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

**Group Members:**

Jiaming Zhang (Data Curator)

Mikhail Kouzminov (Programmer)

Jinghan Huang (Analyst)

**TA:**

Kritika Karri

**Introduction**

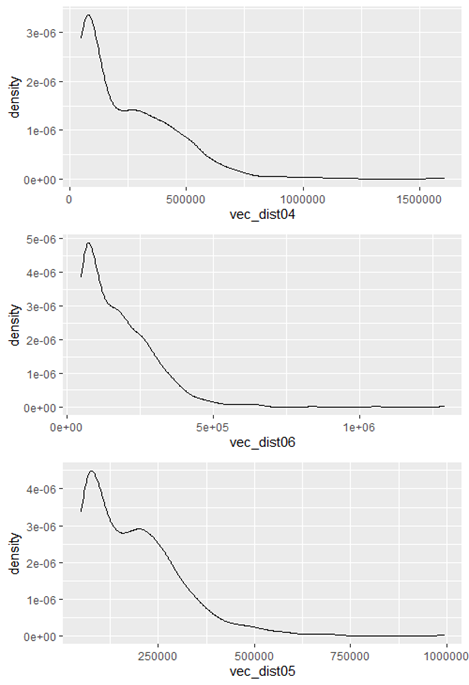
In this project, we followed the method from Baron et al. and analyzed the gene expression profile from single cell RNA seq from a human pancreatic islet sample to identify markers genes that differentially expression between different cell types.

**Data**

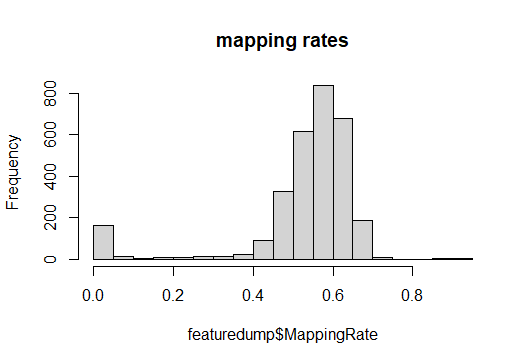
The cells used in this analysis came from human pancreatic islets from a 51 female donor. The cells were encapsulated using the inDrop platform. The sample was lysed as droplets with 0.4% NP-40. The single cell lysates were then reverse transcribed and barcoded with the inDrop(Klein et al.). The single cell seq data contained results from 3 paired end Illumina runs.

**Data Description**

The three fastq files resulting from the illumina runs have been processed so that read 1 has the format of 19 bases barcode followed by 6 bases UMI. The unique reads per barcode has been quantified for each read 1 fastq files. The frequency of the number of unique reads has also been quantified and plotted in R. The paper reported 100000 read per barcode and around 800 cells per batch. To match the paper’s description. We first filtered out barcodes that have a unique read of less than 50000. The density plot (Fig 1) showed the majority of the barcodes had unique reads in the range from 50000 to 250000. However setting a upper unique reads threshold didn’t significantly reduce the final whitelisted barcode. To be on the safe side, all the barcodes with greater than 50000 reads were included in the whitelist. The total unique barcodes whitelisted was 3216 for all 3 runs. The whitelisted barcodes are on the higher side, so we expect some mitochondrial RNA being identified after quantification. Salmon Alevin package was used to quantify reads against a transcript index (from GRCh38.p13) and a transcript to gene map file, using the combined whitelist. To better visualize the UMI count matrix output by Salmon Alevin, we converted the sparse matrix to conventional readable csv format via a python script. Examine the feature dump file from Alevin output, the mapping rate has a mean of 0.527 and a standard deviation of 0.148 among all the barcodes. The histogram was plotted for the mapping rate (Fig.2).



**Fig. 1 Distribution density plot for unique reads frequency among barcodes**

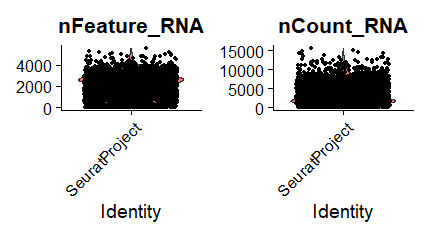


**Fig. 2 Histogram of mapping rate across barcodes**

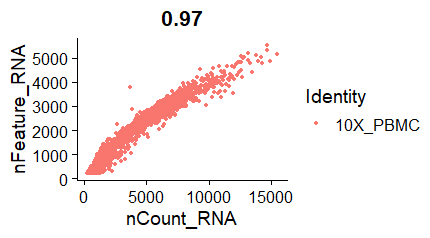
**Data quality control**

We used the R package tx import to import the results from the UMI counts Matrix into Seurat, using biomaRt to convert the Ensembl gene IDs into gene symbols where viable. A small amount of gene symbols were returned blank, so we used the original Ensembl IDs, believing that they were potentially erroneous in the data and would likely be filtered out, and if the ids were needed later, they could be replaced.

Using Seurat/3.1.5 we filtered out low quality cells from the UMI counts Matrix based on having at least 200 genes detected within each cell and each gene detected in at least 3 cells as our metric of quality. In order to ensure quality, we created a violin plot of the data [Fig.3] and a scatter plot comparing number of genes per cell vs number of molecules. [Fig.4] We used this to scatter plot filter out a cell that had unusually high molecule count, nearly equal to the gene count, assuming there was something strange about sampling for that one in order to filter out results from cells that had issues during sampling.



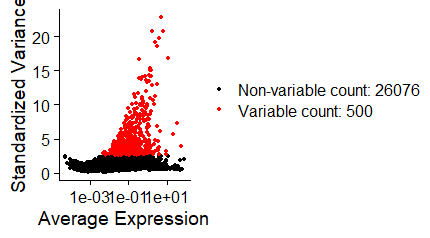
**Fig.3: Violin Plot of number of molecules(left) vs number of features (right)**



**Fig.4 Scatter plot showing number of molecules (vertically) vs number of genes (horizontally) for each cell. We excluded the point on the top left with about 4000 of both molecules and genes**

While the UMI Counts Matrix had 2987 cells and 60233 genes sampled, after performing quality control on these cells, our data had 2797 cells and 26576 genes remaining.

After scaling our data by 10,000 and log normalizing it, we then filtered out genes that had low variance, plotting mean expression values vs variance [Fig.5] This resulted ultimately in a list of 500 features, a number chosen because average expression is noticeably higher in general for those genes, while variance also starts increasing dramatically from that point onwards. We decided that these would be the genes most worth an analysis.



**Fig.5 Plot of Average Expression vs Variance of genes. Note that average expression is noticeably higher for genes using this metric**

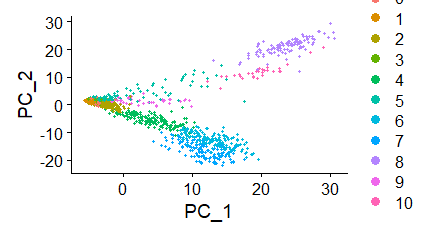
The top 10 genes with the most variance in expression swere:

REG3A, REG1B, REG1A, COL1A1, PRSS2, CTRB2, PPY, PLVAP, IGFBP5, COL3A1

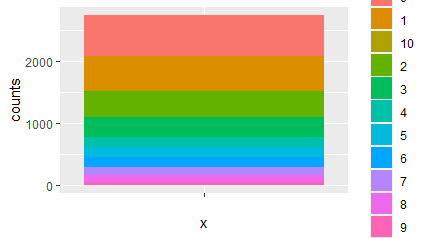
**Methods**

**Identifying Clusters of genes using Principal Component Analysis**

We ultimately used the top 500 genes to perform clustering via Principal Component Analysis with 11 clusters in order to identify clusters of cell type subpopulations, as plotted in Fig.6. Fig.7 shows a bar chart of counts for each group.



**Fig.6 Plot of Principal Component Analysis**

****

**Fig.7: Bar chart of counts of genes in each group**

Since not all groups are clear in the table, they each contained the following amount of elements:

|  |  |
| --- | --- |
| Group | Count |
| 0 | 666 |
| 1 | 540 |
| 2 | 424 |
| 3 | 320 |
| 4 | 162 |
| 5 | 158 |
| 6 | 156 |
| 7 | 138 |
| 8 | 112 |
| 9 | 42 |
| 10 | 29 |

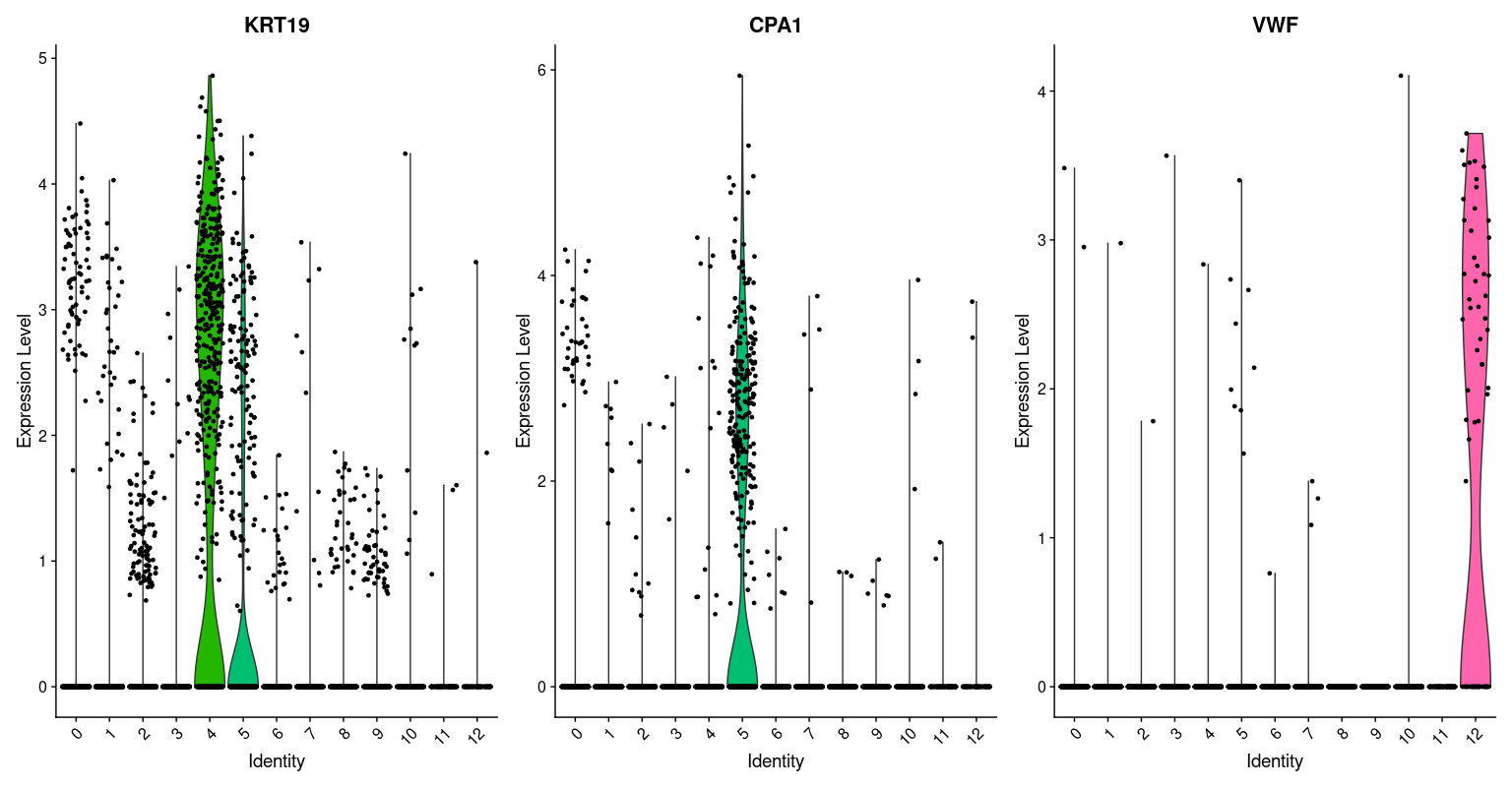
**Obtain marker genes for each cluster using differentially expression method**

The differentially expressed analysis was performed to obtain the marker genes for each cluster using Seurat/3.1.5 [2]. The function FindAllMarkers in Seurat was applied. The parameter ‘min.pct’ was set to 0.25 which indicated that only genes that are detected in a minimum fraction of 0.25 cells in either of the clusters would be tested. The test used was the Wilcoxon Rank Sum test, which was a Non-Parametric test based on ranking all values from the two clusters from smallest to largest and did not require normal distribution of the data. The other parameters were set to default. The markers for each cluster were then used for identification of cell types.

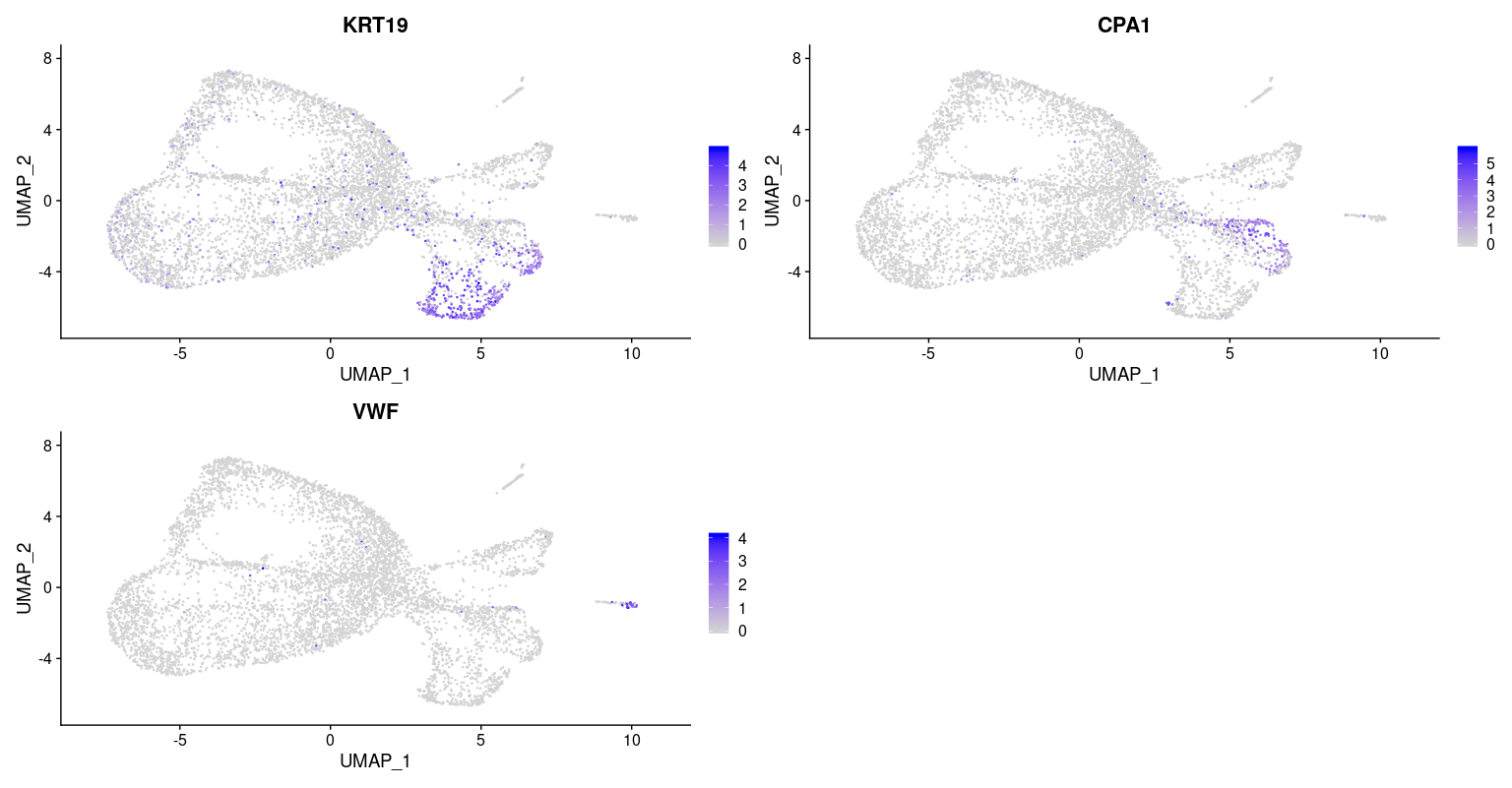
**Label clusters to cell types**

The clusters were labeled to different cell types based on the marker genes provided by the original paper (supplementary information and Fig.6) as well as a database ‘PanglaoDB’ [3] which provided information about gene marker and corresponding cell types. This database allowed users to type in one gene and return the gene enrichment in different cell types as a bar chart. Only positive marker genes were used.

Additionally, two visualization methods were used to help identify the cell clusters. The Vlnplot function in Seurat provided the expression level of the genes in each cluster as Violin plot, while the FeaturePlot function provided the UMAP plot showing the gene expression level in cell clusters. One example was shown in Fig.6 and Fig.7.



**Fig.8 Violin Plot for expression level of three marker genes (KRT15, CPA1, VWF) of Ductal, Acinar and Endothelial cells for 13 cell clusters.**

****

**Fig.9 UMAP Plot showing expression level of three marker genes (KRT15, CPA1, VWF) of Ductal, Acinar and Endothelial cells for 13 cell clusters.**

From Fig.8 and Fig.9, the marker genes for Ductal, Acinar and Endothelial cells provided by the original paper were placed in the VlnPlot and FeaturePlot function. Cluster 4, 5, 12 showed relatively high expression levels of these three genes respectively, and therefore they would be considered as Ductal, Acinar and Endothelial cells. The top 10 marker genes for these three clusters were then used to verify their corresponding cell type using ‘PanglaoDB’. For some of the unknown cell clusters (unable to identify using marker genes in original paper), the top 10 marker genes were placed in ‘PanglaoDB’ and a possible cell type was summarized.

**Visualize the clustered cells using UMAP**

The cell clusters were visualized using the UMAP dimension reduction method. It was based on manifold learning techniques and ideas from topological data analysis and provided a very general framework for approaching manifold learning and dimension reduction as well as specific concrete realizations [4]. The ‘RunUMAP’ function in Seurat was used to obtain the UMAP components by setting the parameter ‘dims’ to 10 (refer to the first 10 PCA components calculated previously). Then the DimPlot function was used to make the 2D plot by setting the parameter ‘reduction’ to ‘umap’ using the first 2 UMAP components. The cell clusters were automatically set to different colors and labeled as the assigned cell types.

**Visualize the top 5 marker genes per cluster**

The top 5 marker genes for each cluster were selected out from differentially expressed analysis and then visualized using heatmap. The DoHeatmap function with parameter ‘feature ’setting to top 5 marker genes in Seurat was applied.

**Find novel marker genes**

The novel marker genes for each cluster were defined as the top genes that were not used to identify the cell type (not the marker genes of the cell type provided by the original paper) and were significantly, solely and positively expressed in that cell cluster. The top 10 significantly positively expressed genes (except the genes that used to identify the cell type) for each cluster were selected out and analyzed in ‘PanglaoDB’ and other research studies. Only the solely expressed genes that found to be relevant to the cell type based on research studies or showed acceptable enrichment in the corresponding cell type in ‘PanglaoDB’ were defined as novel marker genes.

**Results**

Due to a relatively small number of cells in our dataset which resulted in some missing marker genes from the original paper (the genes were possibly filtered out because few cells expressed), we recreated the dataset in this part by increasing the number of cells using a larger whitelist and a less strict threshold for filtering.

**Cluster marker genes comparison**

|  |  |
| --- | --- |
| **Cell type** | **Marker Gene** |
| Alpha | GCG |
| Beta | INS |
| Delta | SST |
| Gamma | PPY |
| Epsilon | GHRL |
| Ductal | KRT19 |
| Acinar | CPA1 |
| Activated stellate | PDGFRA, PDGFRB |
| Quiescent stellate | RGS5, PDGFRB |
| Endothelial | VWF, PECAM1, CD34 |
| Macrophage | SDS, CD163, CD68, IgG |
| Cytotoxic T | CD3, CD8, TRAC |
| Mast | TPSAB1, KIT, CPA3 |

**Table.1 Marker genes for each cell types provided in the original paper**

|  |  |
| --- | --- |
| **Cell cluster** | **Significant Marker Genes** |
| 0 | SST, INS, IAPP, PPY, RPS23P8 |
| 1 | TTR, GCG, TM4SF4, CFC1B, CRYBA2 |
| 2 | TM4SF4, AC092155.1, AC091078.1, VGF, LOXL4 |
| 3 | INS, IAPP, HADH, DLK1, MAFA |
| 4 | KRT19, TACSTD2, MMP7, PMEPA1, SPP1 |
| 5 | CTRB2, PRSS2, REG1B, CTRB1, REG1A |
| 6 | MAFA, PCSK1, C1QL1, DLK1, ABCC8 |
| 7 | COL1A1, COL3A1, COL1A2, COL6A2, BGN |
| 8 | AQP3, PEG10, RBP4, RGS2, PPY |
| 9 | CACNA1A, RPS6KA5, AL022322.2, CNKSR3, HELLPAR |
| 10 | RPS6KA5, CHST12, IFT22, AL022322.2, ACER3 |
| 11 | LAPTM5, HLA-DRA, ACP5, IFI30, CD74 |
| 12 | PLVAP, ESM1, FLT1, ENG, TP53I11 |

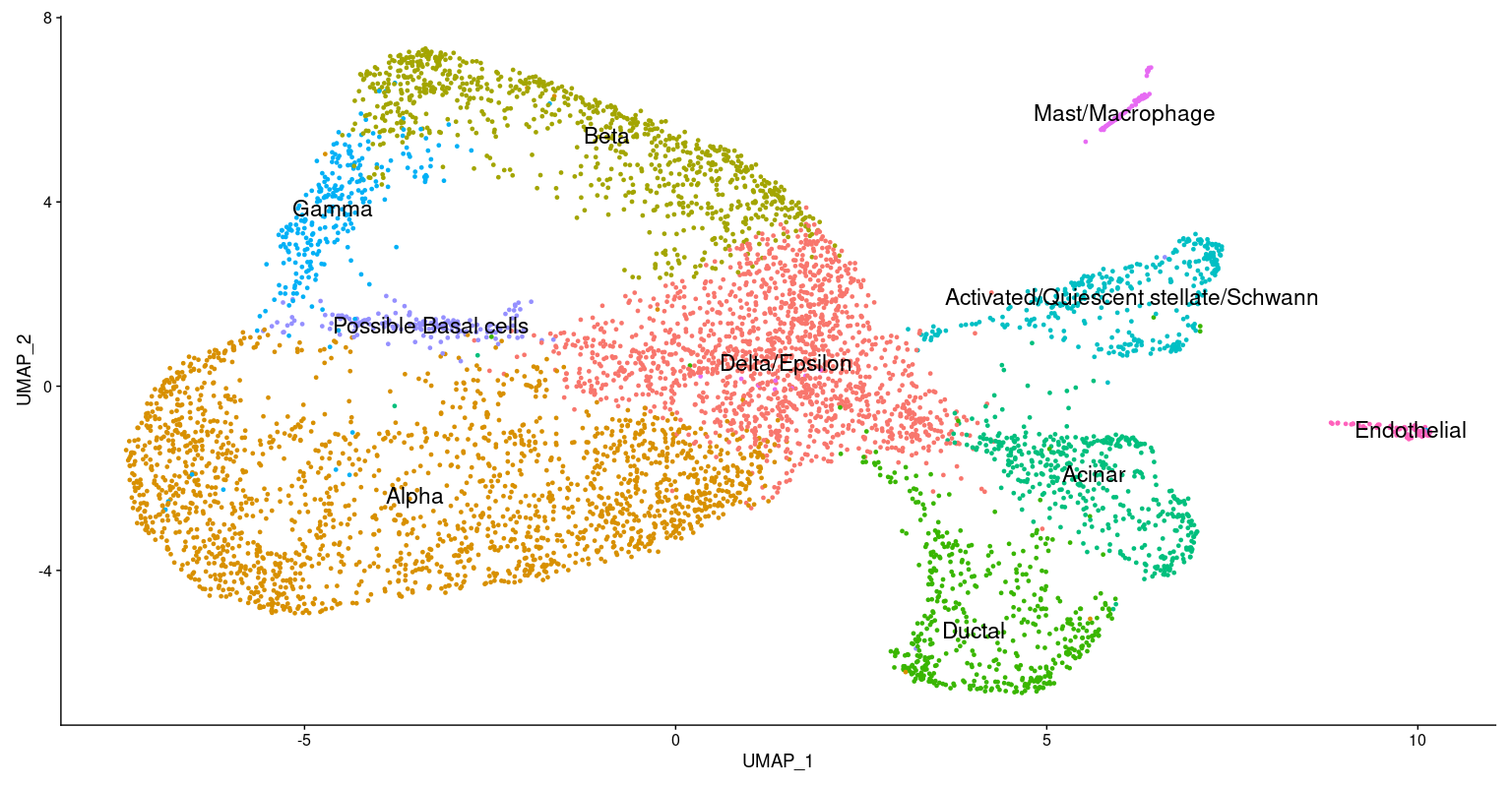
**Table.2 Top5 significant marker genes for each cell clusters using differentially expressed analysis**

|  |  |
| --- | --- |
| **Cell cluster** | **Assigned cell type** |
| 0 | Delta / Epsilon |
| 1 | Alpha |
| 2 | Alpha |
| 3 | Beta |
| 4 | Ductal |
| 5 | Acinar |
| 6 | Beta |
| 7 | Activated / Quiescent stellate |
| 8 | Gamma |
| 9 | Alpha |
| 10 | Basal cells |
| 11 | Macrophage / Mast |
| 12 | Endothelial |

**Table.3 Assignment of the cell types to the cell clusters**

The marker genes provided in the original paper was shown in Table.1. The top5 significant marker genes for each cluster was shown in Table.2 and the assigned cell type for each cell cluster was shown in Table.3. When considering only the top 5 differentially expressed genes for each cluster, only cluster 0, 1, 3, 4 included the marker genes presented in the original paper, and therefore they were assigned to Delta, Alpha, Beta and Ductal cells respectively. For the rest of the clusters, a visualization method described in Method part was applied, which did not depend on the top 5 differentially expressed genes. The cell cluster was assigned to a cell type if the marker genes of that cell type from original paper showed relatively high expression level solely in that cell cluster, regardless whether it was in the top 5 significantly expressed genes or not. For example, from Fig.3 in Method part, the marker gene VWF for Endothelial cells showed relatively higher expression level in cell cluster 12, and therefore cluster 12 was regarded as Endothelial cells, though gene VWF did not rank in the top 5 differentially expressed genes in cluster 12. In this way, the rest of the cell clusters were identified. One interesting point was that some cell clusters (1,2,9; 3,6) were assigned to the same cell types (Alpha / Beta cells respectively) since they both showed relatively high expression levels of the marker genes presented in the original paper. Also, some cell types (Delta / Epsilon; Activated / Quiescent stellate; Macrophage / Mast) in the paper were assigned to the same clusters (0, 7, 11 respectively) as their marker genes both showed relatively high expression levels in these clusters. This would be discussed in the Discussion part. Additionally, one cell cluster (cluster 10) could not be identified based on the marker genes the author provided. Instead, the top10 significantly expressed genes were used in the tool ‘PanglaoDB’ described in Method part to obtain a possible cell type. Some of these genes showed relatively high enrichment in ‘Basal cells’, and therefore we labeled this cluster as ‘Possible Basal cell’. The relationship between these genes and Basal cells could be verified by other research studies. After labeling the clusters, the UMAP plot was made and shown below.

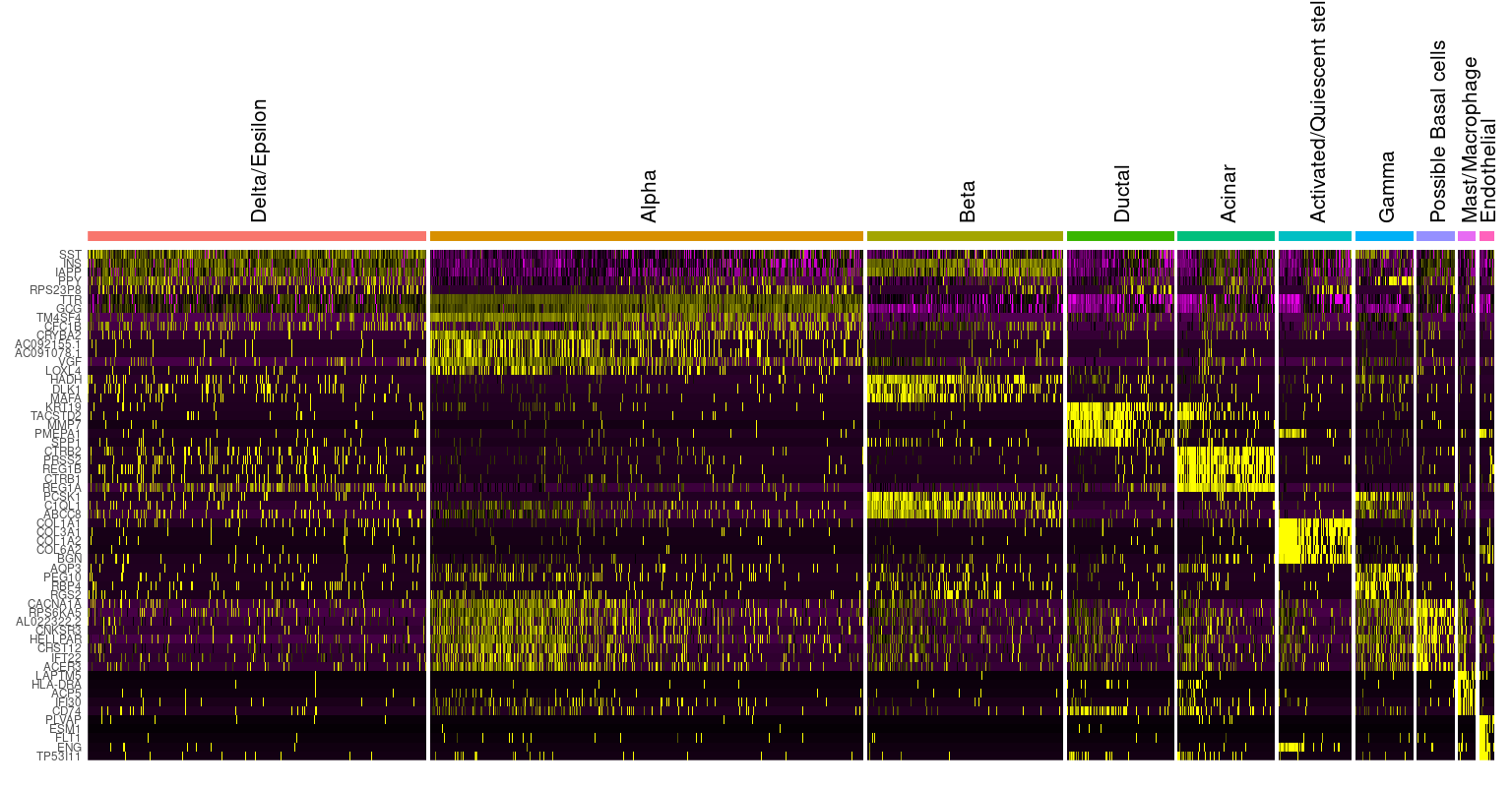
**Projection plot colored by cluster and labeled with assigned cell type**

****

**Fig.10 UMAP plot for cell clusters labeled with different cell types in different colors**

Fig.10 showed the UMAP 2-dimension plot for the cell clusters. Here cluster 1, 2, and 9 were labeled as Alpha and therefore they were combined together as one cluster ‘Alpha’. Cluster 3 and 6 were labeled as Beta and were combined together as ‘Beta’ as well. The comparison between this UMAP plot and Fig.1D in the paper would be discussed in the Discussion part.

**Clustered heatmap of top 5 marker genes for each cluster**

****

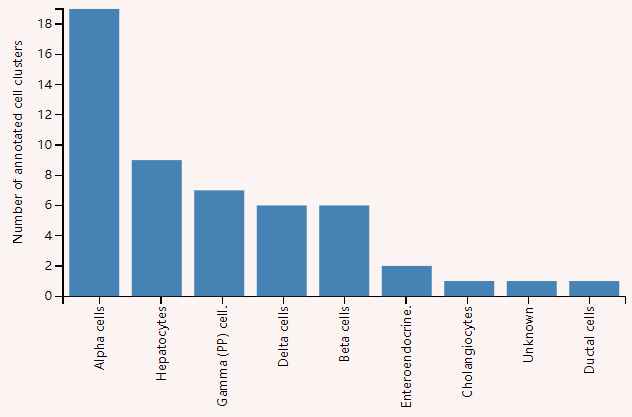
**Fig.11 Clustered heatmap of top 5 marker genes for each cluster**

Fig.11 showed the clustered heatmap of top 5 marker genes for each cluster. For most of the cell types, they seemed to have a clear expression pattern. The top 5 marker genes were solely highly expressed in these cell types. One exception was the Delta/Epsilon cells. They seemed to have a relatively noisy expression pattern, which means the expression levels for the top 5 marker genes in some of the cells were not significantly high. One possible reason was that it was mixed with Delta cells and Epsilon cells and the top 5 significant expressed genes were mostly from one of these two types of cells. The comparison between this heatmap and Fig.1B in the original paper would be discussed in the Discussion part.

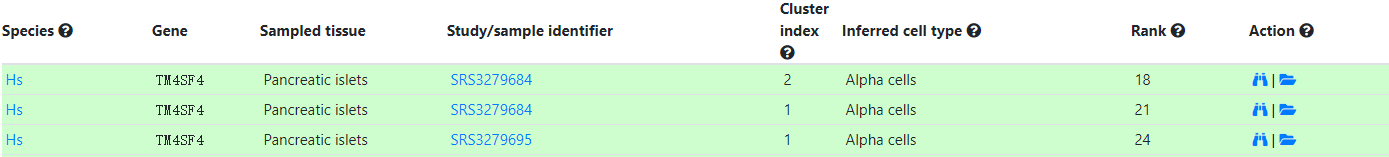
**Obtain novel marker genes for the labeled clusters**

|  |  |
| --- | --- |
| **Assigned cell type** | **Possible novel marker genes** |
| Delta / Epsilon | SEC11C, RPS23P8 |
| Alpha | TTR, TM4SF4, CRYBA2, VGF |
| Beta | IAPP, HADH, DLK1, MAFA, C1QL1 |
| Ductal | TACSTD2, SERPINA5, MMP7, SPP1 |
| Acinar | CTRB2, PRSS2, REG1B, CTRB1, SPINK1 |
| Activated / Quiescent stellate | COL3A1, COL1A2, COL18A1 |
| Gamma | PEG10, PCSK1N |
| Basal cells | RPS6KA5, IFT22, LTBR |
| Macrophage / Mast | LAPTM5, TYROBP, ACP5, |
| Endothelial | PLVAP, ESM1, PECAM1, RGCC, F2RL3 |

**Table.4 Novel marker genes for the labeled clusters**

****

**Fig.12 Bar plot of cell clusters (Y-axis) and cell types (X-axis) where the gene TM4SF4 is expressed in tool ‘PanglaoDB’**

****

**Fig.13 TM4SF4 as a marker gene for Alpha cells**

The novel marker genes for each labeled cluster were identified using the method described in Method part. The top 10 differentially expressed genes in each cluster were selected out as marker genes and then verified in the tool ‘PanglaoDB’. The results were summarized in Table.4. The Fig.7 and Fig.8 were an example about how we verified these marker genes. For the marker gene TM4SF4 for Alpha cells, we placed it in the ‘PanglaoDB’ and a bar plot was returned showing a high enrichment/expression level of this gene in Alpha cells. The Fig.8 from this tool indicated that the inferred cell type for this marker gene was also Alpha cells, which corresponded to our result. With this approach we found out several marker genes for each of the labeled clusters. Some differentially expressed genes for each cluster were not included because they were not special or unique for the cell type and could not be verified by any other research studies or ‘PanglaoDB’.

**Discussion**

In general, most of the cell types presented in Fig.1D of the original paper had been successfully identified in our analysis. One extra possible cell type ‘Basal Cell’ was also identified using the identification strategy. Nevertheless, some of the cell types, Delta / Epsilon, Activated / Quiescent stellate and Macrophage / Mast were assigned to 1 cluster respectively. The possible reason was that these different types of cells were similar to some degree (e.g. subtypes of one cell type) and did not totally separate during the clustering method. Another possible reason was that too many relevant cells were filtered out and thus affected the expression level of the genes and resulted in the failure of the separation. If we would like to separate them well, several possible steps could be taken: (1) When building the index with transcriptome FASTA file, reduce the k parameter (k acts as the minimum acceptable length for a valid match) to some extent. This would allow a relatively higher mapping rate and therefore obtain some marked genes with higher expression level used to define the cell types. (2) Provide a larger whitelist of the barcodes, so that more cells would be included and provide more information about the marker gene expression level. (3) Use a less strict threshold (e.g. minimum total number of genes a cell should have; maximum percentage of mitochondria genes a cell should have - in other cases, since we had no info on mitochondria in this case), don’t filter out results based on relation of genes to molecules for gene and cell quality control. Sometimes a too strict threshold would filter out many genes and cells that might provide useful and interesting information. (4) Use more features - don’t filter out as many for variance filtering (4) Increase the resolution when finding the clusters. Usually when increasing the resolution in FindCluster function in Seurat we would be able to obtain more cell clusters and therefore many subtypes of the cells would probably be given. Thus, it would be helpful to increase the resolution to further separate the clusters into more small clusters and then it would be easier to identify the cell types that were mixed together.

Another interesting information was that some clusters (cluster 1,2,9; cluster 3,6) could be assigned to one cell type (Alpha, Beta cells respectively). This could be well explained by the cell subtypes: the different clusters from one cell type were actually the subtypes of that kind of cells and they shared the same marker genes but also had many other different expressed genes, and therefore they were separated into different clusters. This was often not a big problem since we were able to label them together if we did not need to figure out their specific subtypes. Due to the time restriction, we did not search for information about the corresponding subtypes for these cells.

When comparing the UMAP plot we created with the Fig.1D in the original paper, most of the cell types were successfully identified. Here we used a better visualization dimension reduction method UMAP, while the original paper applied a T-SNE method which could not actually reflect the position and distance between cell clusters. In our plot, the largest cluster was the alpha cluster, followed by Beta and Delta/Epsilon clusters. While in the authors’ plot, the largest cluster was the Beta cluster, followed by Alpha and Delta clusters. The slight difference here might be due to the following reason: (1) Difference between the size of the whitelist. (2) Difference in using alignment method (Salmon vs Bowtie), result in different mapping rates, and therefore some cells would be removed. (3) Different thresholds used in analysis. (4) Different clustering method (shared nearest neighbor in Seurat vs hierarchical clustering). The Activated stellate and Quiescent stellate were separated well in the authors’ plot, while we failed to separate them from one cluster. Additionally, the number of the Endothelial cells in our plot seemed to be fewer than the number in authors’ plot. The reason might be similar as mentioned above. There were two cell types, Mast/Macrophage and Basal cells, identified in our analysis, while they were not presented in the authors’ plot. One possible reason was that these cells are two few and had not been included in their results.

The clustered heatmap of marker genes for each cluster we created was kind of different from Fig.1B in the original paper. In our heatmap, we selected the top 5 differentially expressed genes for each cell type, while the authors’ plot only showed one pre-selected marker gene for each cell type. Here only GCG, INS, SST, PPY and KRT19 shown in the Fig.1B were presented in our heatmap. Nevertheless, these single marker genes usually could not define a cell type, and a reasonable way was to apply a set of the marker genes specifically for the cell type. So here we may see that these marker genes for different cell types would group together, but in fact these cell types were defined using a combination of marker genes, not only one single marker. Since we were only using one sample for analysis, we did not include the color bar for different samples as the Fig.1B showed.

**Conclusion**

The cell types presented in Fig.1 of the original paper were successfully identified using the Seurat R package and some other outer sources described in Method part. Some new possible types of cells such as Basal cells had also been discovered but required further investigation and verification. It could be concluded that usually one single gene marker was unable to truly define a cell type of an unlabeled cluster. To obtain a more accurate result, a group of marker genes of a specific cell type should be used. Additionally, it might also depend on the source of samples (e.g. different tissues). Another challenge was that one cell type might have some different subtypes. These subtypes could also be investigated but required more related biological knowledge to interpret it. It was interesting to note that some cell types were assigned to one cell unlabeled cluster. It was due to the similarity between the two cell types, and one way to separate them was to increase the resolution of the clustering method. Another difficulty was that the analysis of the cells and genes required strong biological background and it was necessary to do some search for the information and relationship between cell types and corresponding genes.

Every step could make some difference on downstream results. It would be interesting to test different settings in each step. For example, different clustering methods, different thresholds, different numbers in whitelist or different tools used for mapping would result in different numbers of cells or clusters and various gene expression levels. But the result should still be similar to some degree if the steps were reasonable. In the future study, these settings could be explored to obtain a deeper and more comprehensive understanding of single cell sequencing analysis.

**References**

[1] 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.

[2] Butler, A., Hoffman, P., Smibert, P. et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36, 411–420 (2018).

[3]Klein AM, Mazutis L, Akartuna I, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161(5):1187-1201.

[4] Oscar Franzén, Li-Ming Gan, Johan L M Björkegren, PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data, Database, Volume 2019, 2019

[5] McInnes, L, Healy, J, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426, 2018