Class: Computational Functional Genomics

Professor: Liana Lareau

Final Project

GitHub of the Project: https://github.com/cagintunc/single_cell_analysis

For the configuration of the project, please visit the GitHub repository above. This report serves as an example usage of the developed desktop application. So, it will only explain the biological meanings of the results. All technical information can be found on GitHub README.md file.

1. Introduction and Rationale

This analysis aims to uncover transcriptional dependencies between liver and skin tissues, focusing on how changes in skin gene expression could indicate liver-related diseases by using a newly developed desktop application for this project. By analyzing their single-cell data, it is possible to identify systemic markers detectable in skin tissue that reflect liver dysfunction, leveraging their shared roles in detoxification, metabolism, and immune responses. It can be used to diagnose liver-related diseases by using skin tissue as a tool. Moreover, finding shared pathways and markers could provide targets for therapeutic intervention.

2. Dataset and Methods

Dataset Description

The dataset used in this analysis was obtained from the *Tabula Muris* project (Consortium et al., 2018), a collaborative effort to create a comprehensive single-cell transcriptomic atlas of mouse tissues. Specifically, we utilized the *Single-cell RNA-seq data from Smart-seq2 sequencing of FACS sorted cells (v2)* dataset.

The database folder includes essential reference files, such as *annotations_facs.csv* and *metadata_FACS.csv*, as well as the *FACS* subfolder, which stores tissue-specific count matrices like *Liver-counts.csv*. These files serve as the primary input data for analysis.

The dataset includes 2,464 unique skin cells and 981 liver cells. Their corresponding FACS-sorted files capture gene expression profiles across 23,434 genes, enabling detailed analysis of tissue-specific transcriptional patterns.

Methods

First, liver and skin tissues were selected on the desktop application. In order to exclude anomalies while preserving the biological meaning, the middle 80% of cells were selected. Not all genes are expressed in every cell, and some genes are expressed at such low levels that they are prone to dropout events where their expression is undetected due to technical noise. With the help of the user-friendly interface, user can easily find the best cutoff value. In this test-case, cut-off value was selected as 10^2. So, genes expressed in fewer than 100 cells were removed. PCA, with 7 selected principal components, was used to identify the most significant components of variation in gene expression between tissues. By looking at the principal components graph (see. **Figure 2**), we can say that there are two distinct clusters which easily separate the two tissue types. Therefore, we did clustering with 2 clusters by using Spectral clustering algorithm. Spectral was decided to be used based on the application comparison between five different clustering algorithms. Spectral identified the clusters effectively, while other methods such as Agglomerative or KMeans clustering could not detect the clusters because of their distinct shapes. To identify genes that are significantly differentially expressed in the selected cluster compared to the rest of the dataset, we perform the t-test method. It is used to compare mean expression levels between cells in the selected cluster and cells outside it, the results are filtered to include only genes with positive log2 fold change (since positive results highlight genes that are specific to the selected cluster).

We selected specific genes representing key biological processes relevant to liver and skin tissues. Cyp1a2 and Fabp1 are liver-specific markers linked to detoxification and lipid metabolism, while Nqo1 and Hmox1 indicate oxidative stress responses common to both tissues. Il6 and Tnf are critical inflammatory cytokines. For skin, Krt14 and Flg highlight structural integrity, Sod1 and Gpx3 mark oxidative defense, and Cd207 and Cxcl12 reflect immune activity. Lastly, Col1a1 and Col3a1 represent extracellular matrix remodeling, bridging the shared fibrosis pathways in both tissues. These genes provide insights into tissue-specific and systemic processes.

3. Results and Key Figures

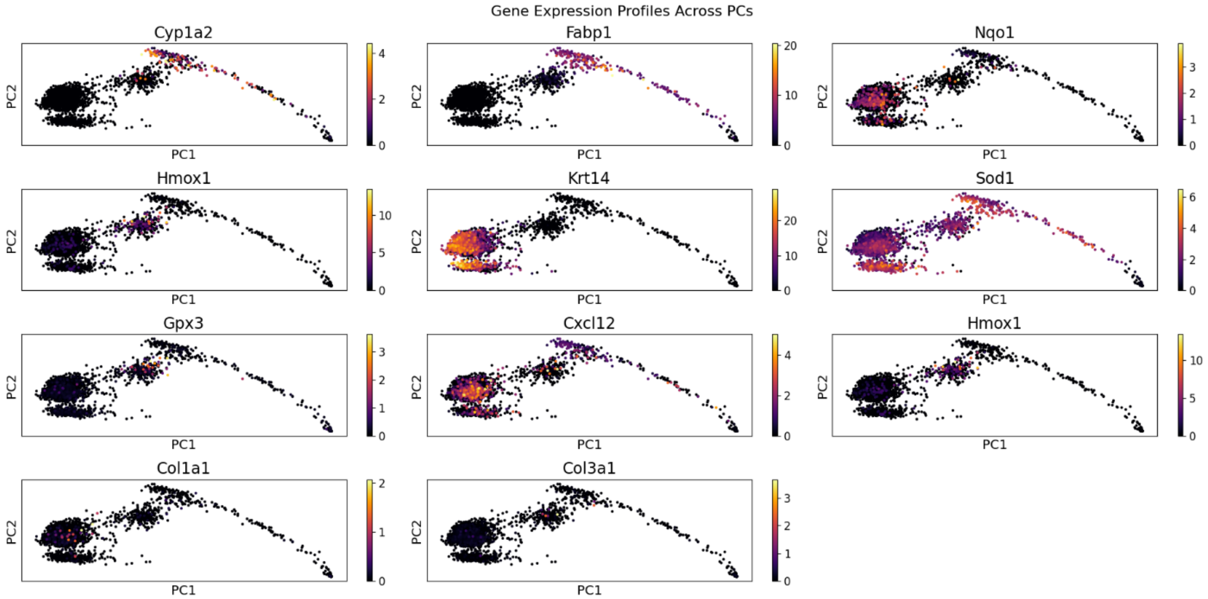

Figure 1: Distribution of specific gene expressions across dimensionally reduced data.

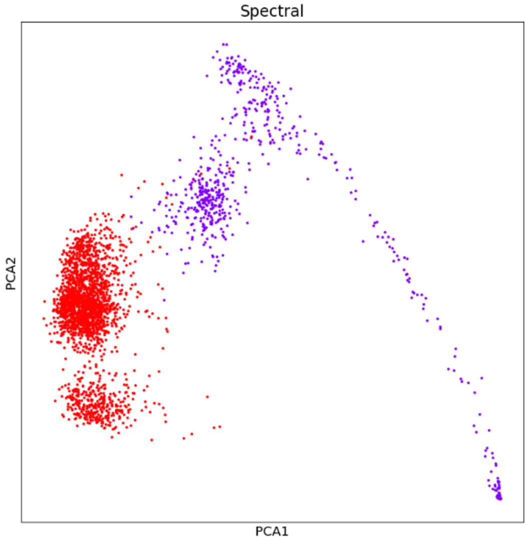

Figure 2: The most important two principal components found in PCA shows two distinct clustering formation. Therefore, the number of clusters was selected as two for the Spectral algorithm.

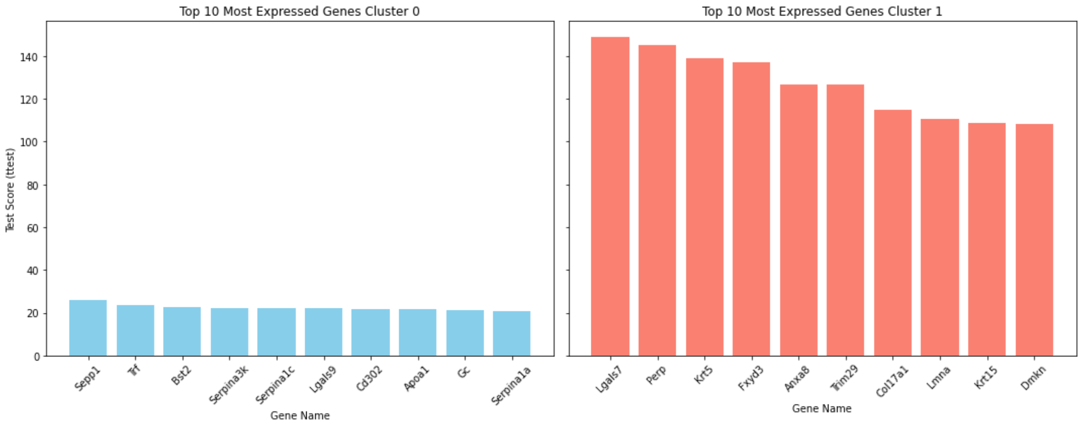

Figure 3: The top ten mostly expressed genes for each class show their expression profiles.

4. Discussion and Interpretation

Based on the top ten highly expressed genes in each cluster (see. **Figure 3**), cluster 0 likely represents liver tissue, while cluster 1 corresponds to skin tissue. In cluster 0, genes such as Sepp1 (Selenoprotein P), which is known for its role in antioxidant defense and selenium transport, and Serpina1c/Serpina3k, members of the serpin family involved in protease inhibition and regulation of inflammatory processes, are highly expressed. Research by Saito (2019) highlights the role of selenoprotein P in regulating pancreatic β cell function supporting that Sepp1 plays a systemic role beyond the liver, indicating the importance of liver-specific proteins in influencing other tissues. Additionally, Apoa1 and Apoc3, apolipoproteins critical for lipid metabolism, are hallmark genes associated with liver function (Lyu et al., 2022). These genes collectively suggest that cluster 0 is enriched in liver-specific metabolic and regulatory processes. On the other hand, cluster 1 exhibits high expression of Lgals7 (Galectin-7), a marker commonly associated with keratinocytes in epithelial tissues, and Krt5/Krt15 (Keratin genes), which are critical for maintaining skin structure and integrity. Genes such as Fxyd3 and Trim29, involved in epithelial differentiation and ion transport regulation, further support the identification of cluster 1 as skin tissue (Edqvist et al., 2015). The distinct gene expression profiles highlight the functional specialization of liver and skin tissues.

The differentiation between the clusters is evident from the principal components (see. **Figure 1**), which reveal distinct spatial distributions of gene expression. It highlights genes involved in shared pathways between liver and skin tissues, such as oxidative stress response and extracellular matrix remodeling. For example, *Hmox1 (Heme oxygenase 1)* and *Sod1 (Superoxide dismutase 1)*, key players in antioxidant defense, are expressed in both clusters, indicating a common response to oxidative stress. Skin is an easily accessible tissue, making it ideal for non-invasive diagnostic approaches. Biomarkers like Hmox1 and Sod1 could be indicators of systemic conditions originating from liver diseases.

5. Limitations and Future Directions

While using skin tissue as a diagnostic tool for liver diseases shows promise, challenges remain. Shared pathways, like oxidative stress, may complicate distinguishing liver conditions from other systemic issues, and environmental factors could introduce variability. Expanding datasets and incorporating machine learning could enhance predictive accuracy. Future efforts could focus on developing skin biopsy protocols and robust models to establish skin analysis as a reliable, non-invasive diagnostic tool.

6. References

- Consortium, Tabula Muris; Webber, James; Batson, Joshua; Pisco, Angela (2018). Single-cell RNA-seq data from Smart-seq2 sequencing of FACS sorted cells (v2). *Figshare*. Dataset. https://doi.org/10.6084/m9.figshare.5829687.v8
- Edqvist, P.D., Fagerberg, L., Hallström, B.M., Danielsson, A., Edlund, K., Uhlén, M., & Pontén, F. (2015). Expression of Human Skin-Specific Genes Defined by Transcriptomics and Antibody-Based Profiling. *Journal of Histochemistry & Cytochemistry*, 63, 129 - 141.
- Lyu, W., Xiang, Y., Wang, X., Li, J., Yang, C., Yang, H., & Xiao, Y. (2022). Differentially Expressed Hepatic Genes Revealed by Transcriptomics in Pigs with Different Liver Lipid Contents. *Oxidative Medicine and Cellular Longevity*, 2022.
- Saito, Y. (2019). Selenoprotein P as a significant regulator of pancreatic β cell function. *Journal of Biochemistry*.