**Individual Project 5: Single Cell RNA-Seq Analysis of Pancreatic Cells**

Author: Preshita Dave (Data Curator, Programmer)

**Introduction**

Dysfunction of the pancreas is clinically important for type 1 (T1D) and type 2 diabetes mellitus (T2D), pancreatitis, and cancer. Beta cells lost in type 1 diabetes have made the pancreas one of the focus organs of current regenerative biology. But without a complete understanding of the gene expression patterns in the pancreas, the efforts are incomplete.

Baron et al [1] 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. The single cell RNA-seq (scRNA-seq) inDrop approach combats the issue of masking any biological variability that may occur while processing bulk gene expression samples. It provides a systematic approach for capturing thousands of cells without pre-sorting using high-throughput droplet microfluidics that barcode the RNA from individual cells. By using single-cell transcriptomics, substructures within cell populations that can classify cell type by function and marker gene expression can be performed.

In this project, I aim to attempt to replicate the study’s results using a 51 year old female human donor’s cells. The reads and barcodes would be processed and then aligned to a reference transcriptome to generate read counts for unique molecular identifiers (UMIs). Clustering will then be performed to determine the number of cell types found.

**Methods**

**Processing Single Cell Sequencing Reads and Choosing Barcodes**

For our study, only SRR files from the 51 year old female human donor were processed and analyzed. To identify the SRA accession number, the metadata was downloaded from the GEO accession number GSE84133, which contains the thirteen samples (sequencing libraries) and four human individuals. From the sample information, it was determined that the short read archive (SRA) accession number for the donor was SRP07832 and included the runs SRR3879604, SRR3879605, and SRR3879606 that were used for further analysis [1].

To determine the number of cells sequenced, the number of reads per distinct barcode was calculated to eliminate reads with infrequent barcodes from consideration. Barcodes from each run were extracted using *awk* and *sed*, and the counts of each distinct barcode were determined using *sort* and *uniq -c* from the pre-processed files that were provided. A whitelist of informative barcodes was then created. Only barcodes with more than 100 reads were included in this list. The objective of this step was to eliminate barcodes that were too infrequent to provide any information to the analysis.

To quantify the reads, a cell-by-gene (UMI) count matrix was generated using the salmon alevin package [2]. But before generating the count matrix, the current human reference transcriptome available on the Gencode website was downloaded (Release 40 GRCh38.p13) [3]. A transcript ID (ENSTXXX…) to gene (ENSGXXX…) map file as described in the salmon alevin documentation was created to enable salmon to collapse from the transcript to the gene level. These transcript sequences were used to build the index of the reference transcriptome using *salmon index* and were inputted using the *-i* argument. Along with the index file generated, the fasta files for both the reads, the whitelist information and custom barcode+UMI arguments to quantify the reads: *--end 5 --barcodeLength 19 --umiLength 6* were provided to generate a UMI matrix, which contains the counts per gene. This was used in the subsequent steps of the analysis.

**Processing the UMI Counts Matrix**

The UMI count matrix was imported with the help of *tximport* [4]. Ensemble gene identifiers were converted to the Gene symbols using *EnsDb.Hsapiens.v79* [5]. The next step was to convert the UMI .mtx counts into a Seurat object which was accomplished by the *Seurat* [6] package. Filtering of mitochondrial genes < 10% and nfeature\_RNA > 200 was carried out based on the data. The data was normalized through the LogNormalize method and the top 2000 variable genes were found through the FindVariableFeatures() function. In order to conduct a dimensional reduction technique to cluster the cells, the data needs to be scaled through the ScaleData() function. RunPCA() function was called on the 2000 most variable features and the PC dimensions can also be visualized. To determine the dimensionality of the dataset, an Elbow Plot was constructed. Based on the dimensions to be considered, a UMAP was plotted to visualize the clusters. Finally, to understand the proportion of cells in each cluster, a bar and pie chart were constructed. This Seurat object was saved as an RDS file to be used for further analysis into marker genes for each cluster.

**Results**

**Read Processing and Barcode Selection**

When merging all reads from the donor, 2914895 distinct barcodes had to be filtered to remove potential bias from low-quality cells. Based on the cumulative distribution graph (Fig 1) barcodes with counts less than 100 were removed and the remaining barcodes were used as input for salmon alevin.

**Chart

Description automatically generated**

Fig 1: Cumulative distribution plot of the barcode frequencies

The results from *salmon* (Table ) show that 11.16% of the cellular barcodes matched a low number of reads (noisy) and were dropped. These noisy reads could be a result of the experimental procedure carried out to prepare the cells for sequencing. Finally, 1324770031 out of a total of 1324837961 barcodes were used for the analysis, discarding 240577 barcodes. The mapping rate was found to be 43.94%. This may be due to the *k*-mer size when building the index, with smaller values of *k* potentially improving sensitivity. However, since the authors did not state the mapping rate, these results may be expected.

|  |  |
| --- | --- |
| **Barcodes Used** | 1324770031 / 1324837961 |
| **Skipping # Barcodes as no read was mapped** | 240577 |
| **Total white-listed Barcodes** | 4189196 |
| **Total % reads will be thrown away because of noisy Cellular barcodes** | 0.0111673 |
| **Total Unique barcodes found** | 4251176 |
| **Used Barcodes except Whitelist** | 1131 |
| **Mapping rate** | 43.94% |

Table 1: Salmon alevin mapping summary statistics

**Processing the UMI Counts Matrix**

The UMI count matrix consisted of 56063 features across 1317415 cells. Based on Fig 2, I decided to subset the data to include cells which had a mitochondrial content of less than 10% and nFeatureRNA > 200. This results in 56063 features 4762 cells.

A picture containing diagram

Description automatically generated

Fig 2: Violin plot of 1) number of genes per cell, 2) number of molecular counts associated with genes per cell, and 3) percent of reads that map to the mitochondrial genome.

A FeatureScatter plot shown in Fig 3 helps us to visualize feature-feature relationships. In our case, I have plotted the percent.mt vs nCount\_RNA, and nFeature\_RNA vs nCount\_RNA. The percent.mt vs nCount\_RNA doesn’t give us a very good correlation score as the data is very noisy. The nFeature\_RNA vs nCount\_RNA gives us a positive correlation score of 0.98.

**Chart, scatter chart

Description automatically generated**

Fig 3: Feature Scatter plot representing correlation between (a) percent.mt vs nCount\_RNA (b) nFeature\_RNA vs nCount\_RNA

Fig 4 identifies 2000 most variable genes, which were found through the FindVariableFeatures() function. In the figure below, the top 10 variable features are labelled. REG13B is the most variable gene, followed by PRSS3P2 and so on.

**Chart, scatter chart

Description automatically generated**

Fig 4: Scatter plot showing highly variable genes in red and naming the top 10 genes.

After scaling the data and running PCA on the 2000 most variable features, the RunPCA() function generates 50 dimensions by default. Fig 5 below represents the first 2 dimensions of the PCA components generated and how the genes capture the data in the cells.

**Chart, scatter chart

Description automatically generated**

Fig 5: Visualization of the 1st and 2nd dimensions of PCA

The top principal components represent a robust compression of the dataset. To overcome the extensive technical noise in any single feature for scRNA-seq data, Seurat clusters cells based on their PCA scores, with each PC essentially representing a ‘metafeature’ that combines information across a correlated feature set. Hence, based on the Elbow plot shown below in Fig 6 we can decide how many components should we choose to include that represents the majority of the data with very little noise. Based on Fig 6, the plot suggests that the majority of true signal is captured in the first 30 PCs.

**Chart

Description automatically generated with medium confidence**

Fig 6: Elbow plot for PC vs standard deviation, Elbow Effect seen near PC30

Clustering algorithm KNN was applied to identify cell clusters. Seurat’s FindNeighbours() function clustered the cells based on the KNN technique. FindClusters() function helped to cluster the cells, based on the modularity optimization technique called Louvain algorithm. The resolution for FindClusters function was chosen to be 0.5. The clustering is based on the first 30PC dimensions, resulting in 13 clusters shown in Figure 7.

**Chart, scatter chart

Description automatically generated**

Fig 7: UMAP plot representing the 13 cell clusters

To have a better understanding of the clusters identified, Fig 8 below shows the bar chart and pie chart that represent relative proportions of the number of cells in each cluster.

**Chart, histogram

Description automatically generated**

**Chart, pie chart

Description automatically generated**

Fig 8: 1) Bar plot, 2) Pie chart representing the distribution of the number of cells in each cluster

**Discussion**

Single cell RNA-seq provides a higher resolution of cellular differences, allowing researchers to understand cellular diversity better. It can help us to better understand disease mechanisms at the gene and cellular level. In this project, I tried to replicate the results obtained from Baron et al. [1] using three samples from a 51-year-old female. The reads were mapped based on customized barcodes, and the low quality readings were eliminated. Using the Seurat package in R [6], the data was pre-processed and cells were clustered based on gene expression. The reproducibility of the findings from Baron et al. [1] the study was not perfectly replicated due to minor inconsistencies. A better result reproduction can be gained when using more samples to perform the analysis. Overall, the study provided a novel method to characterize transcriptome profiling using single cell RNA-Seq and use the result to study disease pathology.

**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. 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
3. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC, Armstrong J, Barnes I, Berry A, Bignell A, Boix C, Carbonell Sala S, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, García Girón C, Gonzalez JM, Grego T, Hardy M, Hourlier T, Howe KL, Hunt T, Izuogu OG, Johnson R, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Riera FC, Ruffier M, Schmitt BM, Stapleton E, Suner MM, Sycheva I, Uszczynska-Ratajczak B, Wolf MY, Xu J, Yang YT, Yates A, Zerbino D, Zhang Y, Choudhary JS, Gerstein M, Guigó R, Hubbard TJP, Kellis M, Paten B, Tress ML, Flicek P.

*Nucleic Acids Res* 2021 : **49** ; d1 ; D916-D923. PUBMED: [33270111](https://www.ncbi.nlm.nih.gov/pubmed/33270111); PMC: [PMC7778937](https://europepmc.org/articles/PMC7778937); DOI: [10.1093/nar/gkaa1087](http://dx.doi.org/10.1093/nar/gkaa1087)

1. *tximport*. Bioconductor. (n.d.). https://bioconductor.org/packages/release/bioc/html/tximport.html.
2. Rainer J (2017). EnsDb.Hsapiens.v79: Ensembl based annotation package. R package version 2.99.0.
3. Hao Y, Hao S, Andersen-Nissen E, III WMM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zagar M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LB, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R (2021). “Integrated analysis of multimodal single-cell data.” *Cell*. [doi:10.1016/j.cell.2021.04.048](https://doi.org/10.1016/j.cell.2021.04.048), <https://doi.org/10.1016/j.cell.2021.04.048>.