FINAL PROJECT

BF528- Applications in Translational Bioinformatics

-Rojashree Jayakumar

Section 1: Project 4: Processing the UMI counts matrix (Programmer)
Section 2: Project 4: Cluster marker genes (Analyst)
Section 3: Project 3: RNA-Seq Sample Statistics and Alignment (Data- Curator)

SECTION 1 - Single Cell RNA-Seq Analysis of Pancreatic Cells

Project 4: Processing the UMI counts matrix (Programmer)

Introduction

Unlike bulk RNA sequencing, single cell RNA sequencing can give information about the transcriptional profile for each cell. Therefore one of the major applications of single cell sequencing is its use in identifying cell type heterogeneity and information about the transcriptional profile of low population of cells (Tang et al., 2019). It has several applications in various fields of medicine such as immunology and oncology. Tumor development is driven by cell type heterogeneity and single cell sequencing can be harnessed to identify and classify cell types in the tumor microenvironment (Zhang et al., 2021).

Baron et al had performed single cell sequencing on 12000 cells from 4 human donors and 2 mouse strains. Using the droplet-based (inDrop) method, they were able to establish the transcriptional landscape of pancreatic cells and identify distinct cell subtypes in the pancreas. They also performed differential expression analysis associated with disease.

The Unique molecular Identifier (UMI) matrix was generated by me as part of the Data Curator task in Project4. As part of the final project, I want to continue with further downstream analysis to gain insights into the single cell analysis workflow using the SEURAT in the Bioconductor package in R. In this section, the UMI matrix is processed to filter out low quality genes and cells with low variance and filtered cells were clustered using UMAP to identify the subpopulations in the pancreatic cells. Filtering the UMI matrix is very important as most of the matrix will be sparse, due to some cells lacking many genes. Sometimes due to sequencing errors, doublets might get sequenced and a single cell will have double the number of transcript counts for all genes. Furthermore some genes will have low variance or will not be expressed

altogether in all the cells and hence they should be filtered out along with mitochondrial genes to reduce noisy data. Finally clustering is important to identify cell type populations.

Methods

Filtering the low quality cells:

I had generated three UMI matrices in Project 4 and I chose a UMI matrix that was built using decoy aware transcriptome at a threshold of cells less than 20 whitelisted. To process the UMI matrix, the UMI matrix previously generated from the sample of a 51 year old female donor was loaded in the form of salmon alevin output into a R session using tximport and fishpond packages. Instead of using the workflow by Baron et al to process the UMI counts matrix, the Suerat package in R was utilized to create a Suerat object which serves as a container containing both the data and the analysis for a single dataset. To identify the mitochondrial genes in the gene set, the ensembl ids that are in the data matrix in the Suerat object were renamed with the corresponding gene names. The standard preprocessing workflow was followed to generate violin plots of the distribution of the number of features (i.e genes) per cell, number of counts for per cell and percentage of mitochondrial genes in each cell (Figure 1). These distributions can be used to assess the filtering criteria for the dataset.

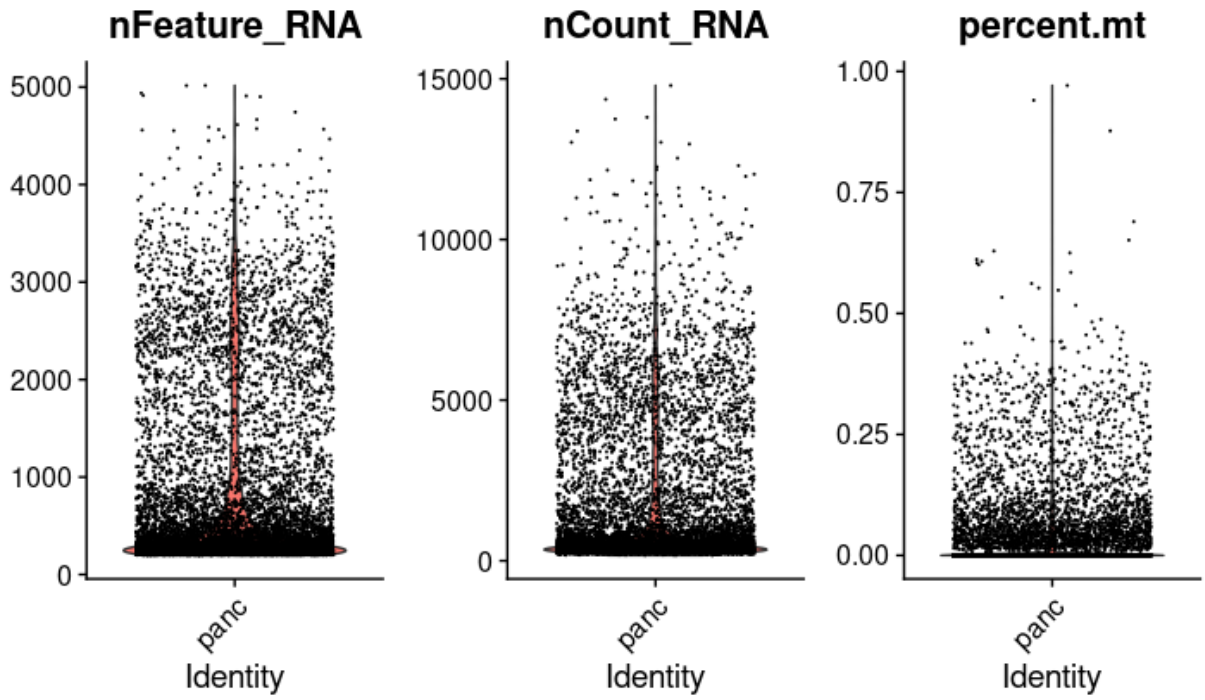

Figure1: Violin plots showing the distributions of the number of features (i.e genes) per cell, number of counts for each gene per cell and and percentage of mitochondrial genes in each cell.

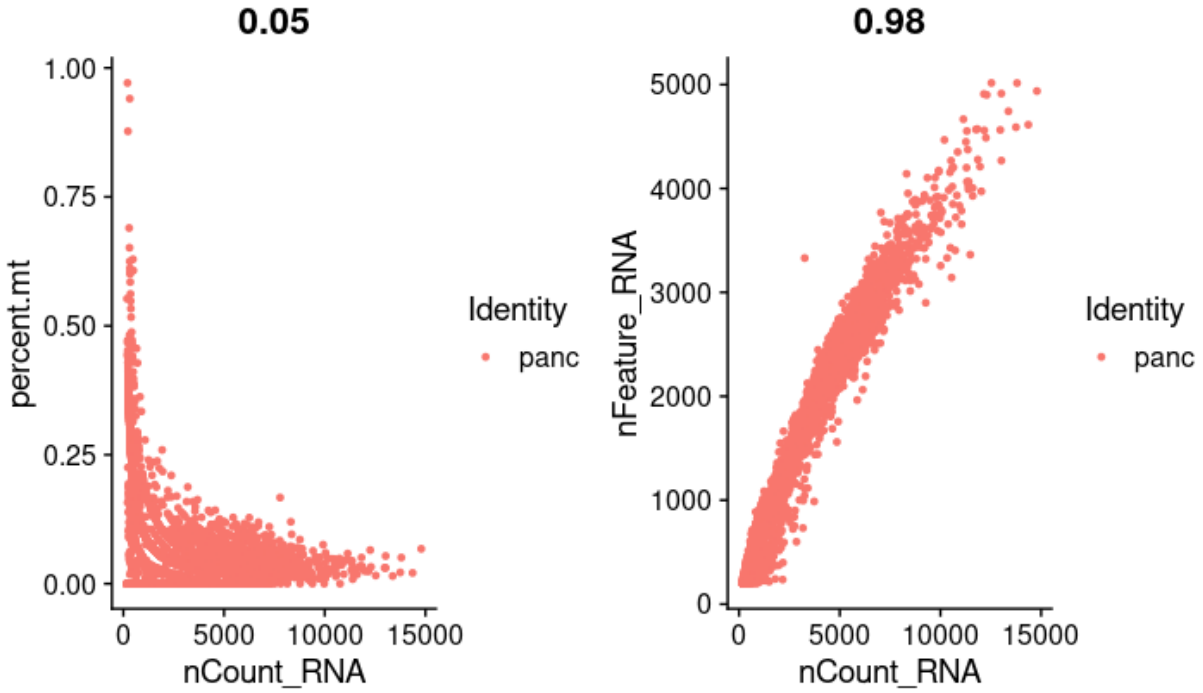

Figure 2: (A) Scatter plot of number of counts in each cell vs percentage of mitochondrial genes, the correlation between them is 0.05. (B) Scatter plot of number of counts in each cell vs number of features (i.e number of genes), The correlation between the number of genes in the cell population and the number of counts for each cell is 0.98 indicating high correlation.

Filtering genes based on variance

To filter the genes the data has to be normalized. The method used in Suerat normalizes the feature expression for each cell by the total expression of that cell and multiplies it by a scaling factor (I used 10000 as the scaling factor, which is the default as per Seurat) and the values are log transformed. The FindVariableFeatures function was used to find genes with highest variance (default =2000). Only these genes will be retained for further analysis.

Clustering:

To cluster the cell types, the dimensionality of the dataset has to be determined. To determine the dimensionality linear dimensionality reduction such as PCA was performed. Either a heuristic approach such as an elbow plot can be used to identify the dimensions which capture most of the variance or a jackstraw plot can be used to identify the dimensionality of the data. The FindNeighbours and FindClusters methods were used to find the clusters and then UMAP was run with the dimension that was detected through the Jackstraw and elbow plots (here it was 10) to visualize the clustering of cells.

Results

Filtering the low quality cells

From Figure1 it can be seen that most of the features lie in the range 200 to 2500 and most cells have a mitochondrial genes percentage at 5%. Therefore cells with features between 200 and 2500 were retained and cells with greater than 5% are removed from further analysis.

Filtering the genes based on variance

By default the top 2000 variable features (genes) are returned using the FindVariableFeatures function which and only these genes will be used in further downstream analysis (Figure 3).

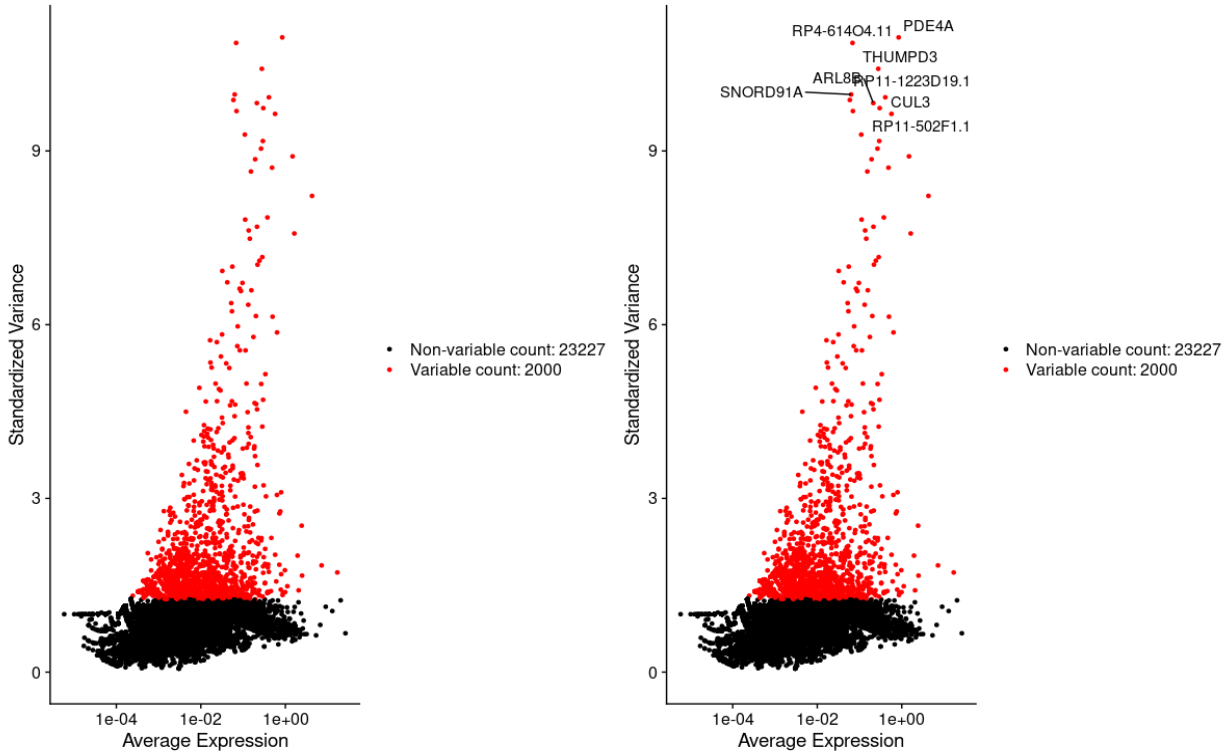

Figure3: Scatter plot showing the average expression of the genes vs their variance across different cell types. The genes with high variance are marked in red. The top 10 genes with high variable expression are labeled on the plot on the right.

The original UMI matrix had the dimensions of 27060 cells with 9928 genes. After filtering the dimensions of the UMI matrix changes significantly (25227 cells and 9005 genes). Though the genes are reduced significantly, in downstream analysis only the top 2000 variable

genes (having the highest variance in the dataset) will be utilized as obtained by applying the FindVariableFeatures method on the UMI matrix.

	Number of cells	**Number of genes**
Original UMI matrix	27060	9928
Filtered UMI matrix	25227	9005
UMI matrix for downstream analysis	25227	2000

Table 1: Dimensions of the UMI matrix before and after filtering.

Clustering

Both the elbow plot and Jackstraw plot (Figure 4 and 5) showed that the most of the variance was captured in the first 10 principal components. These methods estimated the dimensionality of the dataset to be around 10. Hence for further analysis the first 10 dimensions of the data were utilized. Non-linear dimensionality reduction, UMAP, was used to view the clustering (Figure 6). There were a total of 12 clusters, with the first cluster representing the greatest number of cells representing 35% of the total cells (Figure 7 and 8).

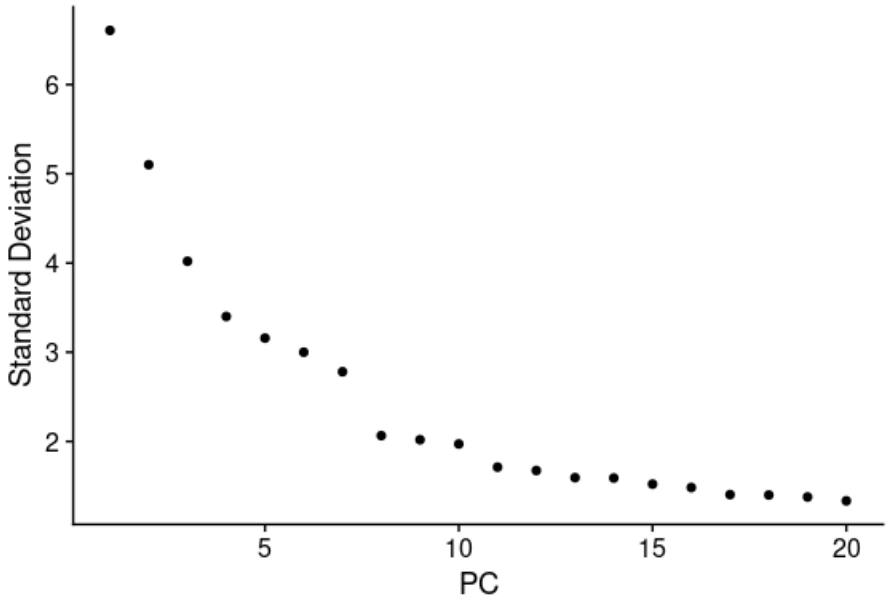

Figure 4: Elbow plot to determine the dimensionality of the data. The elbow occurs near PC 10, which indicates that most of the variance of the data is captured by first 10 principal components

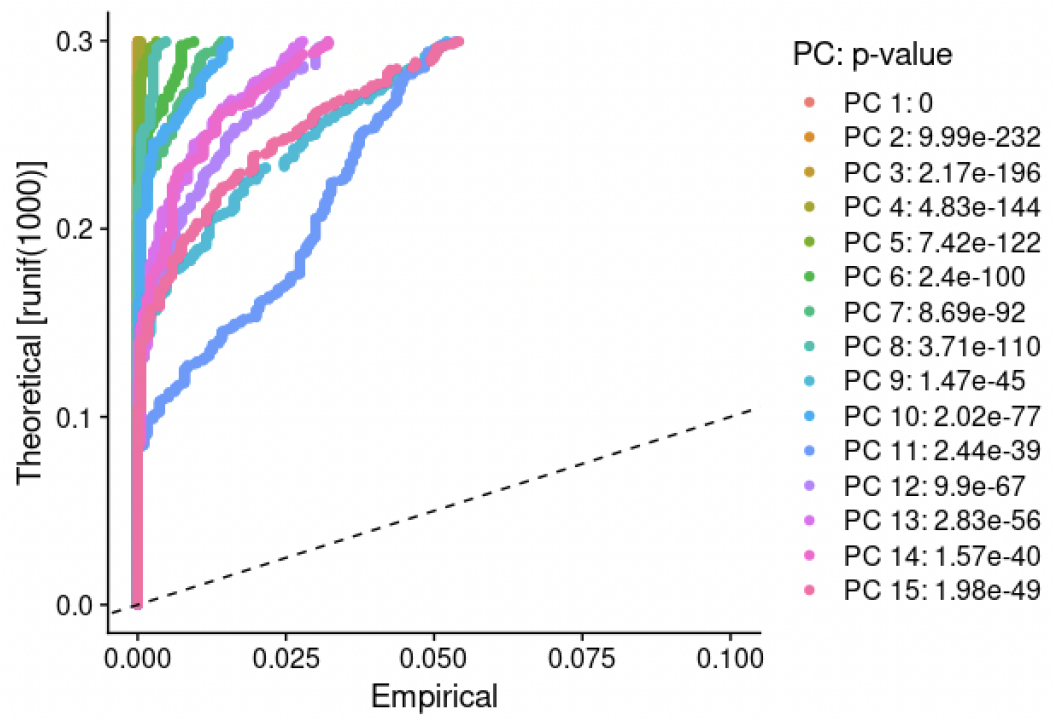

Figure 5: Jackstraw plot to determine the dimensionality of the data. There is a sharp drop in the p-value between PC 10 and 11, indicating that most of the data dimensions are captured by the first 10 principal components.

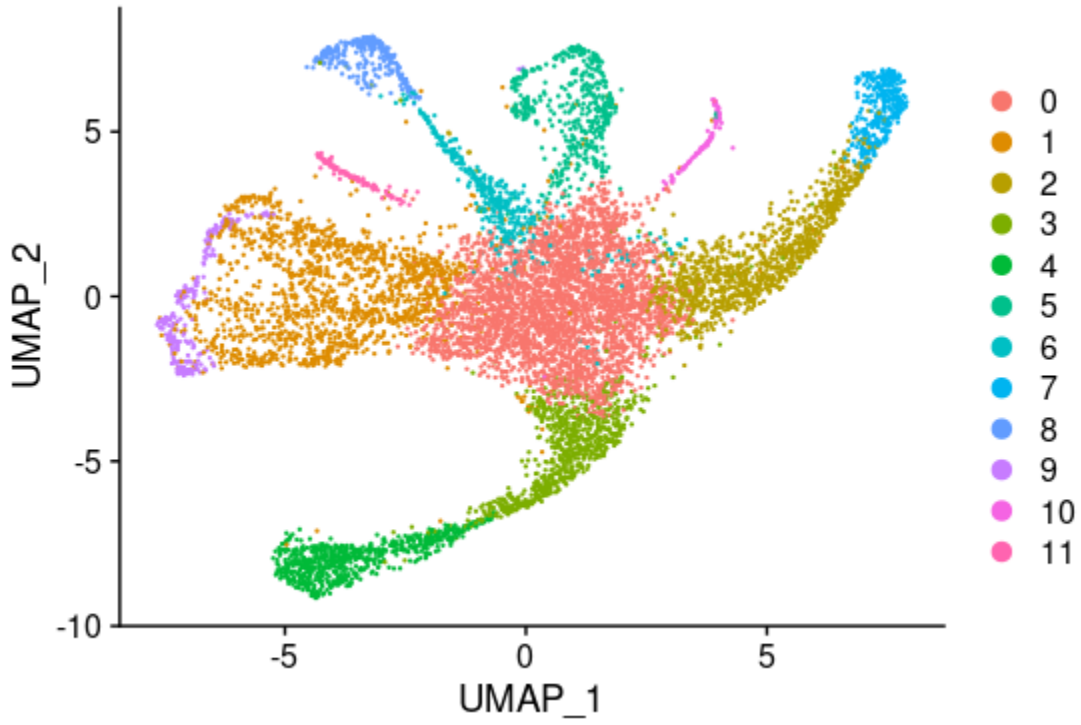

Figure 6: Umap clustering of the cells. A Total of 12 clusters were identified.

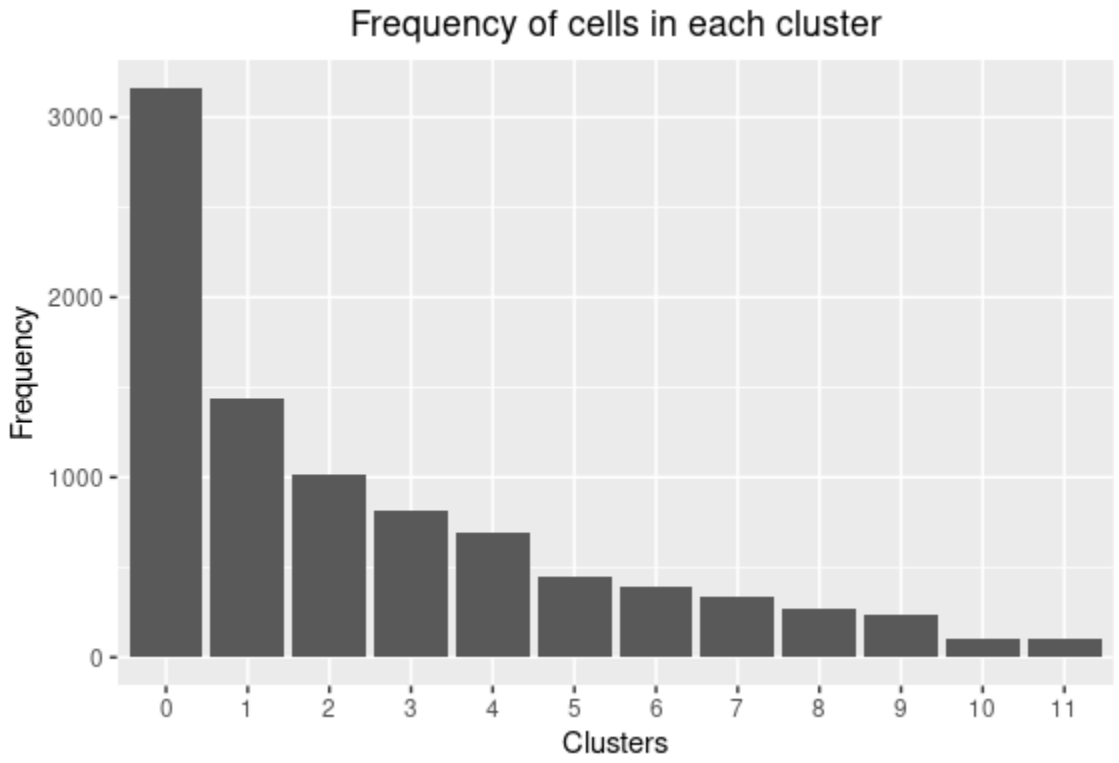

Figure 7: Barplot showing the frequency of cells in each of the clusters.

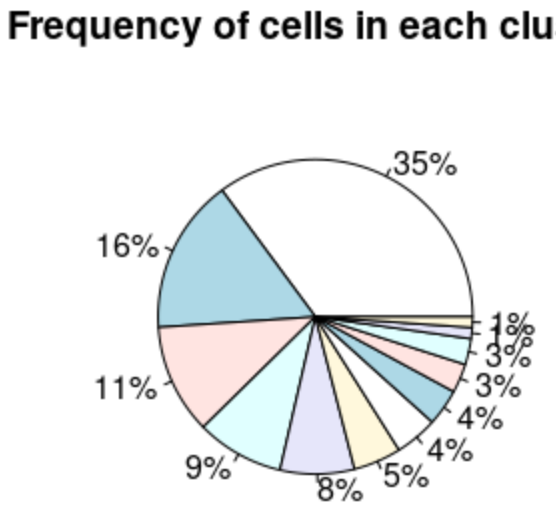

Figure 8: Pie Chart showing the relative frequency of cells in percentages in each of the clusters.

Discussion

The goal of this analysis was to reduce the noise in the dimensions of the UMI matrix so downstream analysis can be performed effectively. Furthermore, cells were clustered using UMAP.

Following filtering the genes and cells the dimensions of the matrix reduced significantly from 27060 cells to 25227 cells and only the top 2000 genes with the highest variance were used for further analysis.

There were a total of 12 clusters, with the first cluster representing the greatest number of cells representing 35% of the total cells (Figure 8). However as per the paper there were 15 clusters. The higher number of cell subpopulations in the paper may be a result of using more samples in the original study and hence a greater pool of cells in the first place. Another reason why my analysis yielded lower cells would be because I used UMAP whereas the original study made use of t-SNE to cluster the cells into distinct subpopulations. Therefore this section was successfully able to reduce the dimensions of the data, however the clustering was not accurately reproduced.

SECTION 2 - Single Cell RNA-Seq Analysis of Pancreatic Cells

Project 4: Cluster marker genes (Analyst)

Introduction

Unlike bulk RNA sequencing, single cell RNA sequencing can give information about the transcriptional profile for each cell. Baron et al had performed single cell sequencing on 12000 cells from 4 human donors and 2 mouse strains. Using the droplet-based (inDrop) method, they were able to establish the transcriptional landscape of pancreatic cells and identify distinct cell subtypes in the pancreas. They also performed differential expression analysis associated with disease.

The Unique molecular Identifier (UMI) matrix was generated by me as part of the Data Curator task in Project4. I have also completed the processing of the UMI counts matrix in the first section of this project. In this section, the top marker or representative genes for each cluster was identified and the cluster was labeled with a cell type name depending on which cell type the marker gene represented. Additionally other representative genes were also identified for each cluster which might be as discriminative as the marker genes for the cell type. Identifying the marker gene is important, as marker genes can be used for assigning the cell type to each of the clusters.

Additionally the marker genes are generated based on differential expression of the genes in each cluster and the genes which have low log2 fold change represent the cluster best. Therefore identifying and assigning the cell types to clusters is very important for further downstream analysis such as gene set enrichment which will aid in gaining biological insights such as which genes in the cluster are enriched for what functions or pathways.

Methods

Identifying marker genes that drive clustering

The filtered Seurat object which was stored in a RDS file was read. The FindAllMarkers function was used to determine the markers that drive each cluster based on differential expression. This method was used to find markers for every cluster by comparing each cluster to every other cluster. The minimum percentage of the genes between each cluster pair was set at 0.25% and the log fold threshold was set at 0.25. The minimum percentage of cells was selected to speed up the time. Only positive markers were reported.

Annotation and visualization of the clusters

To visualize the clustering, UMAP was run and the clusters were annotated with the cell types by knowing which marker genes represent which cell type. Baron et al have described the marker genes corresponding to the cell types and the same were used to label the clusters. Additionally the Panglao database was also used for annotation of the clusters. The clusters were visualized using a scatter plot. A heatmap was used to view the differential expression of the top 10 markers based on average log2 fold change.

Results

Similar to Baron et al the largest cluster was of Beta cells in this analysis, followed by alpha cells. However the clustering was not similar to the one in Baron et al. The marker genes that were defined for macrophages, cytotoxic T, Epsilon and mast cells were not detected in this analysis (Figure 9). Hence the Panglao database was used to identify which cell type the top 2 marker genes of the unmapped (after mapping other clusters with Baron et al) clusters mapped to. Beta cells were represented by two clusters (cluster 0 and 3). The marker genes from both Baron et al and Panglao database were detected in both the clusters. Only the cytotoxic T cell type could not be mapped to any cluster as the marker genes in both the Panglao database and in Baron et al were not detected in the differentially expressed genes.

Cell type	Top 2 novel marker genes	Marker genes (Baron et al)	Marker genes (Panglao database)
Alpha	TTR, CRYBA2	*GCG*	*GCG*
Beta	*INS*, SST, MAFA	*INS*	*C1QL1*, INS2
Delta	XACT, *SST*,ACER3	*SST*	*SST*
Gamma	MTND1P23, *PPY*	*PPY*	*PPY*
Epsilon	CRYBA2	GHRL	GHRL, *ARX*
Ductal	TACSTD2,*KRT19*	*KRT19*	TFF1
Acinar	CTRB2,PRSS1	*CPA1*	PRSS1
Stellate	COL1A1,COL3A1	*PDGFRB*	COL6A1
Vascular	PLVAP, FLT1	*VWF*, PECAM1, CD34	*VWF*
Macrophage	*IFI30*,ACP5	CD163, CD68, IgG	*IF130*, CD68
Cytotoxic T	-	CD3, CD8	CD8A
Mast	REG3A,*REG1B*	TPSAB1, KIT, CPA3	*REG1B*, SLC29A1

Table 2: Marker genes for each cell type from Baron et al and Panglao database. The italicized genes were used as markers for cell types.

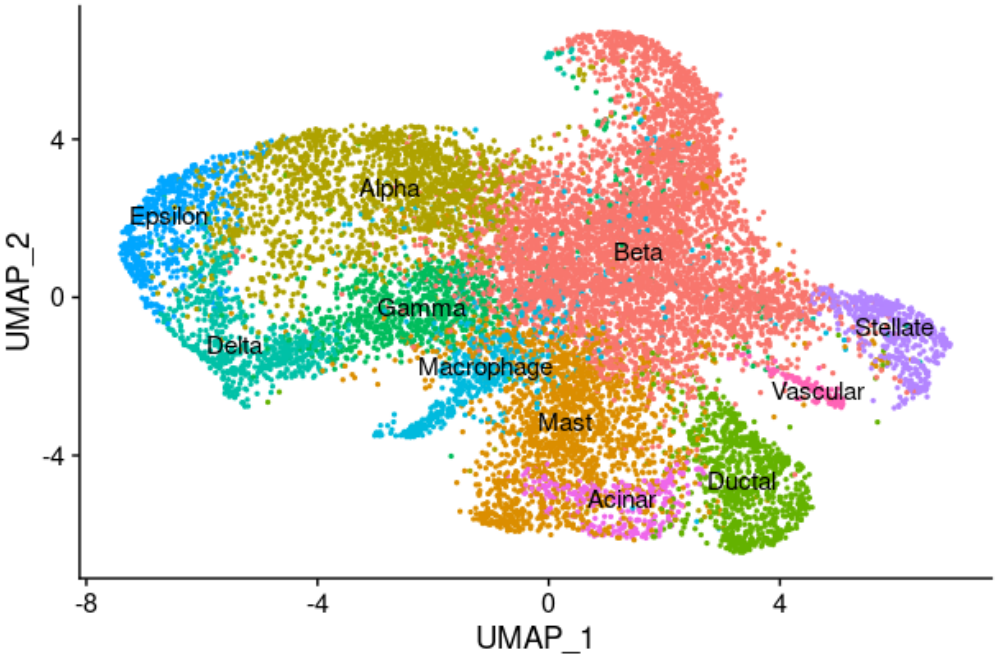

Figure 9: Scatter plot of the first two dimensions of UMAP to identify clusters. The clusters were annotated as per the cell types represented by marker genes of each cluster.

Heat map was generated to visualize the differential expression of the top 10 genes for each of the clusters (figure 10). A clear difference in expression for mast cells, stellate cells, ductal and delta cells was noted. There was not a significant difference in expression in beta and delta cells from other clusters, as these genes are co-expressed in other cell types too.

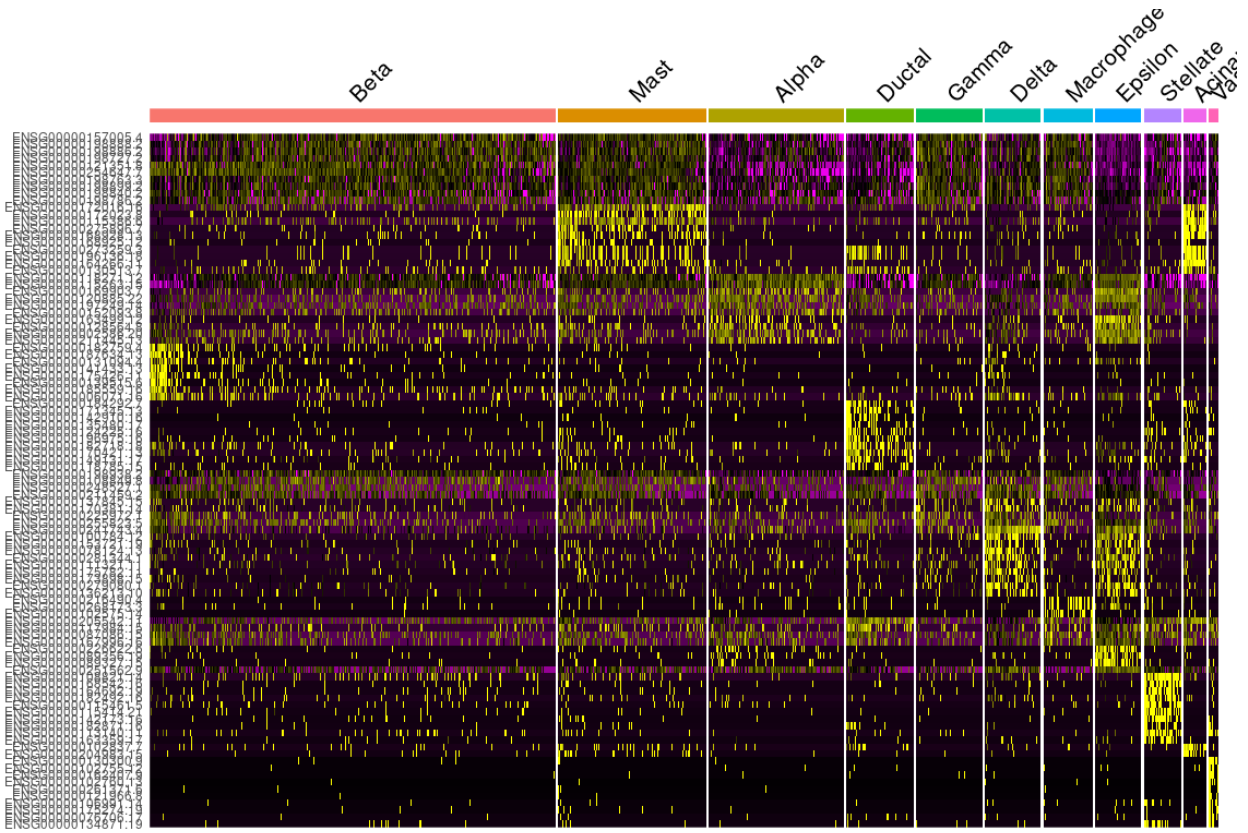

Figure 10: Heatmap of the top 10 differentially expressed genes for each cluster based on average of log2 fold change.

Discussion

The goal of this analysis was to identify which cell type each of the clusters belonged to, so that downstream analysis involving gene enrichment could be performed. A total of 12 clusters were obtained, of which there was an overlap and two clusters were assigned to the same cell type, Beta cells. One of the cell types could not be assigned. The UMAP projection which was obtained was different from the one generated by Baron et al and this could be because they used t-SNE on a larger dataset. However the heatmap of the top marker genes was similar indicating that the clustering though not reproduced may be close to the results obtained by Baron et al.

This analysis was significant because the heatmap of the top 10 genes was similar to the paper, however the clustering and annotation of the clusters was quite ambiguous and hence was not reproduced. Hence overall, I was able to replicate partially the results from Baron et al.

SECTION 3 - Concordance of microarray and RNA-Seq differential gene expression

Project 3: RNA-Seq Sample Statistics and Alignment (Data- Curator)

Introduction:

Microarray technologies have been well established for studying RNA expression across genes. Recently RNA-sequencing has gained popularity for studying the transcriptomic profile of organisms. The study by Wang et al chose to assess RNA sequencing by detecting its concordance with the well established microarray platforms. In this study they treated rat livers with 27 chemicals representing 5 modes of action. They identified that the cross platform concordance with respect to enriched pathways and differential expression was highly correlated with treatment effect size, gene expression abundance and modes of action. They also discovered that RNA sequencing outperforms microarray in differential expression and also in detecting lowly expressed genes. When they ran the same set of analysis on the training set, both platforms performed similarly and hence they concluded that the choice of platform depends on the endpoint studied, biological complexity and transcript abundance.

In this section, I chose to do RNA sequencing statistics and alignment of the RNA sequences. I chose tox-group 2 which includes the samples collected from rats that were exposed to beta-naphthoflavone, econazole, and thioacetamide and also control samples for each of these chemicals. Alignment is obviously the most important step in any sequencing platform as downstream analysis depends on the quality of aligned reads to detect any significant biological meaning.

Methods

Each chemical has 3 samples, therefore a total of nine samples were used for alignment. Preprocessed microarray and RNA sequencing count files for the controls were provided already. To assess the sample quality fastqc was run on all the nine samples. Following fastqc star alignment was performed on all the samples. To run the alignment a reference mouse genome index was used. Finally multiqc was run in the sample directory. Multiqc will princess both the star alignment and fastqc summary statistics into a single report.

Results

Alignment for all reads was completed within 2-3 hours. Multqc report detected that all the samples aligned greater than 80% accuracy (Table 3 and Figure 11). Most of the samples had an alignment almost close to the 90% mark indicating that the alignment was overall good and can be considered for further analysis. The multi mapped percentage was also below 5% and there were no reads that were unmapped for all the samples.

Sample	total_reads	uniquely_mapped	uniquely_mapped_percent	avg_mapped_read_length	num_splices	num_annotated_splices	num_GTAG_splices	num_GCAG_splices	num_ATAC_splices	num_noncanonical_splices	mismatch_rate	deletion_rate	deletion_length	insertion_rate	insertion_length	multimapped	multimapped_percent	unmapped_mismatches_percent
SRR1177966	18974404	15983143	84.24	198.98	8962150	0	8898757	42945	2488	17960	0.68	0.01	1.87	0.01	1.49	807481	4.26	0
SRR1177969	19310282	16494573	85.42	198.9	9862160	0	9795377	46704	2641	17438	0.75	0.01	1.85	0.01	1.46	695194	3.6	0
SRR1177970	19120119	16262555	85.05	198.93	9603101	0	9538674	44477	2534	17416	0.72	0.01	1.86	0.01	1.45	670339	3.51	0
SRR1177993	19823897	17094859	86.23	198.67	10137342	0	10059623	55925	2894	18900	0.75	0.01	1.85	0.01	1.53	858349	4.33	0
SRR1177994	21044593	18514298	87.98	199.04	11145326	0	11066585	54699	2821	21221	0.61	0.01	1.78	0.01	1.48	859510	4.08	0
SRR1177995	20151588	17661373	87.64	198.76	10515768	0	10439851	53184	2523	20210	0.7	0.01	1.8	0.01	1.49	794497	3.94	0
SRR1177998	23610281	20972287	88.83	199.03	13111485	0	13030035	56133	2994	22323	0.65	0.01	1.84	0.01	1.48	936865	3.97	0
SRR1178001	24936439	22219635	89.11	199.11	13591693	0	13503127	62043	3085	23438	0.67	0.01	1.84	0.01	1.46	974756	3.91	0
SRR1178003	24721331	22041739	89.16	199.07	13763863	0	13675438	62928	3017	22480	0.68	0.01	1.88	0.01	1.46	926439	3.75	0

Table 3: Read and alignment statistics from STAR alignment

The Fastqc statistics for all the samples were also acceptable. The mean Phred quality scores fall towards the 3’ end for all samples and this is expected due to RNA degradation at the 3’ end. However most of the reads have a score above 30 in all samples and hence the analysis can be processed further (Figure 13). The GC counts for all the samples follow a normal distribution which is a good quality metric (Figure 14).

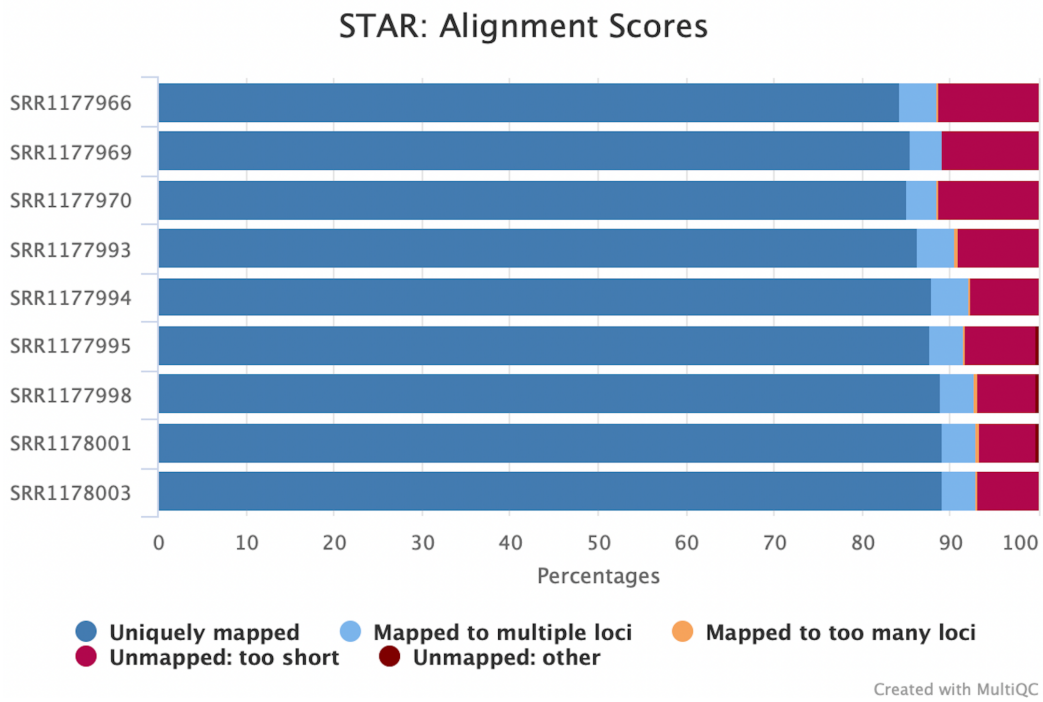

Figure 11: Figure displays the star alignment statistics as percentages

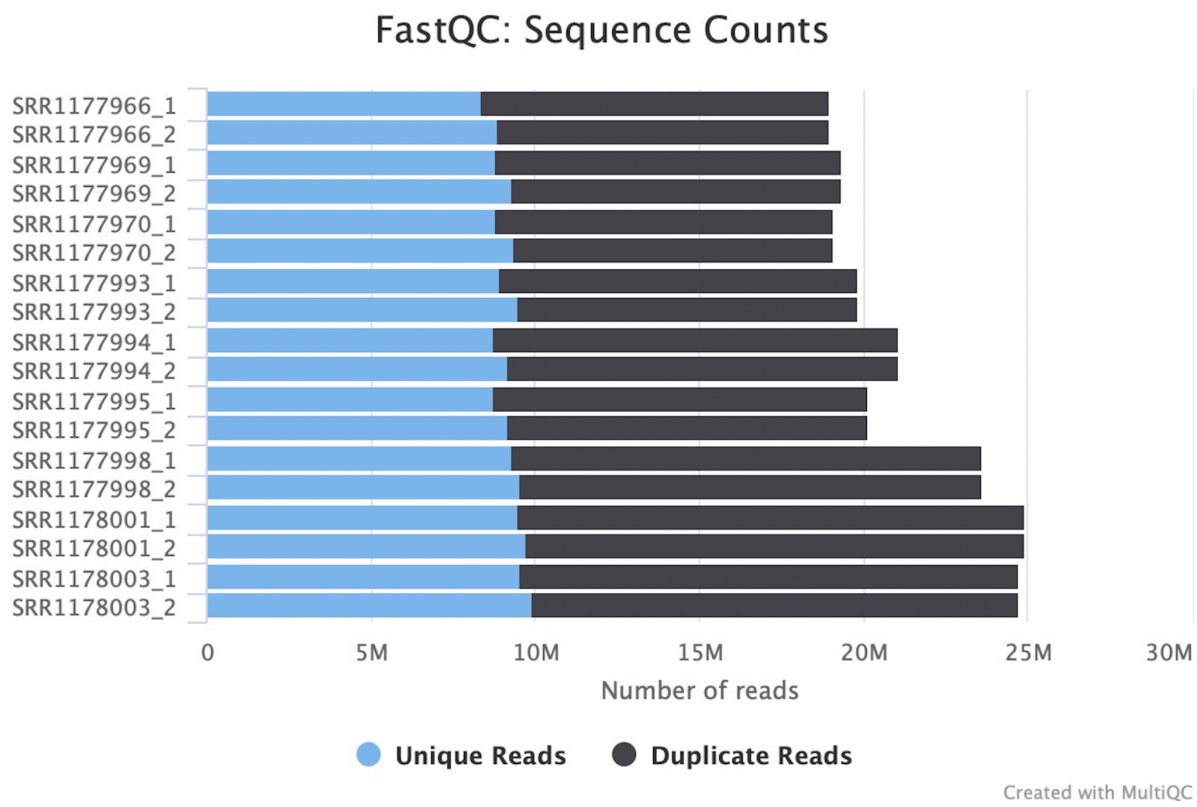

Figure 12: Sequence counts for each sample. Duplicate read counts are an estimate only

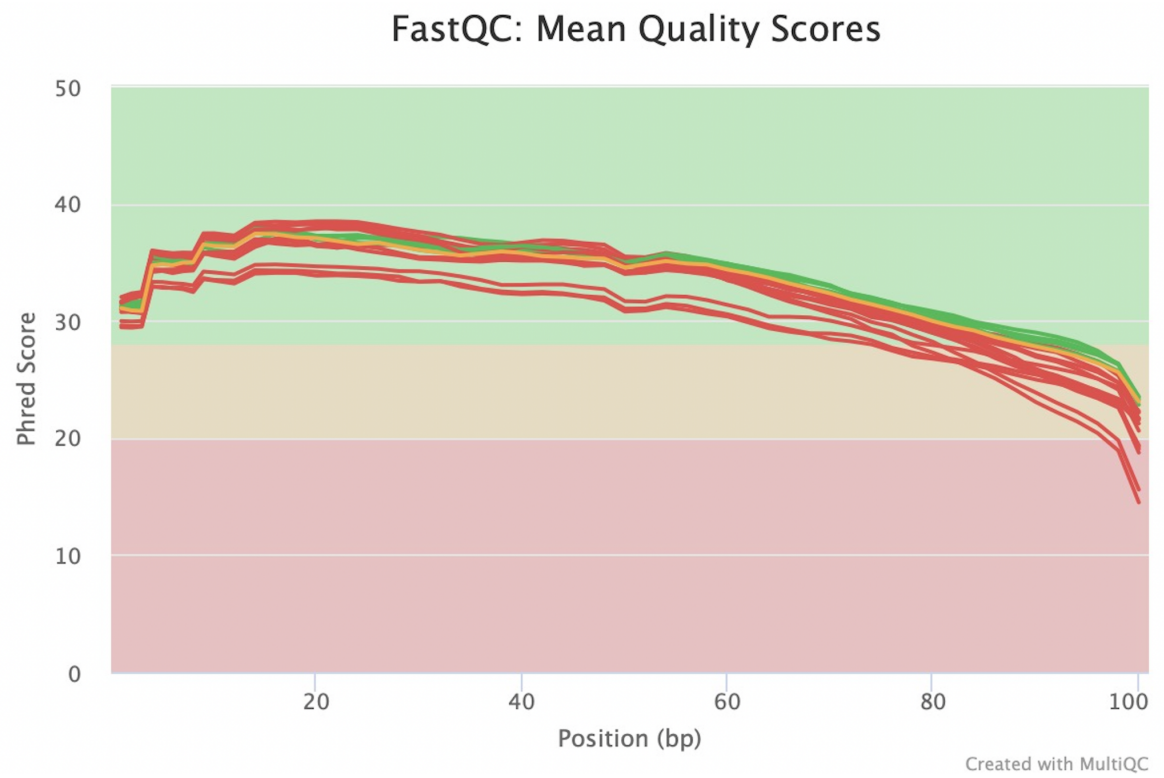

Figure 13: The mean quality value across each base position in the read for all the samples

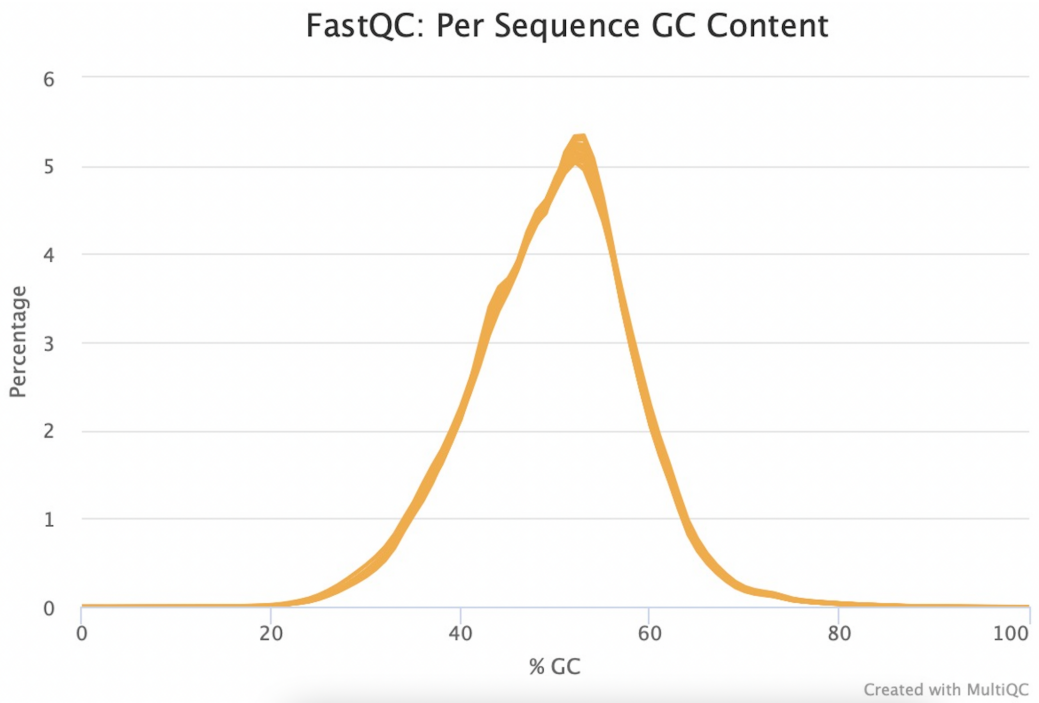

Figure 14: The average GC content of reads.

Discussion

The goal of this analysis was to align the sequences to the reference genome using STAR and assess the quality of the fastq files and the alignment of the sequences to the reference genome.

The alignment statistics and the fastqc statistics are overall acceptable. The quality of the reads in the fastq files are acceptable due to good mean Phred score, GC content for most of the reads in all the samples. Further the percentage of reads aligned using STAR for all samples was greater than 80% which is a good enough metric for downstream analysis. Therefore the alignment can be used for further downstream analysis.

References

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. doi:10.1016/j.cels.2016.08.011

Tang, X., Huang, Y., Lei, J. *et al.* The single-cell sequencing: new developments and medical applications. *Cell Biosci* 9, 53 (2019). https://doi.org/10.1186/s13578-019-0314-y

Wang C, Gong B, Bushel P, Thierry-Mieg J, Thierry-Miegg D, et al. A comprehensive study design reveals treatment- and transcript abundance–dependent concordance between RNA-seq and microarray data. Nat Biotechnol. 2014; 32(9):926–932.

Zhang, Y., Wang, D., Peng, M. *et al.* Single‑cell RNA sequencing in cancer research. *J Exp Clin Cancer Res* 40, 81 (2021). https://doi.org/10.1186/s13046-021-01874-1