**UNDERSTANDING BIOINFORMATICS PIPELINES**
**Week 5 assignment report: Single-cell Multiome**
**Group 2: Kelly J. Cardona | Alejandra López V | Giulia Sansone**

## I. ASSIGNMENT COMMON TO ALL

a. **Why is transcriptomics used as a cornerstone to the different multiome technologies?** Transcriptomics offers a broad range of insights into diverse biological questions. Firstly, it provides a dynamic view of gene expression, capturing the active part of the genome that is being expressed at any given time and condition. This, in turn, helps us understand the changes in protein production and how they ultimately affect cellular functions. Moreover, transcriptomics can be applied universally across different cell types and conditions, and it bridges the gap between the static information contained within the genome (genomics) and the functional proteins.

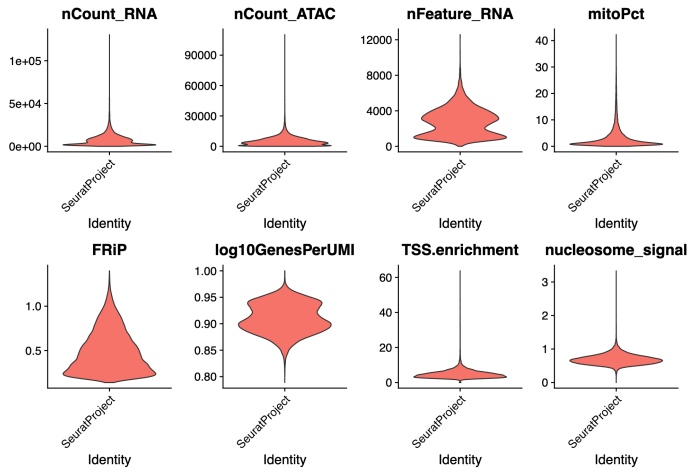
b. **Does focusing on single cells within the context of multi-omic data provide a complete picture of biological processes, or are there important aspects that this approach might miss?** Focusing on single cells within the context of multiomic data is a powerful approach that offers high-resolution insights into the heterogeneity and complexity of biological systems. This can lead to a more precise understanding of developmental processes, immune responses, and disease mechanisms. However, this approach may only provide a partial picture of some biological processes. For example, it might miss interactions between cells (cell-cell communication), the role of the extracellular matrix, and spatial context within tissues. Additionally, some omics layers, like metabolomics, are technically challenging to measure at the single-cell level due to current technological limitations, potentially leaving out important aspects of cellular function and regulation

c. **How might the future of multiomic single-cell technology contribute to our understanding of complex biological phenomena?** As these technologies continue to evolve, they are expected to become more sensitive, comprehensive, and integrated, allowing for the simultaneous analysis of multiple layers of biological information from single cells. This will enable a more complete view of cellular function, revealing not only the presence of specific biomolecules but also their interactions and dynamics within the cell. Advances in computational and bioinformatics tools could also play a role in integrating and interpreting this complex data, uncovering new biological insights and mechanisms. Furthermore, multiomic single-cell approaches can significantly impact personalized medicine, allowing for more precise disease diagnosis, prognosis, and tailored therapeutic strategies.

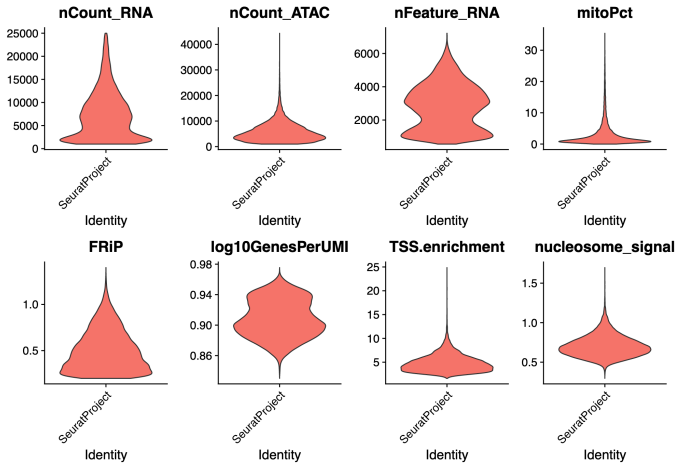
## II. ASSIGNMENT GROUP 2

### a. Annotate the cells using the differentially expressed genes in GEX, perform a similar analysis with the ATAC assay and compare the results.

The first fundamental step of this pipeline is **the quality control analysis**. This analysis is essential for ensuring the quality and reliability of the single-cell data before further analysis. QC metrics like the number of detected molecules and genes can help identify and exclude low-quality cells or nuclei. Metrics such as the percentage of mitochondrial genes are used to detect dying cells or cells with damaged membranes, which are often excluded from downstream analysis. The FRiP score, TSS enrichment, and nucleosome signal are crucial for evaluating the quality of ATAC-seq data, indicating the efficiency of chromatin accessibility mapping and the enrichment of signal at transcription start sites and nucleosome positioning, respectively. Is important to perform QC at the beginning of the analysis to understand the general status of our datasets **(1)**, and then perform some filtering to ensure the exclusion of low-quality cells **(2)**. This allows us to adjust the data quality as much as possible to guarantee the success of the downstream analysis. For instance, from Figure **(2),** we can observe that we can still adjust the filtering on the number of ATAC peaks (nCount_ATAC) and the percentage of mitochondrial genes (mitoPct) since from the violin plot, we can observe there are still many genes with outlier values.

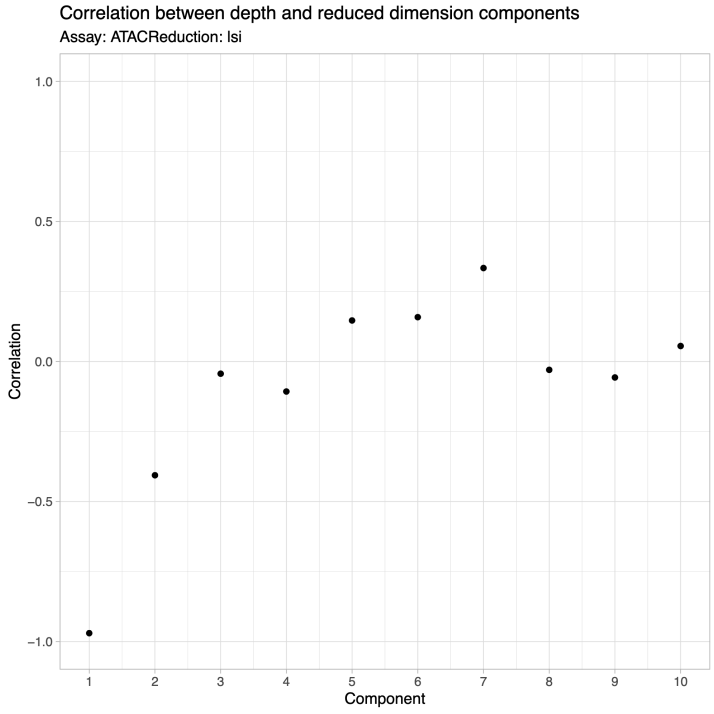



**(1)**



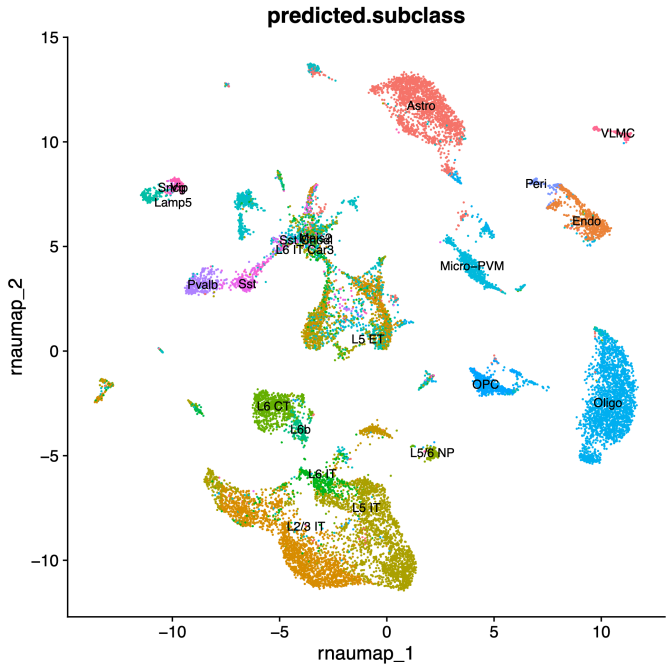

(2)

**The correlation analysis** (3) is important for understanding the influence of sequencing depth on the extracted features or dimensions after dimensionality reduction, which is critical in single-cell ATAC-seq data analysis. A high correlation between sequencing depth and certain components might indicate that those components capture variance primarily due to technical factors (like sequencing depth) rather than biological variance. Identifying and possibly correcting for such biases are crucial steps to ensure that downstream analyses, such as clustering and differential accessibility analysis, reflect actual biological differences rather than technical artifacts.



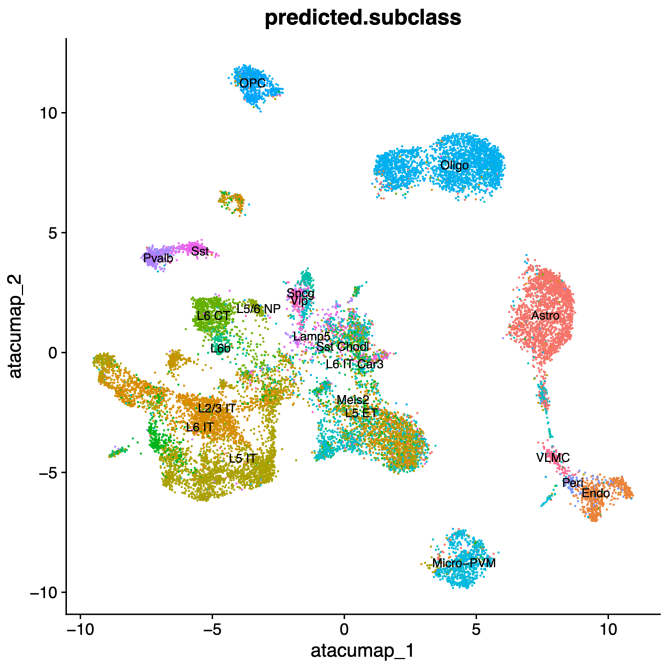

(3)

Then, utilizing **UMAP for visualizing RNA-seq data** is a powerful approach to uncovering the structure and diversity of cell populations in complex tissues like the brain. By applying this technique, we can identify distinct cellular subpopulations, facilitating a deeper understanding of their functional roles and interactions within the tissue context. The prediction of subclasses allows us to annotate clusters with specific cell types or states based on their gene expression profiles. This step is crucial for dissecting the cellular complexity of the mouse brain and for linking particular cell types or states to physiological functions. The generation of the UMAP-RNA plot **(4)** implies that the analysis successfully identified and visualized many distinct cell populations within the mouse brain based on their transcriptomic profiles.



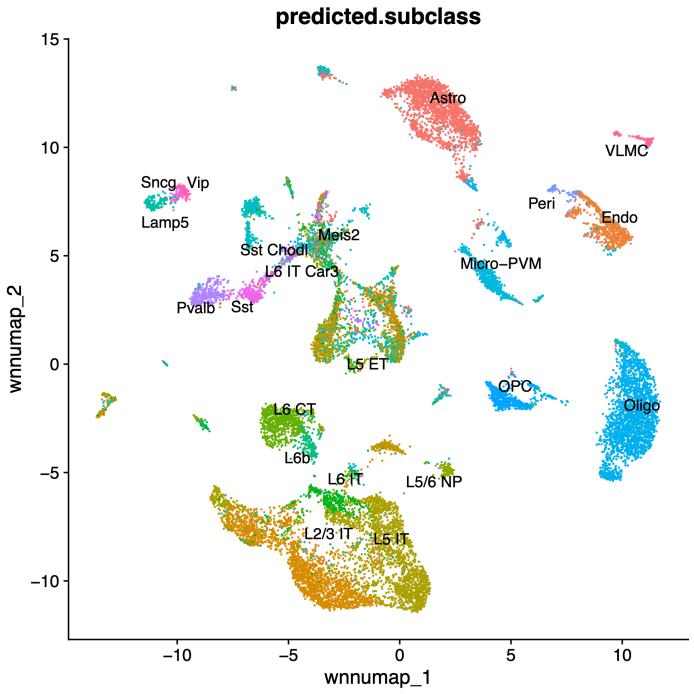

**(4)**

In addition, **UMAP for visualizing ATAC-seq** is important for identifying distinct regulatory landscapes that correspond to different cell types or states, highlighting the heterogeneity at the epigenomic level. Identifying these subclasses through ATAC-seq data **(5)** complements RNA-seq analyses by providing insights into the regulatory mechanisms driving gene expression differences. This can uncover potential transcription factor binding sites active in different cell types, contributing to our understanding of the gene regulatory networks in the brain.



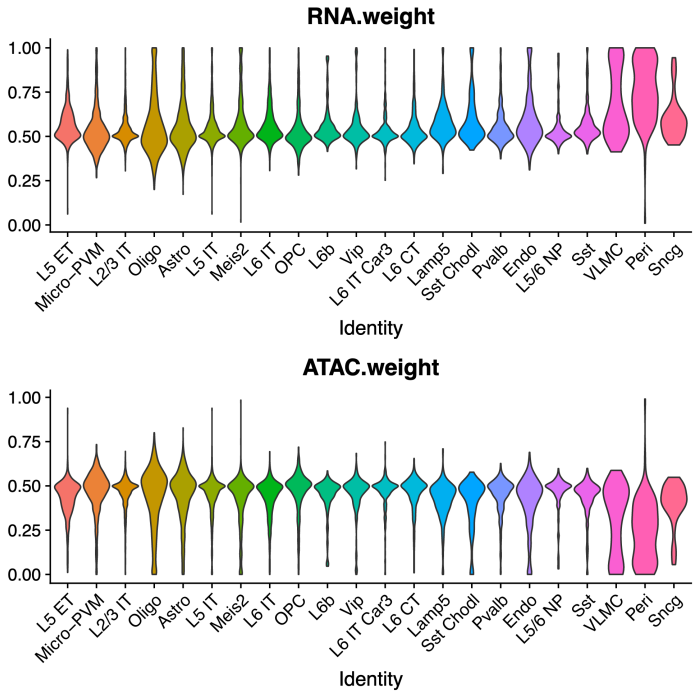

**(5)**

Then, we performed the integration of single-cell data from the mouse brain, combining RNA and ATAC-seq datasets to illustrate cellular subclasses, **the integrated UMAP plot (6)** reveals the spatial distribution and relationship between different predicted subclasses, including various neuron types and glial cells, based on both gene expression and chromatin accessibility. This dual-omic approach enhances the resolution and accuracy of cell type identification, enabling a more comprehensive understanding of the cellular landscape and the regulatory networks underlying cellular diversity and function in the mouse brain.




(6)

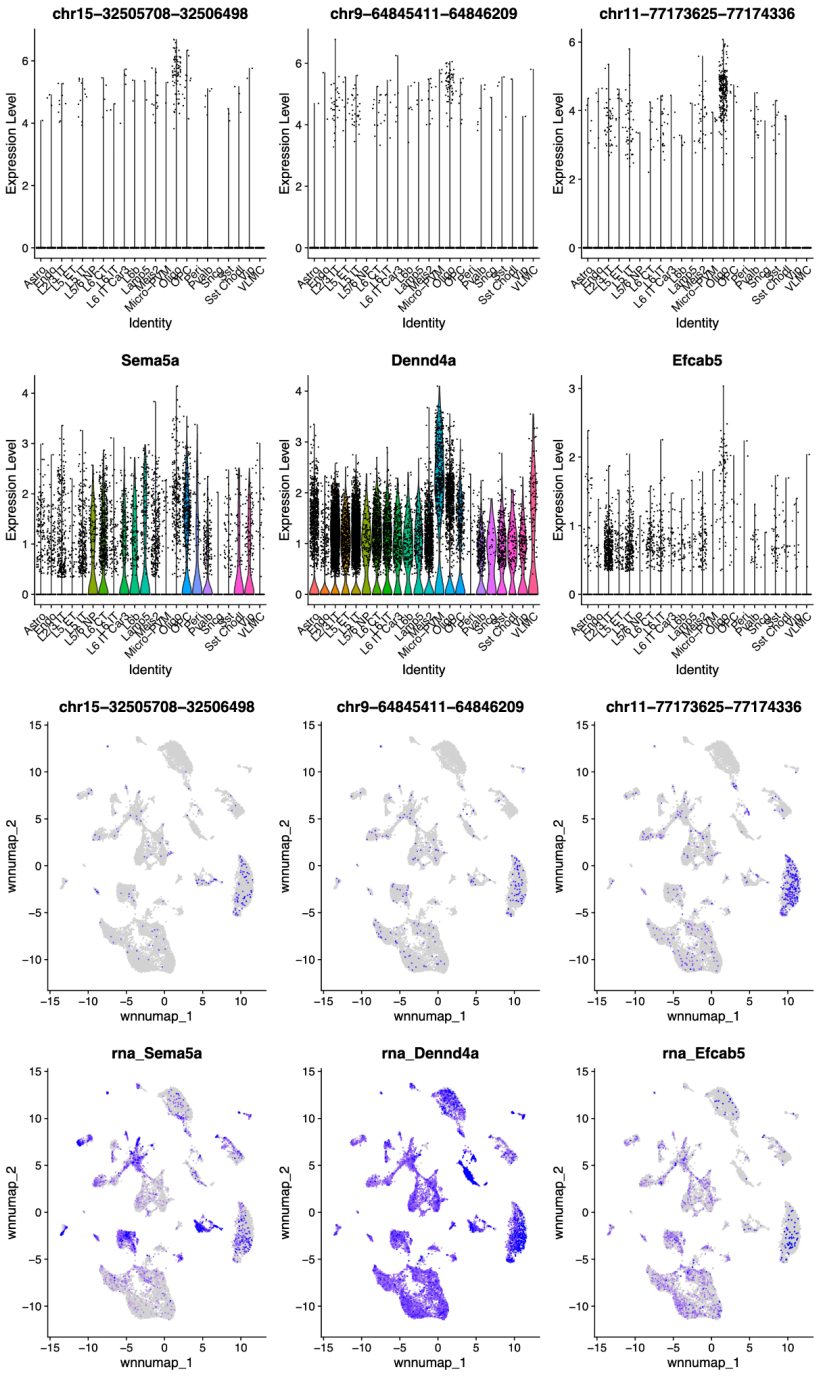
Now, it is important to consider that the previous integrated UMAP derives from the **integration weights calculated on RNA and ATAC-seq** data of mouse brain cells (7). The weights plot shows how each dataset (RNA and ATAC) contributes to the identification and visualization of cellular subclasses in the integrated analysis. This approach allows for a balanced consideration of gene expression and chromatin accessibility data, ensuring that both molecular layers are appropriately represented in the study. The integration weights reveal the relative influence of each omic layer (RNA and ATAC) in defining the cellular landscape, enhancing the interpretability and robustness of the identified cell subclasses. From this, we can see that some cell types (specifically, the last three) have higher median weights for RNA, suggesting a stronger contribution of gene expression data to their identification. In contrast, others might rely more heavily on ATAC-seq data, indicating the regulatory landscape is more informative.




**(7)**

**The differential analysis** across different cell types or subclasses in the mouse brain (in this example Oligodendrocytes and Astrocytes) is represented in the figure shown below **(8)**. The top panel displays the distribution of accessibility peaks in the top three differentially expressed genes in the contrast oligodendrocytes - astrocytes, following there is a panel showing the distribution of the expression levels for the top DE genes in each cell type cluster identified on the multiome assay. Below are UMAP plots that represent the distribution of ATAC peaks and the gene expression of these top genes across the integrated UMAP space, highlighting which cell subclasses express these genes the most or in which cell types these genes are more accessible by the chromatin machinery.

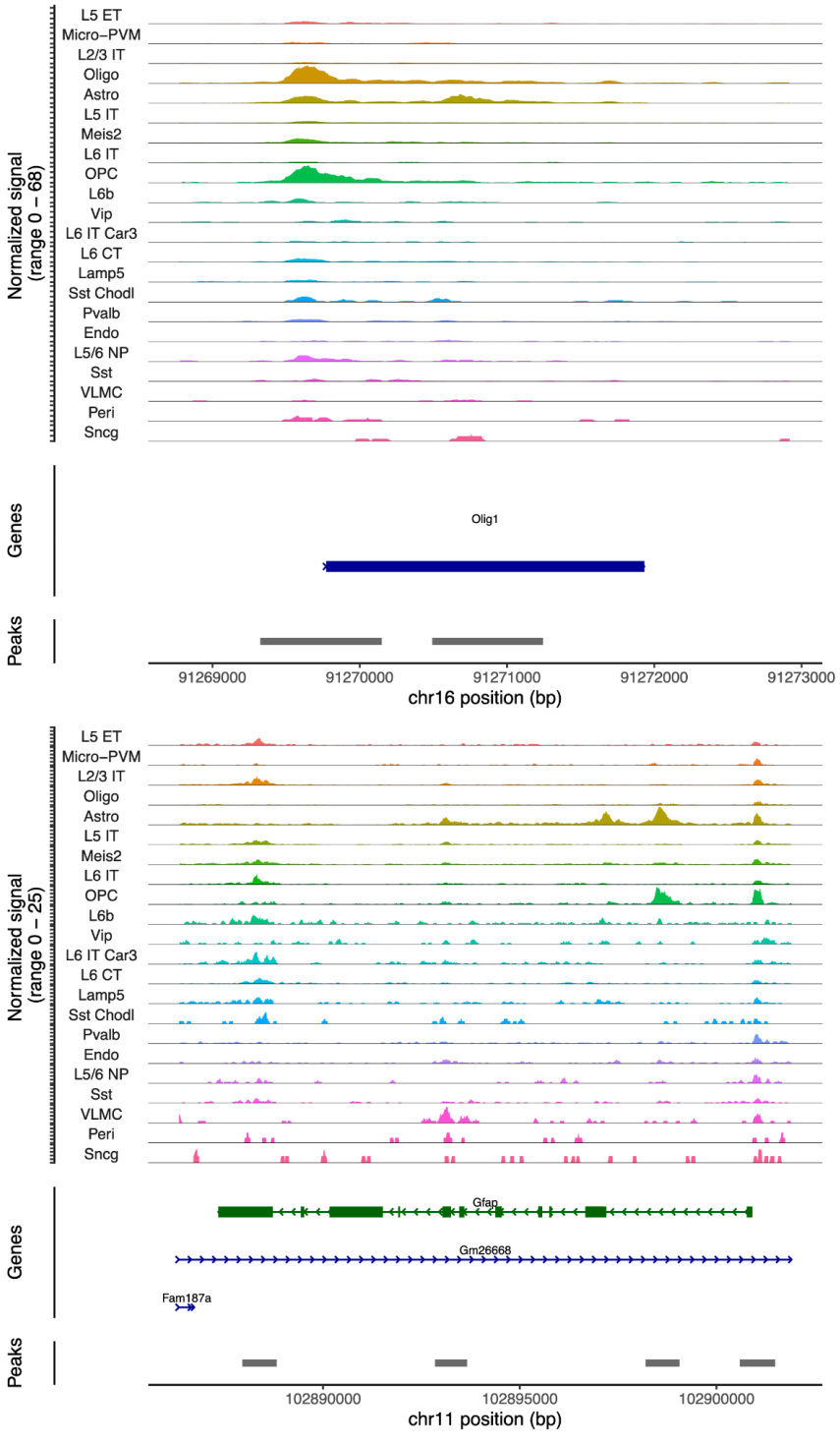
This figure is essential because it helps to identify which genes are specifically upregulated in different cell types, giving insights into each cell type's functional specialization. It's a relevant step in characterizing the cellular diversity within the mouse brain, linking gene expression to cell identity and function. Moreover, these plots are essential for the identification of biomarkers to discriminate between different cell types and establish correlations with the accessibility. In a real life experiment, this plot should be generated on cell markers to discriminate between cell populations and verify the effect of epigenetic regulator (ATAC data) on the expression of cell markers,




(8)

Another informative plot for our analysis is the coverage plot (9) that is displaying normalized chromatin accessibility signals across selected genomic regions. These plots are critical for comparing chromatin openness in various cell types from the mouse brain, indicating active regulatory regions like enhancers or promoters. By examining these accessibility patterns, we can infer potential cell-type-specific regulatory mechanisms, which are instrumental for understanding the diverse cellular functions in the brain. Also is a second step verification to corroborate the correlation between gene expression and chromatin accessibility initially

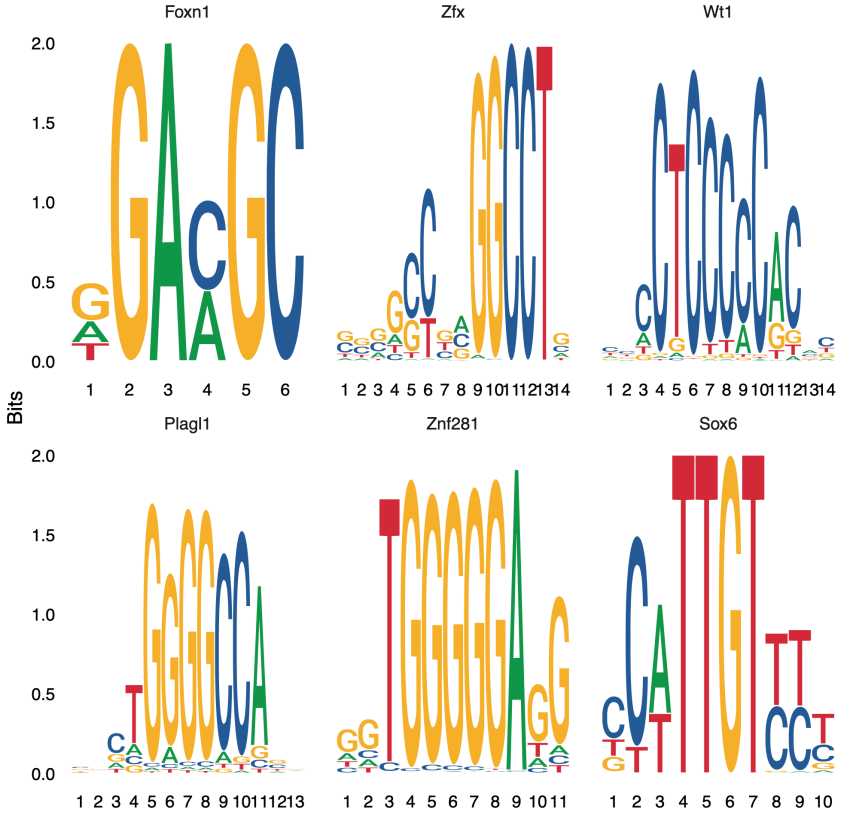
explored on the figure **(8)**, so is useful for revealing a deeper layer of cellular identity and specialization.




**(9)**

Finally, another essential step in this pipeline is **the binding motif enrichment analysis**; a process crucial for deciphering which transcription factors might be actively involved in cell-type-specific gene regulation. By identifying motifs that are significantly enriched in the accessible regions of chromatin, we can have insights into the regulatory networks shaping the identity and function of various cell types in the mouse brain. Understanding these motifs' roles can lead to a better understanding of the molecular regulators of cellular diversity and the specific genetic pathways involved in brain function and development in mice.

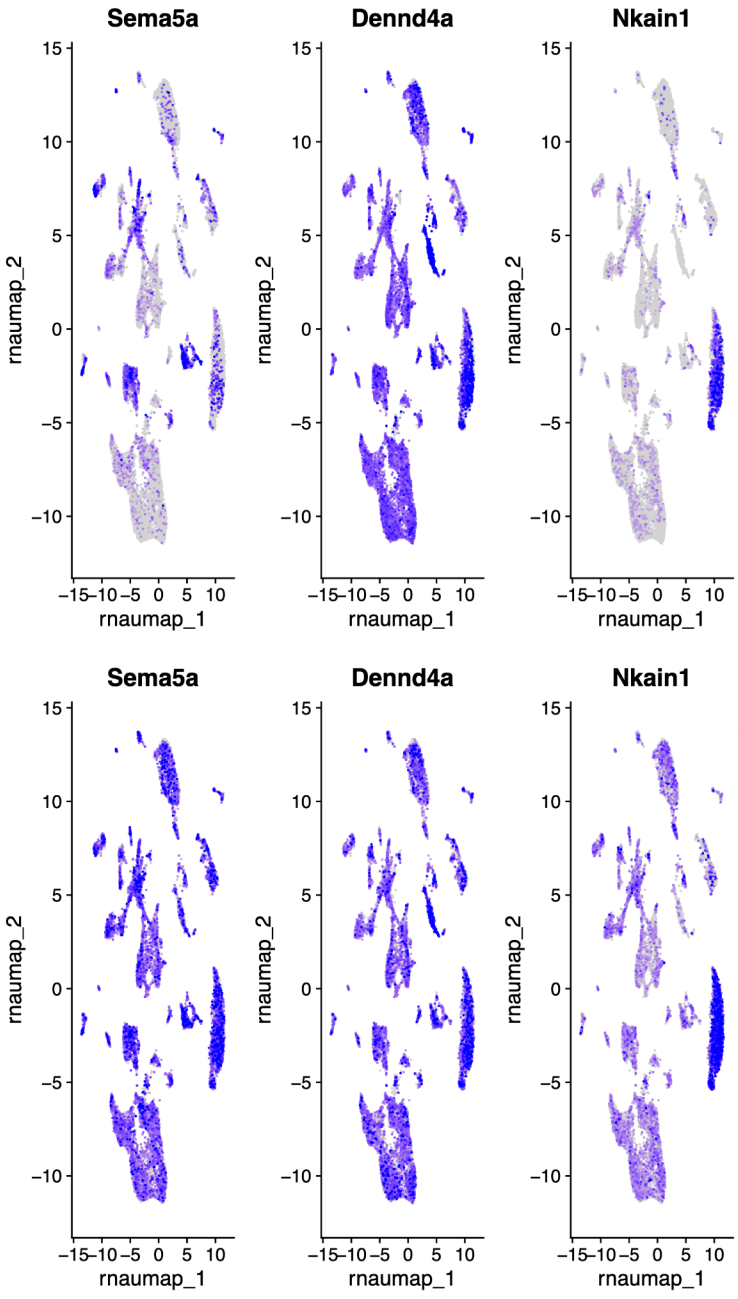
Figure **(10)** displays the motif logo derived from the binding motif enrichment analysis where binding motifs for specific transcription factor (Foxn1, Wt1, Plagl1, Znf281, and Sox6) are highlighted. The enrichment scores, displayed in bits, show a range from 0 to 2, with higher scores indicating stronger enrichment and therefore highly probability of binding/regulation from the specific motif.




**(10)**

Finally, we obtained the last UMAP plot from the RNA activity matrix on the ATACseq analysis; the RNA activity matrix annotates the open chromatin accessible regions detected with ATACseq over the genes; in this way, the genes that are differentially overexpressed should be the same genes with open chromatin, therefore accessible for transcription, the UMAP plots bellow **(11)** represent the accessibility of three specific genes (Sema5a, Dennd4a, and Nkain1) across the clusters of cell types in the mouse brain, previously identified from the single-cell

RNAseq data. So it would be expected that these genes with open chromatin where also overexpressed in the corresponding cluster. However, this correlation is not always observable, as in this case. This type of analysis enables the identification of gene expression and accessibility signatures that are unique to specific cell types or states, providing insights into their functional roles within the brain's cellular ecosystem.




(11)

In conclusion, the comprehensive multi-omic analysis we conducted has successfully annotated various cell types in the mouse brain (oligodendrocytes and astrocytes) using differentially expressed genes from gene expression (GEX) data. Furthermore, a parallel examination using ATAC-seq data has allowed for comparing transcriptional and regulatory

landscapes, highlighting the different contributions of these layers to cell identity. This approach improves the resolution of cellular characterization and provides a robust framework for future investigations into molecular biology through single-cell multiomics.

**b. Can we use the same marker genes? Are they equally specific? Why?**

Based on the analyses conducted for GEX and ATAC assays, the specificity and applicability of the same marker genes across these assays should be evaluated with caution. While specific genes may serve as markers in the GEX data, indicating specific cell types or states based on their expression levels, their direct correlation with ATAC data, which measures chromatin accessibility, might not always be straightforward. The specificity of markers in one assay does not guarantee their relevance in another due to differences in the underlying biological information each assay provides. Therefore, while some markers may overlap in their ability to identify cell types across both assays, a detailed comparison to assess their equivalence and specificity is essential to understand the complementary information each dataset offers about cell identity and function.

## III. REFERENCES


- Vandereyken, K., Sifrim, A., Thienpont, B. et al. Methods and applications for single-cell and spatial multi-omics. Nat Rev Genet 24, 494–515 (2023). https://doi.org/10.1038/s41576-023-00580-2
- Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 24, 550–572 (2023). https://doi.org/10.1038/s41576-023-00586-w
- Miao, Z., Humphreys, B.D., McMahon, A.P. et al. Multi-omics integration in the age of million single-cell data. Nat Rev Nephrol 17, 710–724 (2021). https://doi.org/10.1038/s41581-021-00463-x