the FeatureScatter function to create a violin plot for 2 comparisons. The first plot shows the number of RNA cells and the percentage of mitochondrial. The second plot is the RNA cell count Vs the number of unique genes in the cell which shows a very high correlation of 0.98. Again the correlation plot between the mitochondrial percent and the RNA count shows the NA stats because of the null values in the dataset as mentioned previously.

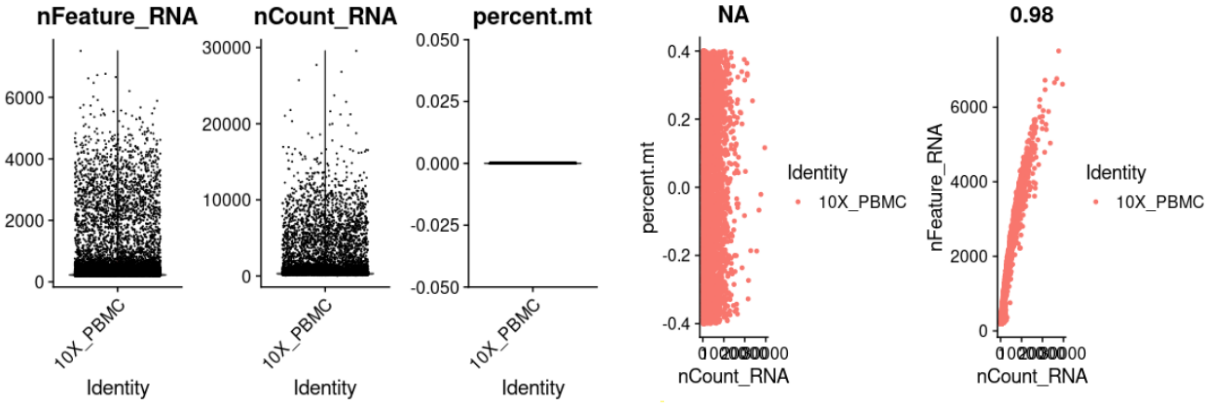


Figure3.a&b. Violin plots for the number of genes detected which are unique in each cell, the total number of molecules detected within each cell and the number of mitochondrial genes per cell.4th figure represents the number of unique genes Vs the MT percentage of the cell and the last figure portrays the feature relationship between the number of genes and the molecules in each cell.

<u>Filtering out low variance genes</u>

We perform further analysis to include only highly variable features which are useful to us. Thus we normalize the data using the "LogNormalize" function that normalizes the feature expression measurements for each cell by the total expression, multiplies this by a scale factor of 10,000, and log-transforms the result. Before carrying out the PCA Analysis, we calculate a subset of features that exhibit high cell-to-cell variation in the dataset. Using the FindVariableFeatures function, we obtain 2,000 features per dataset.

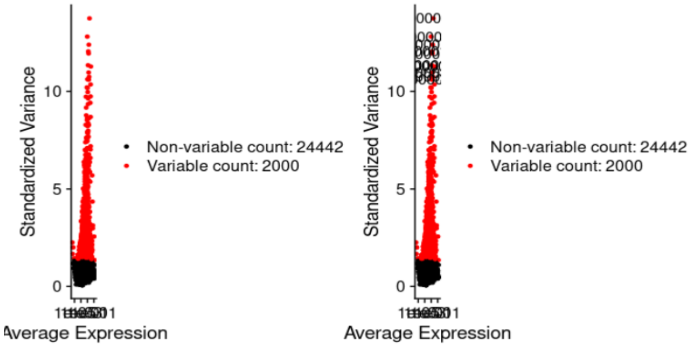


Figure 4: Plot variable features with or without the labels. Scatter plot shows 24442 features with a total of 2000 variable features which are highlighted in red.

Identifing clusters of cell type subpopulations

Now we will scale the data so that the mean expression across cells is 0 and the variance is 1. Then we performed the linear dimensionality reduction via the PCA technique and we obtain the first 5 Principal Components with their unique genes in each cell. Visualization using Seurat can be done using VimDimLoading in our study for the first 2 dimensions. The Jackstraw plot depicted that there's no significant drop off in significance within the 1st 15 PCs visualized. JackStraw procedure was primarily used to handle the noise in any single feature for single-cell RNA-seq data.

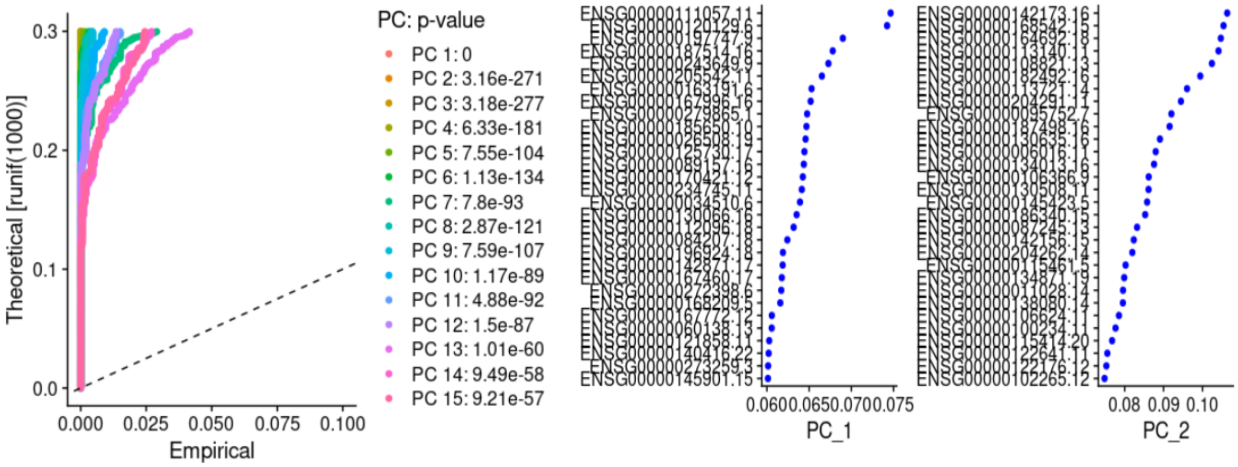


Figure 5 part A depicts the Jackstraw plot which shows 1st 15 Principal Components after PCA Analysis.Part B represents the 1st 2 PCs after PCA Visualization using VimDimReduction.

We have used the KNN graph to cluster the cells in our study. Using the Euclidean distance with 1st 10 PCs as cutoffs, the cellular distance matrix was partitioned into clusters where highly interconnected cell communities were segregated as a cluster and edge dawn between cells with similar feature expression patterns. Modularity optimization was performed using the Louvain algorithm. Then we looked at the clusters of the first 5 cells containing 12 levels obtained. We further performed the UMap technique for nonlinear dimensionality reduction because it is better for visualizing the high dimensional data as compared to the tSNE. We input the same 10 PCs to UMap plot to retrieve 13 distinct clusters of cell type subpopulation. We would be saving the Seurat object in .rds file with our file name as "programmer_output.rds”.

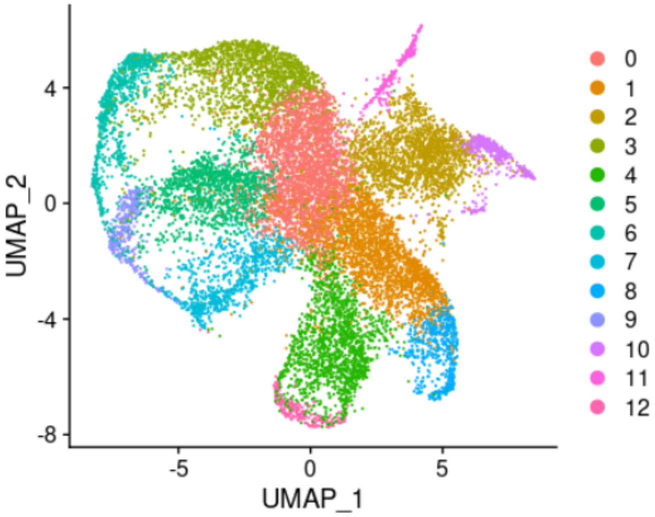


Figure 6: UMap plot for the dataset which represents 13 distinct clusters of cell types subpopulations.

We further calculated the number of clusters and depicted the relative proportions of cell numbers via a barchart.This identified the number of cells in each cluster.

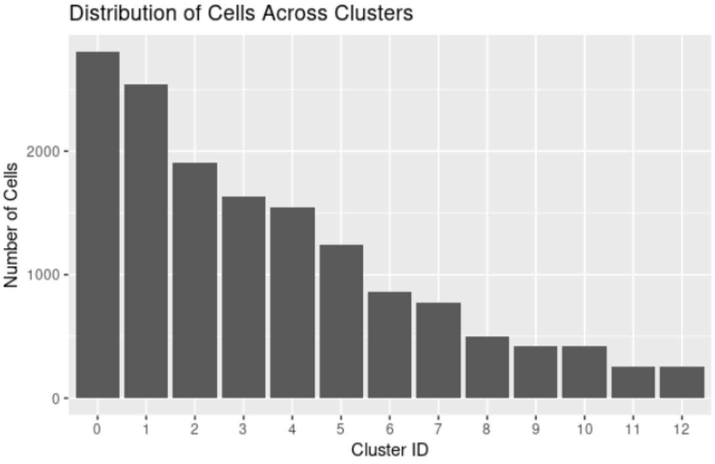


Figure 7: barchart showing the distribution of cells across different clusters

<u>Identifying and labeling marker genes</u>

Differential expression analysis was performed on the genes in the sample dataset to identify biomarkers for each cluster. The function "FindAllMarkers" from Seurat was used, where it identifies positive and negative genes compared to the rest of the genes in the cluster. In this study, only the positive markers were reported. Furthermore, only those genes with a minimum of 25% coverage in either group of cells and those genes that pass a 0.25 log2 fold change threshold were considered. Next, the cell type of each cluster is identified by researching the cell type of the top 3 highest average log2 fold change of each cluster. Research sources include scientific papers and the Human Protein Atlas website. To begin, we looked at the top 1 average log2 fold change and check the cell type with those reported by Baron et al. If the author also used the gene as a marker gene, then the cell type of this cluster would be the same as indicated by the author. If not, we searched on the Human Protein Atlas website to see which cell type this cluster is closest to. In the case when it is unclear for the top 1 average log2 fold change gene, we take the top 3 genes into account. The cell type of each cluster is shown in table 1 and figure 1.

Visualizing clustered genes

After each gene is labeled and categorized into a specific cluster, we visualized the clustered cells using a projection method – UMAP. UMAP is an unsupervised learning method that performs dimension reduction and helps us visualize the gene clusters in a 2D space. The "RenameIdents" function helps us to rename the labels of each cluster. However, because of the many clusters and long names of each cluster, we decided to leave the cluster labels as numeric and identify them further in a table as indicated in table 1. The UMAP was plotted using the "DimPlot" function with the reduction parameter set to "umap" and the label set as true. The resulted UMAP is shown in figure 1.

Visualizing top marker genes per cluster and identifying novel marker genes

The highest 5 marker genes of each cluster were used to create a heat map to visualize a single-cell gene expression. In order to achieve this, we first grouped the top 5 marker genes according to average log2 fold change and applied the function "DoHeatmap" to obtain the heat map. The result is shown in figure 8. In addition, we plotted a violin plot using the function "VlnPlot" to show the expression probability distribution across clusters. In this study, the top 2 highest average log2 fold change genes were plotted as shown in figure 9. Novel marker genes are further identified by evaluating the expression probability distribution across different clusters. If the genes show high expression probability in one specific cluster and was not previously identified as a marker gene for that specific cluster, we consider this gene as a novel marker gene. The novel marker genes of each cluster are presented in table 3.

Results

To further understand the cell type in each cluster, we identified the marker genes in each cluster by differential expression analysis. The cell type of each cluster is shown in table 1 and visualized in figure 1. The marker gene mentioned in table 1 is the genes that we used to identify the cell type of the cluster. The largest cell types in size are the beta cells (cluster 1 and 6), delta (cluster 0), and alpha (cluster 4). There are two clusters in which we were unable to identify their cell type, cluster 7 and 8. From the Human Protein Atlas website, it just indicates that most of the cells that are present in these clusters are hepatocyte related. The resulting UMAP shows that it is not the best way to cluster the genes as there most clusters have a small portion that overlaps with each other. The different expression level of the top five marker genes in each cluster is shown in figure 8. The highest expression level is indicated by yellow, whereas the lowest is indicated by purple. According to the figure, the yellow is obvious in a different section for each cluster, indicating the different gene expression levels for each cluster. It is interesting to see that some clusters have similar gene expression levels as another. For example, the genes in gamma and endothelial clusters or those in unknown_1 and unknown_2. This might be because marker genes do not have to be only expressed in one cell type and could often be expressed in many other cell types. This is further proved to be true when evaluating the expression probability distribution across different clusters as indicated in figure 9. This figure shows the expression probability distribution of the top 2 average log2 fold change genes. 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. The cluster in this table corresponds to the clusters in figure 1. Cluster 7 and 8 are not mentioned in this table because the top 5 marker genes do not seem to belong to the same category.

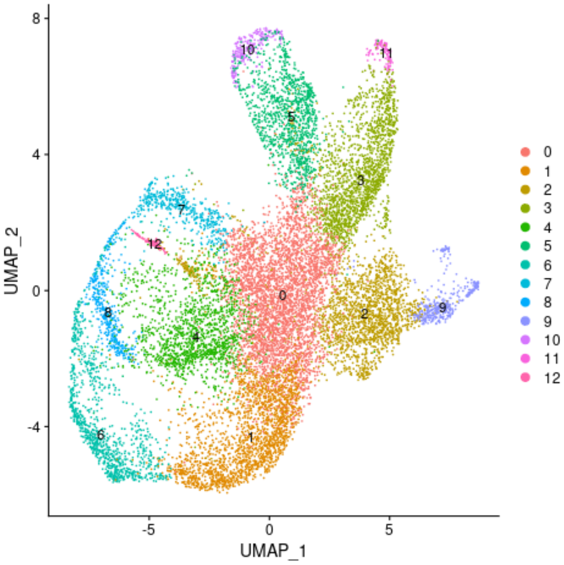


Figure 8. UMAP plot for cell clusters labeled with different colors. The corresponding cluster names can be found in table 1.

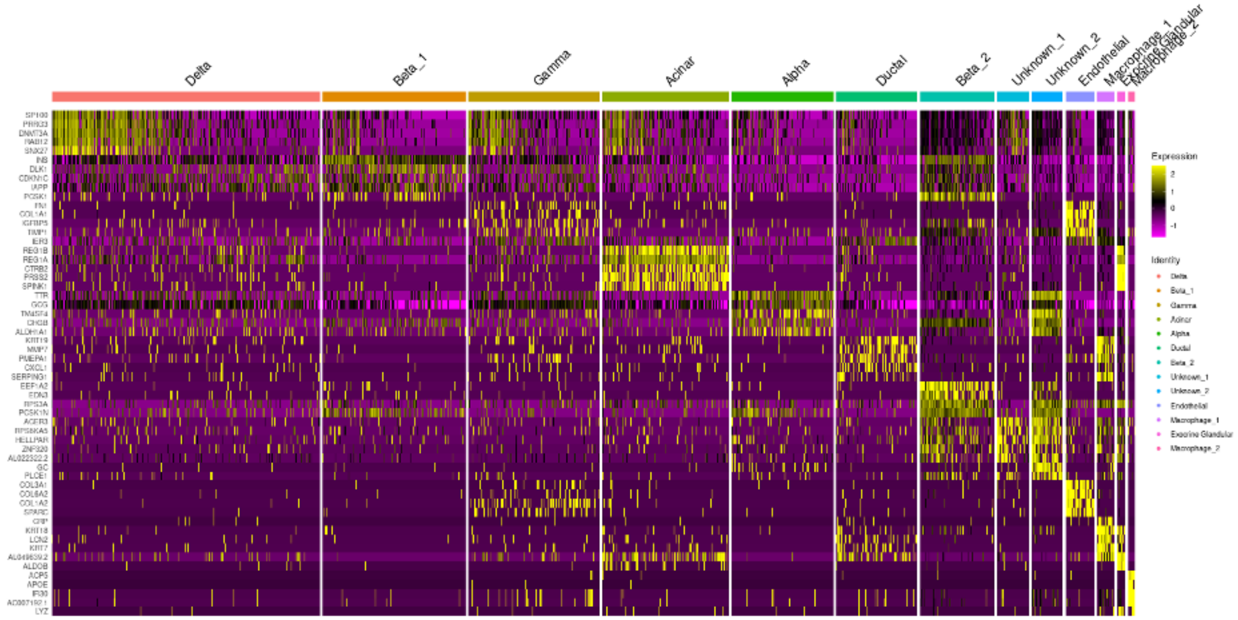


Figure 9. Top 5 marker genes in each cluster. The top 5 is obtained by average log2 fold change value. The expression level of each gene (each row) is indicated by different colors, ranging from yellow to purple representing high to low expression.

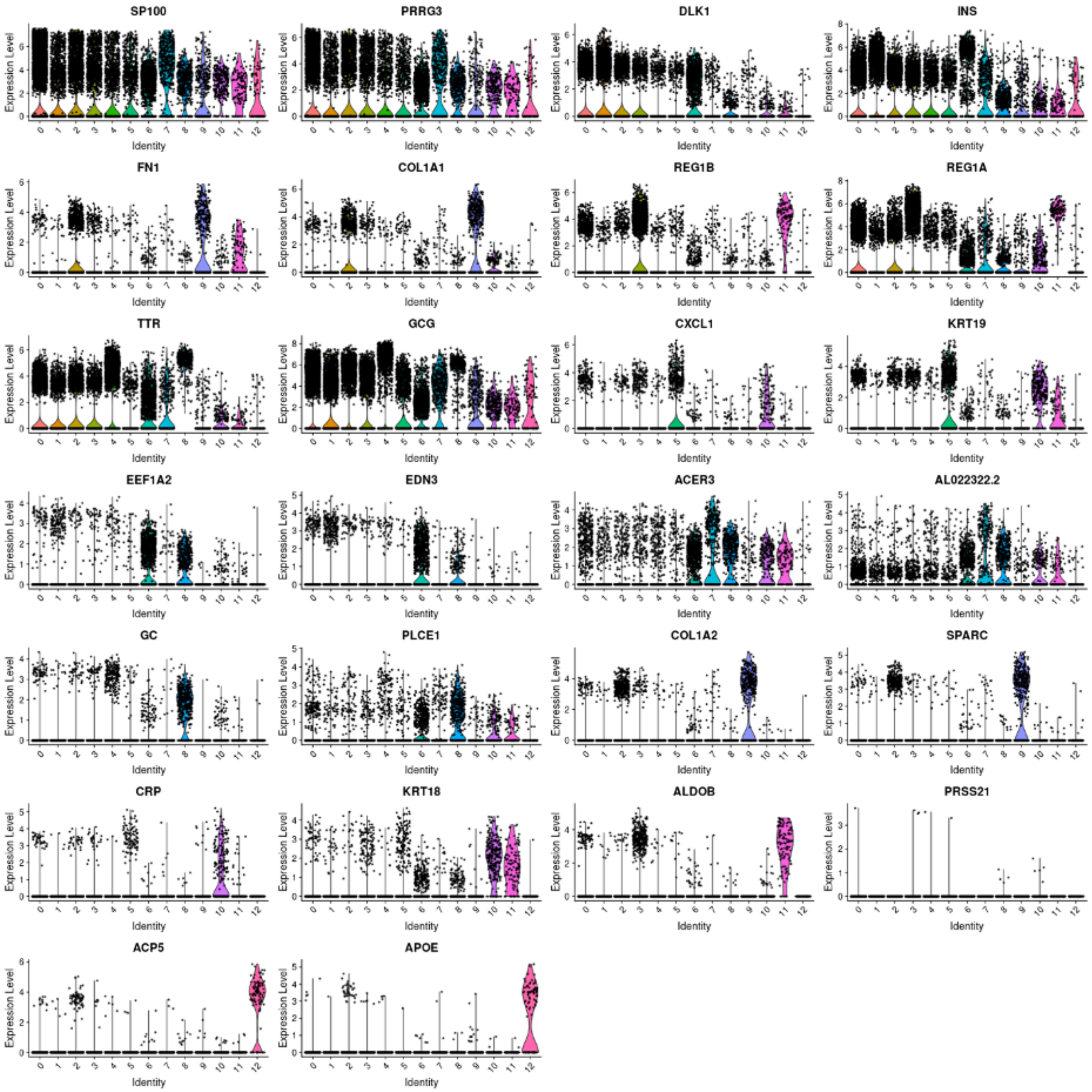
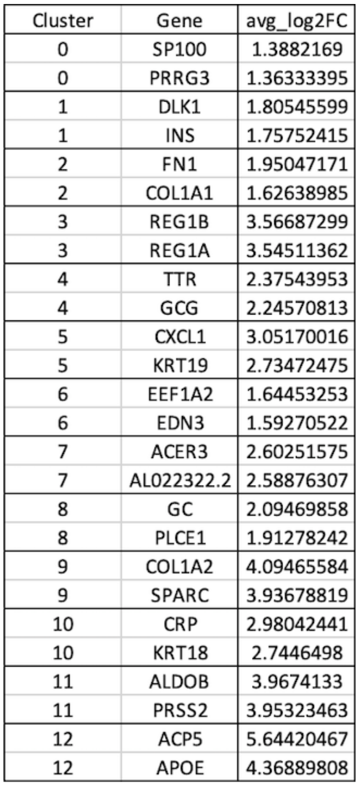


Figure 10. Violin plot of the top 2 highest average log2FC marker genes from each cluster.



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

Table 3. Top 2 highest average log2FC marker genes from each cluster.

To determine cluster identification in an unbiased approach, marker genes determined through differential expression analysis (location: projectnb/bf528/users/dachshund/project_4/Analyst/marker_genes.csv) were used to perform gene set enrichment analysis. For each cluster, the marker genes were sorted based on the adjusted p-value and run through DAVID.

Cluster 0 was heavily enriched for genes and pathways pertaining to aerobic respiration and thermogenesis, as well as neuronal activity, suggesting the endocrine function; as the top marker gene was SST, these are delta cells. Cluster 1 involved genes for insulin processing and secretion, and cluster 6 included genes involving insulin receptor binding and circadian entrainment; both of these clusters are therefore suggestive of beta cells, with cluster 6 perhaps being the less mature cell population. Cluster 2 highlights terms of extracellular matrix enrichment and GCG suggests gamma cells. Cluster 3 shows enrichment with Reg gene family and pathways for digestion, suggesting labeling as acinar cells. Cluster 4 consists of pathways of transport vesicles and secretion; TTR as the marker gene clearly suggests alpha cells. Cluster 5 involves contraction and secretion of various hormone types, with the presence of KRT19 marker gene suggesting ductal cells. Cluster 7 is enriched for oxidative phosphorylation, neurogenesis and contraction; presence of genes such as GPR82 suggest exocrine glandular cell type. Cluster 8 is enriched in a multitude of pathways from thermogenesis and circadian entrainment to GABAergic synapse; although GHRL(ghrelin) is not present among the marker genes, PLCE11 and ARX suggest epsilon cells as the label for this cell population. Cluster 9 involves transport, immune response and receptor internalization, with SPARC and SERPINE suggesting endothelial cell type. Cluster 10 shows enrichment for secretion, immune response and lysosome; LCN2 and KRT18 among other marker genes suggest macrophages, more specifically differentiating from THP-1 monocytes. Cluster 11 involves lipid metabolism, digestion and contraction, and marker genes appropriately suggest exocrine glandular cells. Finally, cluster 12 shows lysosomal activity and immune defenses, with ACP5 and SDS suggesting macrophages.

Table 4. Enriched gene sets for cell clusters

Cluster	No. of marker genes	Cluster label	Enriched gene sets
0	30	Delta cells	respiratory chain, TCA cycle and respiratory ETC, oxidative phosphorylation, thermogenesis, pathways of neurodegeneration (multiple diseases)
1	36	Beta cells	secretory granule, signal transport vesicle, peptide hormone metabolism, insulin processing/secretion, ribosome, gastrulation
2	22	Gamma cells	extracellular matrix, cell adhesion, fibronectin, ECM-receptor interaction
3	66	Acinar cells	cytoplasmic translation, ribosome, rRNA processing, digestion, metabolism of amino acids, pancreatic secretion
4	30	Alpha cells	cytoplasmic vesicle, secretory granule membrane, endosome, integral component of membrane
5	68	Ductal cells	extracellular space, secreted, S100 protein binding, endosome, cardiac muscle contraction, insulin secretion, salivary secretion, gastric acid secretion
6	1070	Beta cells (possibly a less mature population)	cytosolic ribosome, oxidative phosphorylation, thermogenesis, insulin receptor binding, lipid/FA metabolism, circadian entrainment
7	147	Exocrine glandular	respiratory chain, oxidative phosphorylation, thermogenesis, ATP/GTP binding, neurogenesis, contraction

8	1474	Epsilon cells	translation, thermogenesis, lipid metabolism, extracellular matrix organization, cytoskeleton organization, circadian entrainment
9	262	Endothelial cells	Ribosome, translation, GTPase activity, transport, adaptive immune response, receptor internalization
10	1932	Macrophages	ribosome, viral receptor activity, steroid biosynthesis, proteasome, lysosome, protein transport
11	1117	Exocrine glandular	lipid metabolism, FA metabolism, fructose and mannose metabolism, adaptive immune response, contraction
12	102	Macrophages	Lysosome, oxidative phosphorylation, immune response, endocytic recycling

Discussion

When comparing our cluster marker genes to those reported in the paper by Baron et al, it seems like most marker genes, such as GCG, INS, KRT19, and more, from the more commonly seen cell types overlap. Those cell types are alpha, beta, gamma, and delta. However, marker genes such as VWF, RGS5, and others that represent cell types such as activated stellate and quiescent stellate are missing from our marker genes and therefore not detected. We successfully identified the alpha, beta, gamma, delta, acinar, ductal, and endothelial cell types, but we did not find the epsilon, vascular, cytotoxic T, activated stellate, and quiescent stellate cell types present. In addition, we further identified two cell types, macrophages and exocrine glandular, that were not mentioned in Baron et al. However, there are also two clusters, clusters 7 and 8, that we were unable to identify initially. When projecting these clusters to a map as in figure 1, our plot is slightly different from figure 1D from Baron et al. The beta-cell type remains to be the largest cluster in both plots. However, the genes in figure 1D from Baron et al seem to be more separated and distanced from one another compared to our figure 1 in which there are many overlaps between clusters. This might be because the authors used a different clustering method to cluster the genes. Another possible reason for this result is because the authors used a t-sne plot and we used a UMAP to present the clusters, so it is reasonable to see such a difference when comparing the two plots. When comparing the heatmaps from the paper to ours, it is surprising to see the large difference in the size of the delta cell type in our data compared to the size from Baron et al. After comparing the heatmap from the paper to ours, we became interested as to how the authors clustered their genes. The heatmap shown in figure 1B from Baron et al was very clean and every marker gene seems to be specific to one cell type, whereas our heatmap appears to be a bit messy. Although we can clearly see the different expression levels of genes in different cell types, there are some overlaps. Another thing we noticed was that the gamma and delta cell types seem 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. As for novel marker genes, we only identified a few as most genes show active expression in many different cell types, as shown in figure 3. It would be interesting to see the novel marker genes that we identified being used as an identifier in future research.

Upon exploration of enriched gene sets using the identified marker genes with differential expression analysis, we gained further insight into the possible identities of cell clusters: with circadian entrainment showing for cluster 6 marker genes, we expect this to be a more immature cell population of beta cells; clusters 7 and 11 both appear to be exocrine glandular cells; and cluster 8, with the presence of PLCE11 and ARX, suggests epsilon cells.

To sum up, we were 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. Because of the different gene markers present, we end up using different marker genes to label each cluster.

References

1. Baron, Maayan, Adrian Veres, Samuel L. Wolock, Aubrey L. Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, et al. 2016. “A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-Cell Population Structure.” Cell Systems 3 (4): 346–60.e4. PMID: 27667365

2. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161:1187–1201

3. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015 May 21;161(5):1202-1214. doi: 10.1016/j.cell.2015.05.002. PMID: 26000488; PMCID: PMC4481139

4. The R Project for Statistical Computing.” R, www.R-project.org/

5. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. 2017 Mar;27(3):491-499. doi: 10.1101/gr.209601.116. Epub 2017 Jan 18. PMID: 28100584; PMCID: PMC5340976.

6. 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