UNDERSTANDING BIOINFORMATICS PIPELINES
Week 6 assignment report: Integrative approach
Group 2: Kelly J. Cardona | Alejandra López V

Propose and conduct an integrative approach to understand Monocyte and/or B cell Differentiation when using Single-cell RNA and Bulk RNA

You need to work on identifying data and question; maybe not all questions can be addressed, so set a question and the associated strategy.

1. DATA

Three datasets will be relevant for this proposal. GSE188819, GSE65146, and GSE109227, are divided as follows.

Single-cell RNA data: Four samples from the GSE188819 dataset divided into two groups. The control group includes two samples from healthy mice, serving as the baseline. The cerulein group consists of two samples from mice with cerulein-induced acute pancreatitis, representing the diseased state.

Bulk RNA data: GSE65146 (n=8) and GSE109227 (n=11) datasets from the GEO database will complement the single-cell data by offering a broader gene expression profile across the pancreas in mice models of induced acute pancreatitis. These datasets represent expression profiles from mice with cerulein-induced acute pancreatitis at different time points, allowing for the examination of dynamic changes in gene expression associated with the progression of the disease.

2. QUESTIONS

Through the integrative analysis of single-cell RNA and bulk RNA sequencing data, we can attempt to answer some questions aimed at understanding the monocyte-to-macrophage differentiation in acute pancreatitis, which are mentioned below. We decided to design a strategy to address the fourth question.

a. How does monocyte differentiation contribute to the progression of acute pancreatitis?
b. What are the key molecular markers of monocyte differentiation in acute pancreatitis?
c. Can integrated scRNA-Seq and Bulk RNA-Seq data reveal differences in monocyte differentiation between mild and severe cases of acute pancreatitis?
d. **What signaling pathways are predominantly involved in monocyte differentiation in acute pancreatitis?**

3. STRATEGY

Bulk RNA-seq analysis

The initial phase of our project is the bulk RNA-seq analysis, as shown in ***Figure 1.*** Since the data we will use is publicly available, our first step is the data **loading and preparation**, utilizing the GEOquery and data.table libraries in R for efficient handling of the raw counts***.*** Important to our analysis, we will obtain additional biological information, such as gene length and chromosome mapping, from resources like Ensembl Biomart. This step ensures that our dataset is not only comprehensive but also accurately annotated, facilitating further analysis.

Following data preparation, our **quality control** (QC) measures will involve using NOISEQ, a tool that will allow us to assess the quality of the RNA-seq data. Through QC, we want to identify and mitigate potential sources of technical bias, ensuring the integrity of our dataset. The incorporation of biological parameters, such as gene length and GC content, will facilitate the understanding of the dataset's quality, allowing us to proceed with confidence to subsequent analysis stages.

Normalization and **differential expression analysis** are the most relevant parts of our bulk RNA-seq pipeline. We have chosen DESeq2 for this purpose due to its robust statistical framework and flexibility in handling complex experimental designs. This phase will enable us to identify genes differentially expressed across different conditions within our dataset, such as comparing samples from mouse with acute pancreatitis against controls. This differential expression analysis will be a foundation for identifying potential biomarkers and therapeutic targets involved in monocyte differentiation.

Moving forward into the biological interpretation of our DEA to address our selected question, we apply **Over-Representation Analysis (ORA) and Gene Set Enrichment Analysis (GSEA)**, which use a ranked list of differentially expressed genes (DEGs) and comprehensive gene sets from databases like MSigDB and KEGG. The clusterProfiler, limma, and enrichplot packages in R become our tools of choice for these analyses, providing a statistical backbone to identify enriched pathways among our DEGs.

ORA necessitates a list of significantly differentially expressed genes and a background gene set representing all genes that could be differentially expressed in our experiments. We use topTable from the limma package to obtain this list and define our background set as genes with very low expression ($|Log2fc| < 0.05$). The enricher function from clusterProfiler is employed to conduct ORA, focusing on KEGG pathways to uncover biological processes enriched in our list of interest. Visualization of ORA results through dotplot and upsetplot offers insightful views into the pathways significantly represented in our dataset.

On the other hand, **GSEA** provides a more nuanced approach by considering the entire ranked list of genes to determine the statistical overrepresentation of gene sets at the top or bottom of this list. We conduct GSEA to compare the biological significance of pathways across different

states of monocyte differentiation, further ilustrated through gseaplot visualizations highlighting pathways with the most positive and negative Normalized Enrichment Scores (NES).

GeneSetCluster offers a unique perspective by clustering GSEA results, thereby facilitating the understanding of the pathway enrichments. This method allows us to investigate complex pathway interactions and pinpoint specific clusters of interest for further investigation. We utilize the GeneSetCluster package to perform this analysis. This helps us break down significant clusters for granular analysis of pathway interactions and their implications in monocyte differentiation during acute pancreatitis.

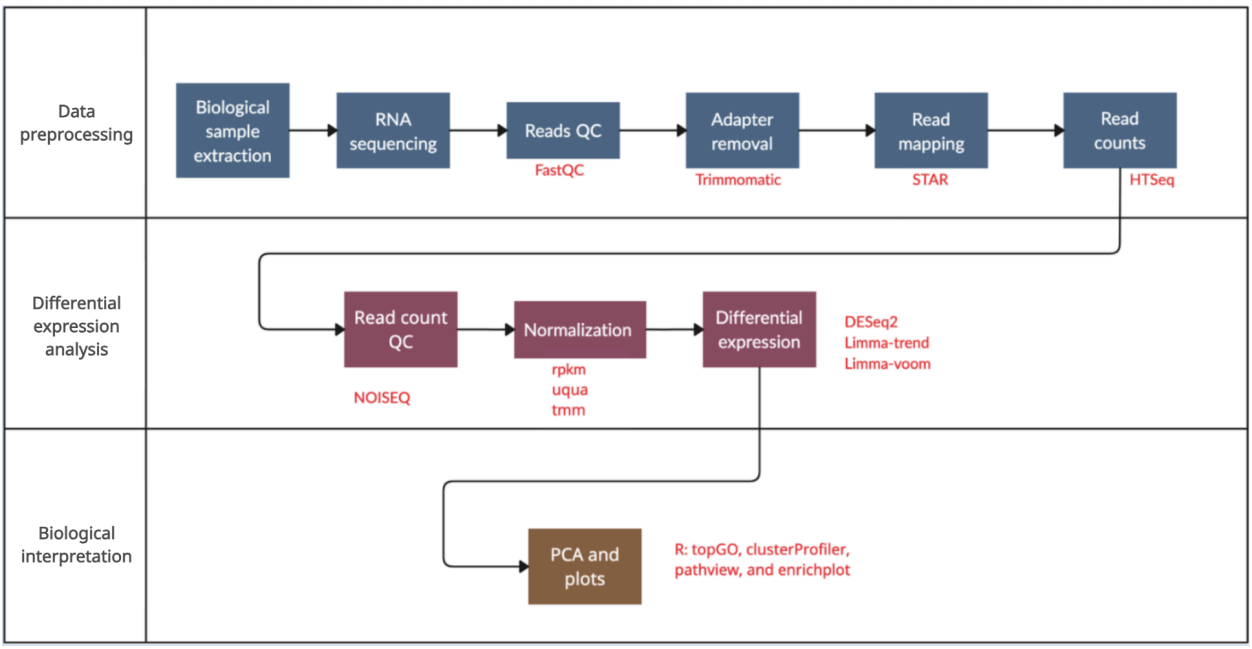

Figure 1. Bulk RNA-seq pipeline

Single cell RNA-seq analysis

Following our approach to understanding monocyte differentiation in acute pancreatitis through bulk RNA-seq analysis, we include single-cell RNA sequencing (scRNA-seq). This analysis will allow us to understand the heterogeneity within monocyte populations and point the signaling pathways at the single-cell level, offering insights that complement findings from bulk RNA-seq analysis.

The first step in our scRNA-seq analysis, as shown in ***Figure 2,*** is **quality control** (QC) to ensure the reliability of our data. Utilizing the Seurat package, we load our scRNA-seq dataset and apply QC metrics to filter out cells with an abnormal number of detected genes, high mitochondrial gene expression (>10%), or any other indicators of low-quality cells or doublets. This process is important for reducing technical noise and focusing on biologically meaningful

signals. The output of this step will be a Seurat object that we can use for downstream analysis.

Following QC, we align our reads to the mouse reference genome to get each cell's gene expression profile. This alignment step is important for identifying the genetic material present in each cell and attributing it to specific genes. Subsequently, we perform mapping QC to evaluate the efficiency of our alignment process, ensuring that a high proportion of reads map uniquely to the reference genome (>80%).

After the alignment, we conduct further **cell QC** to identify and exclude any remaining low-quality cells or potential doublets, refining our dataset for accurate analysis. We then proceed with normalization to correct for differences in sequencing depth and technical variations between cells, using Seurat's normalization methods.

With our normalized dataset, we can start the **differential expression analysis (DEA)** to identify genes that are differentially expressed across different cell populations within our dataset. The DEA results will highlight genes and pathways potentially involved in monocyte differentiation and the pathogenesis of acute pancreatitis.

Following DEA, we employ **clustering** techniques to group cells based on their gene expression profiles, identifying distinct cell populations and states within our dataset. Moreover, we will utilize heatmaps to visualize the expression patterns of key genes across different cell populations, especially focusing on those identified as differentially expressed in monocytes. We will also look into patterns and clusters of gene expression that signify distinct states of monocyte differentiation.

Volcano plots will allow us to visualize the results of differential expression analysis for each comparison of interest, such as monocytes in acute pancreatitis versus controls, so that we can identify genes that show significant differential expression with considerable log-fold changes, highlighting potential key players in monocyte differentiation and the inflammatory response characteristic of acute pancreatitis.

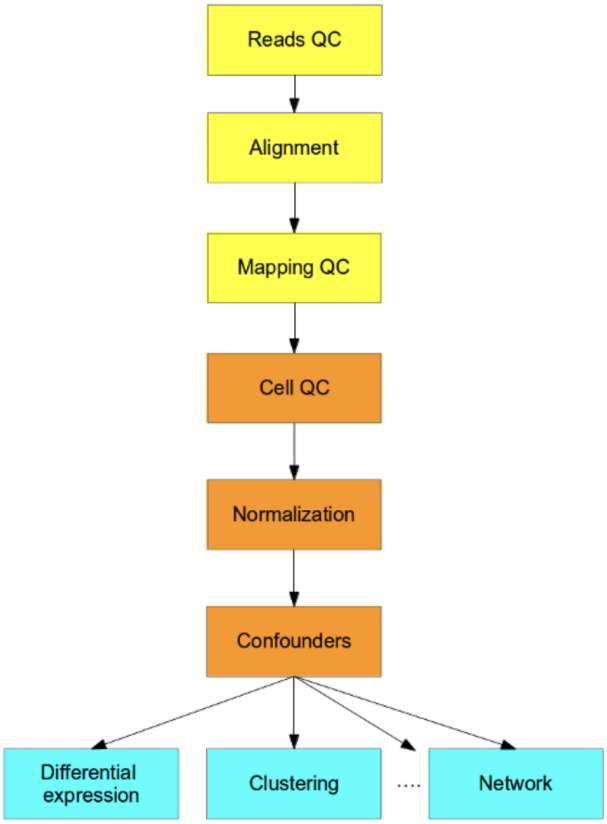

Figure 2. Single-cell RNA-seq pipeline

Integration of Bulk and Single cell RNA-seq data

Given the different resolutions of these datasets, bulk RNA-seq providing a panoramic view of gene expression across the entire tissue, and scRNA-seq offering a high-resolution insight into individual cell profiles, we will use advanced computational tools like Seurat's integration methods or LIGER to align and integrate these datasets. This harmonization will let us combine the broad expression profiles with the cell-specific expression patterns, helping us to understand the gene expression dynamics in monocyte differentiation.

With an integrated dataset at hand, we will identify signaling pathways that are conserved across both datasets as well as those that are uniquely highlighted in either bulk RNA-seq or scRNA-seq analyses. This step involves analyzing the integrated dataset for differential expression, followed by pathway analysis using tools like GSEA and ORA, focusing on the pathways involved in monocyte differentiation. By comparing and contrasting the pathways identified from each dataset, we will highlight the key signaling cascades driving monocyte differentiation in the context of acute pancreatitis.

Applying the high-resolution data from scRNA-seq, we will conduct cell trajectory analysis using Monocle2 to map the differentiation pathways of monocytes into their macrophage states. This analysis will be enriched with insights gained from the bulk RNA-seq data,

particularly regarding the temporal dynamics of gene expression changes associated with disease progression. The trajectory analysis will clarify the sequential activation of signaling pathways as monocytes differentiate and respond to the inflammatory environment.

The final step involves visualizing integrated analysis results, using heatmaps for gene expression patterns, volcano plots for highlighting significant genes, and trajectory plots for visualizing the differentiation pathways. These visualizations will not only illustrate the complex interplay between different genes and pathways but also facilitate the interpretation of how these molecular mechanisms contribute to monocyte differentiation in acute pancreatitis.

This integrative analysis strategy complements the strengths of bulk RNA-seq and scRNA-seq data, providing a multifaceted view of the molecular mechanisms behind monocyte differentiation in acute pancreatitis. By identifying and validating key signaling pathways involved in this process, our strategy can potentially discover novel insights into the pathogenesis of acute pancreatitis, guiding the development of targeted therapeutic interventions.

REFERENCES

Fang, Z., Li, J., Cao, F., & Li, F. (2022). Integration of scRNA-Seq and bulk RNA-Seq reveals molecular characterization of the immune microenvironment in acute pancreatitis. *Biomolecules*, *13*(1), 78.

Melendez E, Chondronasiou D, Mosteiro L, Martínez de Villarreal J et al. Natural killer cells act as an extrinsic barrier for in vivo reprogramming. Development 2022 Apr 15;149(8). PMID: 35420133

Norberg KJ, Nania S, Li X, Gao H et al. RCAN1 is a marker of oxidative stress, induced in acute pancreatitis. Pancreatology 2018 Oct;18(7):734-741. PMID: 30139658