C249 Final Project

1. Introduction and rationale:
   1. Both the most common primary tongue cancer and trachea cancer are squamous cell carcinoma [1][2]. However, the prevalence difference between these two primary cancers is significant, with primary trachea malignancy occurring at 0.142 per 100,000 people [3]. On the other hand, primary tongue cancer, also a squamous cell carcinoma, has a prevalence of 3.6 per 100,000 people. Given this large difference, I want to investigate whether there is a gene-related factor that explains the disparity in malignant transformation between these two cell types. The best way to explore the genetic basis of these differences is to collect the full genomic data of squamous cell carcinoma in both cell types. However, before diving into cancer cells, it is crucial to first identify the basic genetic differences in normal cells. By understanding these baseline differences, we can later subtract them from the differences observed in cancer cells, allowing us to identify the true genetic distinctions in cancer cells.
   2. I used the same dataset as the assignment [4]. The data comprises a compendium of single-cell transcriptomic data from the model organism *Mus musculus*, encompassing over 100,000 cells across 20 organs and tissues. The original study [5] employed two distinct methodologies: microfluidic droplet-based 3’-end counting for high-throughput sampling and fluorescence-activated cell sorting (FACS) for high-coverage characterization. It highlights previously under-characterized cell populations and provides open access to protocols, scripts, and an interactive data browser to facilitate reproducibility and further exploration. I extended my project using this dataset because it enables direct comparison of cell types across tissues and provides a foundational atlas for cellular and molecular biology.
   3. As I mentioned earlier, the significance of this project is to establish a baseline difference between tongue cells and trachea cells. This baseline will allow us, once we obtain squamous cell carcinoma (SCC) data from both cell types, to subtract the normal cell differences from the cancer cell data. This approach will help us identify the true distinctions between the two SCC cell types.
2. Datasets and Methods:
   1. This dataset contains gene-count tables for FACS-sorted cells sequenced with Smart-Seq2 from 20 organs of 7 mice. The cells are categorized by their tissue of origin. It also includes data for 53,760 cells, 44,879 of which passed a quality control (QC) threshold of at least 500 genes and 50,000 reads. Cell annotations, based on the Cell Ontology controlled vocabulary, are provided in a separate CSV file.
   2. My preprocessing steps: Instead of using the differential\_expression function of scprep, I implemented this function as the first step in my process. I aim to identify the most significant gene expression differences between tongue cells and trachea cells and also compare the results of differential\_expression with those of Principal Component Analysis (PCA). Before performing differential\_expression, I filtered cells by library size, removed lowly expressed genes, normalized the data, and transformed gene counts—steps I consider crucial regardless of the downstream processes to follow.
   3. C. My analytic methods: Unlike the approach taken in the assignment, I performed a PCA with a focus on identifying the specific components within each principal component (PC) using the SparseInputPCA function in scprep. I then conducted a comprehensive comparison of the genes identified through differential\_expression, PCA, and those found in literature reviews. I chose SparseInputPCA over the standard PCA used in the assignment because it provides the \_component attribute, which is essential for this project.
3. Result and Key Figures:
   1. Finding: From the basic PCA analysis, we can observe that tongue cells have a more compact distribution, whereas trachea cells exhibit a more diverse distribution (Figure 1). There are some slight differences between the results of PCA and those of the gene expression difference function, but both methods highlighted some of the most significant gene differences between tongue and trachea cells. In the previous assignment, it was evident that starting from PC4 and beyond, the ability to distinguish between the two cell types gradually diminishes. Therefore, instead of selecting only the first principal component (pca\_components.argmax[0]) from the first 10 principal components, we chose the top 6 components (np.argsort(np.abs(pca\_components))) from the first 3 principal components. The choice of 6 components was made for the convenience of generating plots. The genes with the most significant contributions are listed in Table 1.

Table 1. The genes with most contribution in different analysis methods.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Significance of contribution | PCA | | | Gene expression difference |
| pc1 | pc2 | pc3 |
| No. 1 | Gsn | Gsn | Cytl1 | Krt14 |
| No. 2 | Krt14 | Cd74 | Chad | Gsn |
| No. 3 | Mgp | H2-Ab1 | Mgp | Krt5 |
| No. 4 | Dcn | Dcn | Wif1 | Krt6a |
| No. 5 | Krt6a | H2-Aa | Sparc | Mgp |
| No. 6 | Krt5 | H2-Eb1 | Clu | Perp |

* 1. From Figure 1 we can see that the gene resulted from both PCA and differential\_expression function that scprep provided showed really differentiation ability in seperating tongue cells and trachea cells. These genes have great difference between tongue cells and trachea cells. Even within each cell types the epression also differes as shown in Figure 2, both the Cd74 and H2-Ab1 showed differences in trachea cells.

1. Discussion and interpretation
   1. The genes that contribute the most to differentiating trachea cells from tongue cells are Gsn (identified in PC1, PC2, and expression difference) and Krt14 and Krt5 (identified in PC1 and expression difference). The relationship was clearly showed in Figure 1 and Figure 3. These genes are also highlighted in literature reviews. In previous studies, including an investigation into the squamous–columnar junction—considered a hotspot for precancerous lesions—Krt14 was used as a marker for squamous cells [6, 7]. This indicates that our results align with findings from previous research. However, if we used ‘Krt14’, ‘Krt5’ and ‘Gsn’ as the component to create the 3D plot provided in scprep, the effectiveness wouldn’t be as good as using the original PC components. (Figure 4.)
   2. An interesting phenomenon is that, although the predominant epithelial type in the trachea is pseudostratified columnar epithelium, squamous epithelium can still be found in the trachea and other parts of the respiratory tract. This squamous epithelium often undergoes malignant transformation, eventually becoming the most common malignant tumor in these regions [1]. Several previous studies have shown that TP53 is a key gene involved in precancerous changes in both tracheal squamous cell carcinoma and oral (including tongue) squamous cell carcinoma. This finding inspired me to further investigate gene differences. However, this dataset does not include information on this particular gene in its columns.
2. Limitations and Future Directions
   1. This final project aims to identify the normal differences between tongue cells and trachea cells so that, in the future, when single-cell data of SCC from both regions becomes available, we can determine the true differences in cancer cell changes. However, achieving this goal requires quantitative information rather than just qualitative data, which represents the biggest limitation of this project.
   2. In the future, if we could obtain more than just single-cell data—such as an abundant amount of full genomic data from cancer cells—it would enable a much more powerful analysis of cancer gene differences. Additionally, for investigating precancerous changes, information on TP53 would be crucial.

Figure 1. Using PCA and UMAP to show the gene expression difference. Genes were picked from both differential\_expression and SparseInputPCA functions provided in scprep.

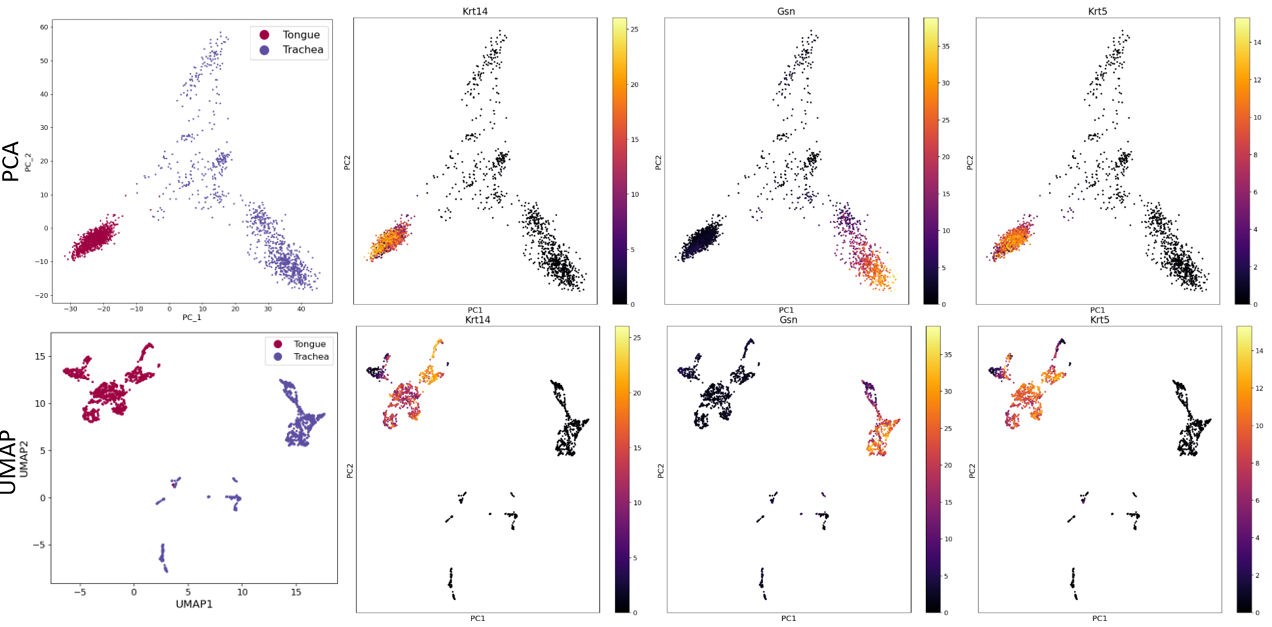


Figure 2. Using PC\_2 picked gene showed that gene expresssion not only showd in different cell types but within the same cell type as well. Cd74 in the middle showed expression difference within trachea cells.

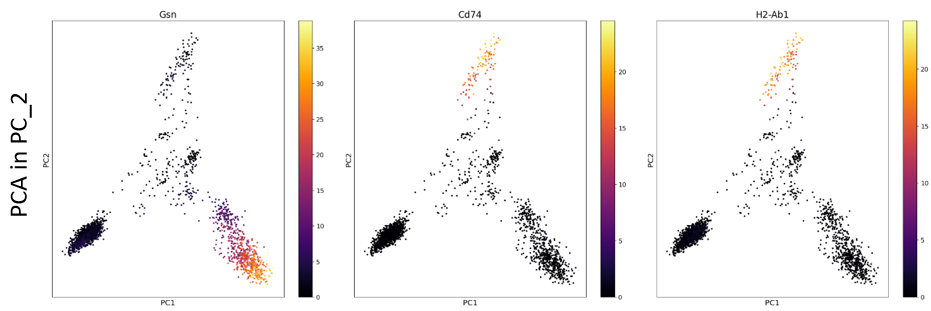


Figure 3. Jitter plot of the selected genes which has the most contribution.

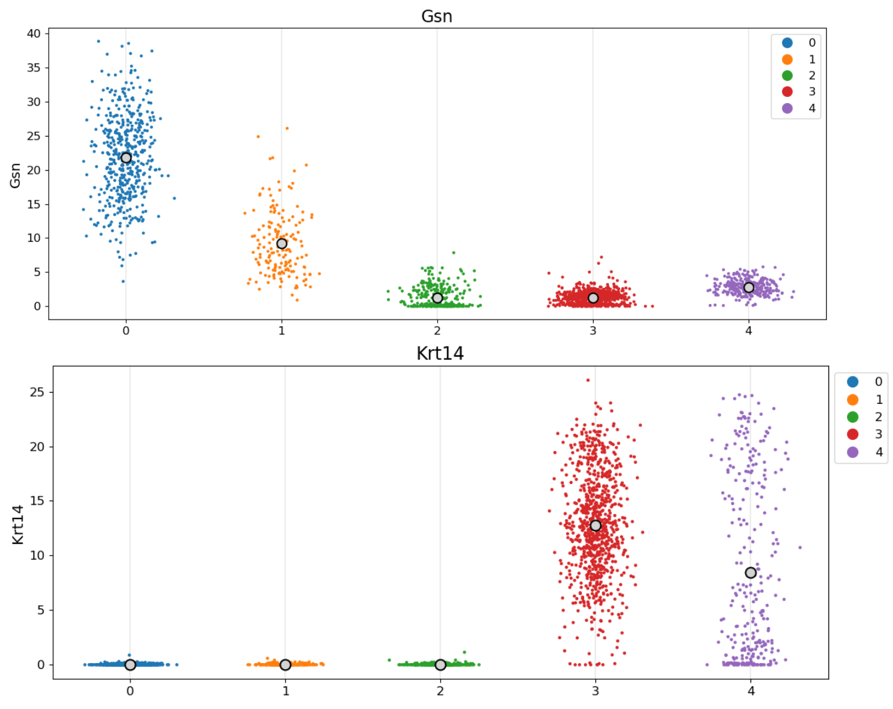
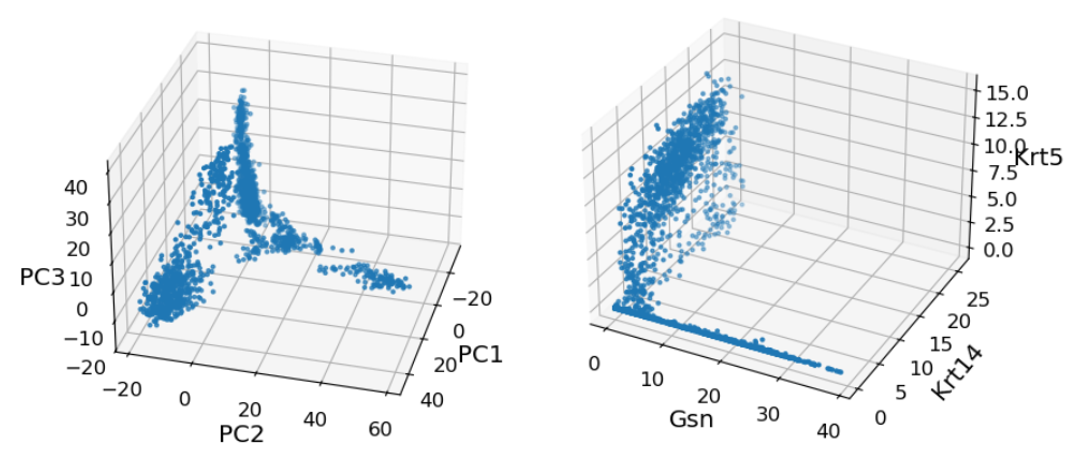


Figure 4. Plotting 3D plots with PC components created by PCA (left) and picked genes from a single PC component.



Reference

[1] Li W, Ellerbroek NA, Libshitz HI. Primary malignant tumors of the trachea. A radiologic and clinical study. Cancer. 1990 Sep 1;66(5):894-9. doi: 10.1002/1097-0142(19900901)66:5<894::aid-cncr2820660514>3.0.co;2-w. PMID: 2386916.

[2] Fatima J, Fatima E, Mehmood F, Ishtiaq I, Khan MA, Khurshid HMS, Kashif M. Comprehensive Analysis of Oral Squamous Cell Carcinomas: Clinical, Epidemiological, and Histopathological Insights With a Focus on Prognostic Factors and Survival Time. Cureus. 2024 Feb 18;16(2):e54394. doi: 10.7759/cureus.54394. PMID: 38505442; PMCID: PMC10949903.

[3] https://seer.cancer.gov/statfacts/html/tongue.html

[4] 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

[5] The Tabula Muris Consortium., Overall coordination., Logistical coordination. et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature 562, 367–372 (2018).

[6] Sankoda, N., Tanabe, W., Tanaka, A. et al. Epithelial expression of Gata4 and Sox2 regulates specification of the squamous–columnar junction via MAPK/ERK signaling in mice. Nat Commun 12, 560 (2021).

[7] https://www.proteinatlas.org/humanproteome/single+cell/single+cell+type/squamous+epithelial+cells