Name: Abdelsalam Helala

ID: 7056985

Name: Ahmed Lamloum
ID: 7003029

---

1. Preprocessing and Quality Control

1.1 Set Up the Environment

We begin by configuring the environment, loading the required R libraries, and specifying the reference genome for analysis.

1.2 Read the Data and Apply Filtering

We read the fragment files into **Arrow files**, applying lenient thresholds:

- **Fragments:** > 500
- **TSS Enrichment:** > 4
  A tiling matrix with a window size of **1000 bp** is computed to prepare for downstream analyses.

1.3 Identify and Remove Doublets

Doublets were identified using the `addDoubletScores` function. The distribution of doublet scores is visualized in the histogram below.

Doublet Score Distribution

The histogram shows that the majority of cells have low doublet scores, indicating they are singlets. A small subset of cells with higher scores are classified as doublets.

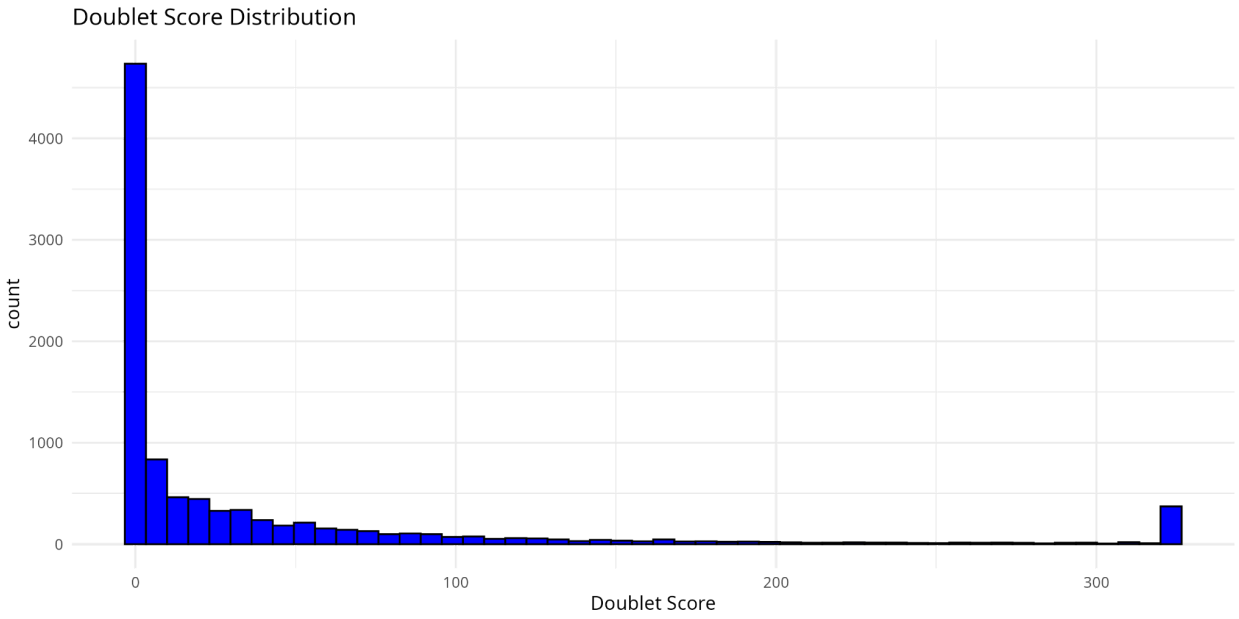
Doublet Removal

We remove the identified doublets and summarize the results per sample.

Table of Doublets Removed:

Sample	Doublets Filtered	Percentage
dc2r2_r1	139	3.7%
dc1r3_r1	119	3.4%
dc2r2_r2	69	2.6%

In total, **327 doublets** were removed from the dataset, ensuring cleaner and higher-quality data for subsequent analysis.



1.4 Collect All Samples into a Joint Data Structure

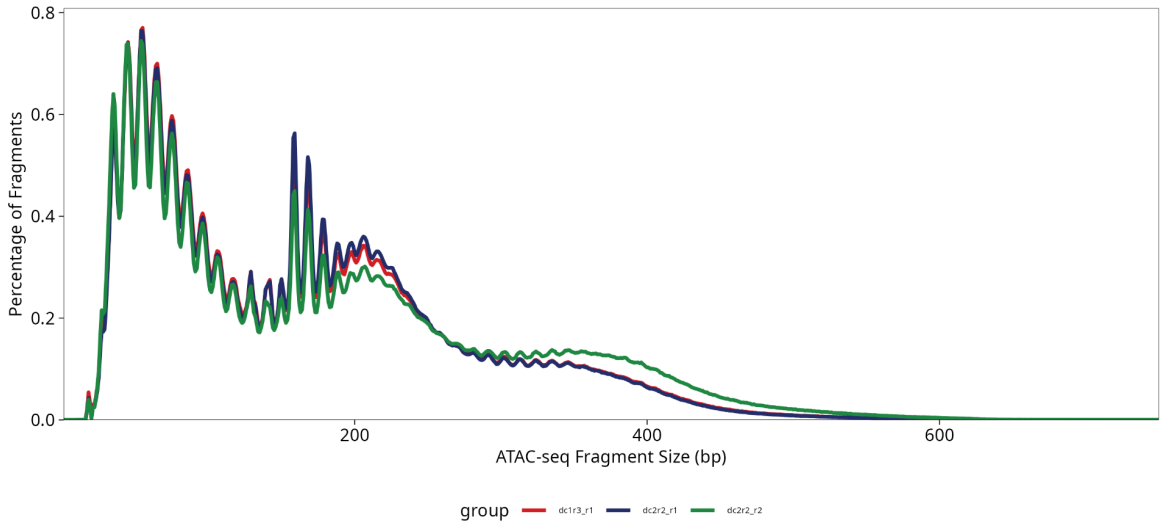
We combined all samples into a joint ArchR project and inspected the metadata.

Results:

Metric	Value
Total number of cells	9500
Median TSS Enrichment value	19.895
Median number of fragments	9202
Dimension of the peak set	

Quality Control Plots: Summary and Observations

1. Fragment Length Distribution



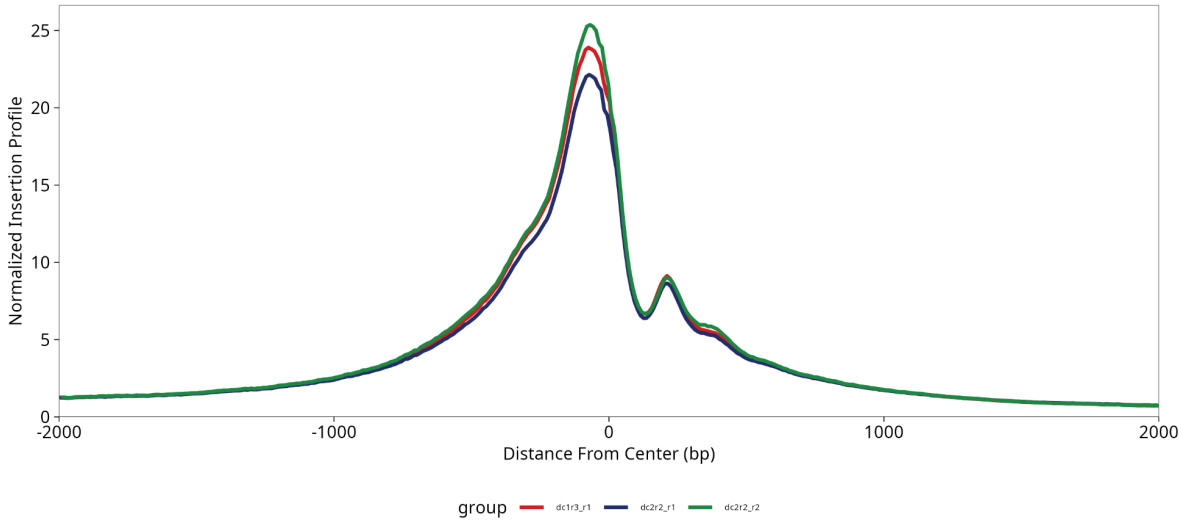
Interpretation

The **fragment length distribution plot** displays the percentage of fragments grouped by their sizes.

- **Key peaks**:
  - **~147 bp**: Represents **nucleosome-free regions (NFR)**, indicating open chromatin regions.
  - **~300 bp**: Reflects **mono-nucleosome fragments**, further suggesting high-quality chromatin accessibility data.
- **Sample Comparison**:
  - All three samples (`dc1r3_r1`, `dc2r2_r1`, `dc2r2_r2`) show **consistent fragment length distributions**.
  - Minimal deviations between the samples confirm reproducibility and the absence of significant anomalies.

And the observed fragment length patterns indicate high-quality ATAC-seq data with expected peaks and minimal deviations.

2. TSS Enrichment Distribution



Interpretation

- The **TSS enrichment distribution plot** shows the normalized insertion profile around **Transcription Start Sites (TSS)**.
- A **sharp peak** at the center (0 bp) confirms enrichment of reads at active TSS regions.
- The profiles across all three samples exhibit the expected "peak shape":
    - A sharp central peak (active TSS).
    - Symmetrical decline on both sides (flanking regions).
- **Sample Comparison**:
    - All three samples display **high TSS enrichment**, reflecting good data quality.
    - Slight differences in peak heights are minor and can be attributed to biological or technical variations.

Conclusion: High enrichment near TSS regions indicates robust and high-quality ATAC-seq data across all samples.

3. Fragments vs TSS Enrichment Scatter Plot

Observations

- The scatter plot highlights the relationship between **fragment count** (log scale) and **TSS enrichment scores** for individual cells across all three samples.
- **Majority of Cells**:
  - TSS enrichment scores between **10–30**.
  - Fragment counts in the range of **10310^3 to 10510^5**.
- **Key Patterns**:
  - **High TSS enrichment (> 30):** Likely represents high-quality cells with good chromatin accessibility.
  - **Low TSS enrichment (< 10):** Potential low-quality or dead cells.
  - **Low fragment counts (< 3200):** These may correspond to cells with insufficient sequencing depth or technical artifacts.
- **Potential Outliers**:
  - **Low TSS and High Fragment Count**: Cells with high fragment counts but low enrichment scores may reflect problematic cells or noise.
  - **Low TSS and Low Fragments**: These cells are likely low quality and candidates for removal.

Conclusion: The scatter plot highlights a **clear clustering of high-quality cells**, along with identifiable **outlier populations** that can be targeted for stricter filtering.

Filter the Dataset

After applying stricter filtering, the updated QC metrics are:

Metric	Value
Total number of cells	6176
Median TSS Enrichment	19.75
Median number of fragments	10,841.5

Interpretation

1. **Total Cell Count**:
    - The number of cells decreased from **9500** (initial dataset) to **6176**.
    - This reduction is due to the removal of cells that did not meet stricter quality thresholds.
2. **Median TSS Enrichment**:
    - The median TSS enrichment is **19.75**, which remains consistent with the earlier value.
    - This indicates that the remaining cells still have high-quality TSS enrichment profiles.
3. **Median Number of Fragments**:
    - The median fragment count increased to **10,841.5**, reflecting the exclusion of cells with low fragment counts.

So in conclusion

By applying stricter filtering based on:

- **Number of Fragments**: > 3200 and < 100,000
- **TSS Enrichment**: > 10
- **Doublet Score**: < 50

The dataset has been refined to include **6176 high-quality cells**. These cells exhibit:

- Consistent enrichment at TSS regions.
- Adequate fragment counts.

Dimensionality Reduction, Clustering, and Peak Calling

2.1 Iterative LSI: Why LSI Over PCA?

Rationale:

- **Latent Semantic Indexing (LSI)** is better suited for scATAC-seq data compared to PCA due to the **sparsity** and **binary nature** of chromatin accessibility data.
- PCA assumes continuous and normally distributed data, which is not ideal for sparse accessibility matrices.
- LSI incorporates **TF-IDF normalization** (Term Frequency - Inverse Document Frequency), emphasizing features (regions) that are specific to a particular cell rather than globally frequent, enhancing biological signal detection.

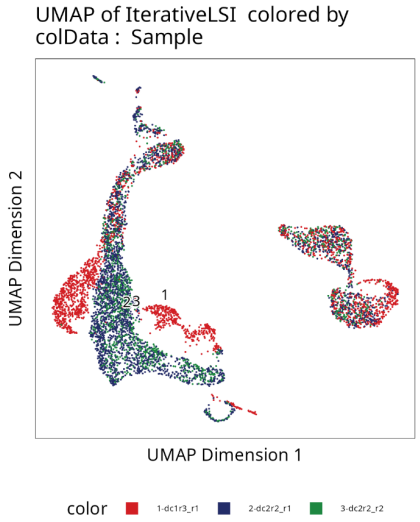
2.2 UMAP Visualization with Annotations

Purpose: UMAP projections are computed to visualize the cells in a lower-dimensional space. We colored the UMAP plots based on:

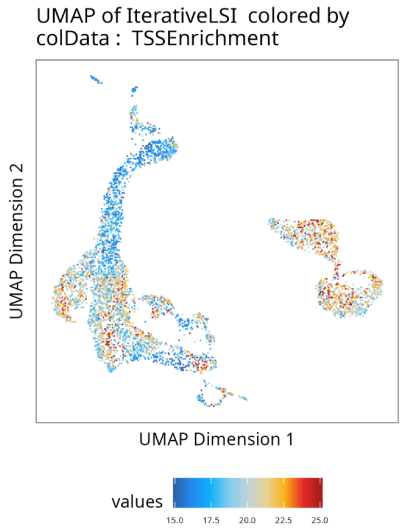
1. **Sample Annotation**: Visualize the distribution of cells from different samples.
2. **TSS Enrichment Scores**: Assess the TSS quality of the cells.
3. **Fragment Counts**: Evaluate the accessibility signal strength per cell.

---

1. **UMAP with Sample Annotation**:
    - Cells from the three samples (`dc1r3_r1`, `dc2r2_r1`, `dc2r2_r2`) cluster into distinct regions of the UMAP space.
    - Some intermixing is observed, which may indicate shared biological states across samples.

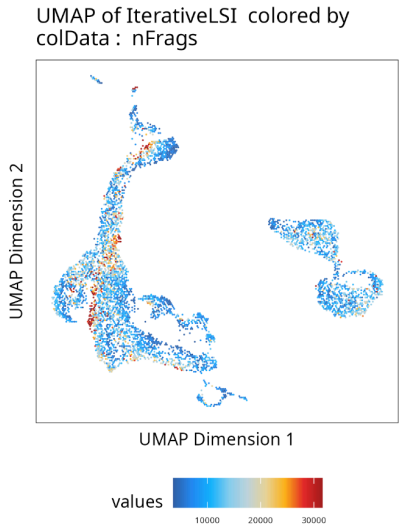


2. **UMAP with TSS Enrichment**:
    - Cells with higher **TSS enrichment** scores (red) cluster predominantly in central regions of the UMAP.
    - Lower-quality cells (blue, lower TSS scores) tend to appear more dispersed or outlying.

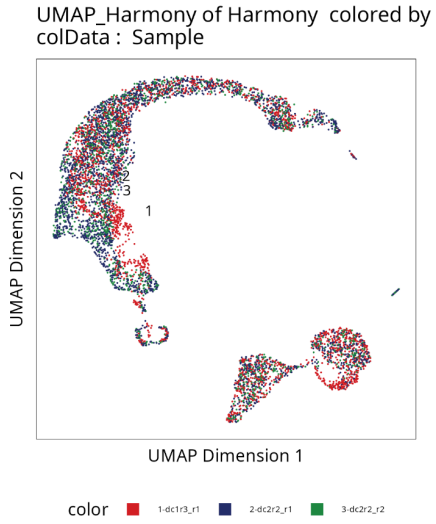


3. **UMAP with Fragment Counts**:
    - Cells with higher fragment counts (red) correlate with regions of higher density in the UMAP, as expected.

- Cells with low fragment counts appear scattered, suggesting lower data quality or sparsity.



2.3 Batch Effects



---

1. UMAP Before Batch Correction:

- The first UMAP plot shows clear batch effects with cells clustering by sample (red, blue, and green regions).
- This clustering by **Sample** suggests batch-related variation dominates the data.
- **Before Batch Correction:** The data exhibited noticeable batch effects, where cells clustered by sample instead of biological signals.

2. UMAP After Batch Correction with Harmony:

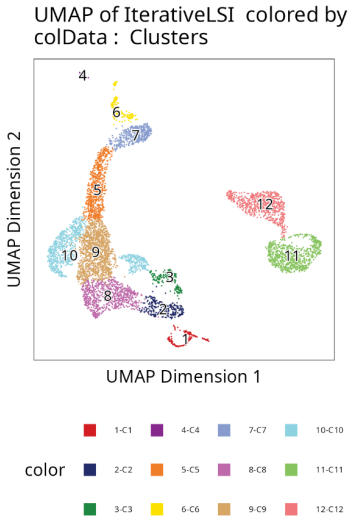
- After applying Harmony, the UMAP plot is more integrated.
- Cells from different samples (red, blue, green) are no longer segregated but overlap, indicating batch effects have been mitigated.
- **After Batch Correction:** Harmony successfully aligned cells across samples. The new UMAP shows improved integration, where clusters are mixed across samples.

in conclusion: Batch correction with Harmony resolves sample-specific variation, enabling more reliable downstream analyses such as clustering and marker identification.

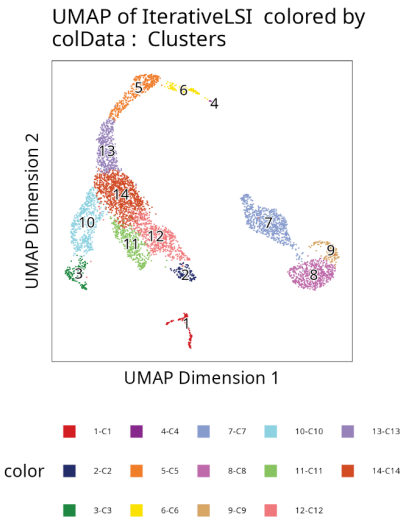
---

3. Clustering

UMAP embedding showing 12 clusters labeled as C1 to C12.



-


UMAP of IterativeLSI colored by
colData : Clusters
UMAP Dimension 2
UMAP Dimension 1
5
6
4
13
14
10
12
11
7
9
3
2
8
1
color
1-C1
4-C4
7-C7
10-C10
13-C13
2-C2
5-C5
8-C8
11-C11
14-C14
3-C3
6-C6
9-C9
12-C12

Cluster Cell Counts

The total number of cells across 12 clusters is **6176**. Here is the breakdown of cells per cluster:

Cluster	Number of Cells
C1	132
C2	279
C3	168
C4	37
C5	690
C6	158
C7	405
C8	751
C9	1202
C10	871
C11	773
C12	710

Cluster Distribution:

- Clusters **C9, C10, and C8** contain the largest number of cells, suggesting these clusters represent dominant cell states or populations.
- Clusters **C1 and C4** have the fewest cells, potentially representing rare cell populations.

2. **Sample Representation**:
   - Observing the earlier UMAP (colored by `Sample`), clusters appear to include cells from all three samples (`dc1r3_r1`, `dc2r2_r1`, and `dc2r2_r2`).
   - This indicates that clustering is likely driven by biological variation rather than technical batch effects.
3. **Cluster Spatial Separation**:
   - The UMAP shows distinct spatial separation between clusters, which suggests that the **Louvain clustering** effectively identifies groups of cells with shared accessibility patterns.

4 Peaks

4.1 Peak Calling

1. **Group Coverages**
   The data was grouped by **clusters** to generate per-cluster coverage tracks, which are essential for improving peak calling sensitivity.
   - the reason:
     Aggregating single-cell data at the cluster level ensures sufficient signal-to-noise ratio for peak detection.
2. **Peak Calling Using MACS2**
   Reproducible peaks were called for each cluster using **MACS2**, a widely used peak caller for ATAC-seq.
   - **Parameters**:
     - `groupBy = "Clusters"`: Call peaks for each cluster.
     - `pathToMacs2`: Path to the MACS2 binary.
     - Peaks were optimized based on replicates and cluster-level coverage.

Results

Cluster	Total Cells	Reproducible Peaks	Max Peaks Allowed
C1	132	106	53,000
C2	279	276	138,000
C3	168	168	84,000
C4	37	35	17,500
C5	690	690	150,000
C6	158	158	79,000
C7	405	405	150,000
C8	751	749	150,000
C9	1202	882	150,000
C10	871	540	150,000
C11	773	773	150,000
C12	710	710	150,000

4.2 Cluster Marker Peaks

Objective

Identify and visualize **cluster-specific peaks** using appropriate statistical cutoffs and display gene accessibility for selected marker genes.

---

1. **Add Peak Matrix**
   Quantify the chromatin accessibility across all identified peaks.

- **the reason is because:**
  The peak matrix provides a quantified view of chromatin accessibility for each peak and cluster.

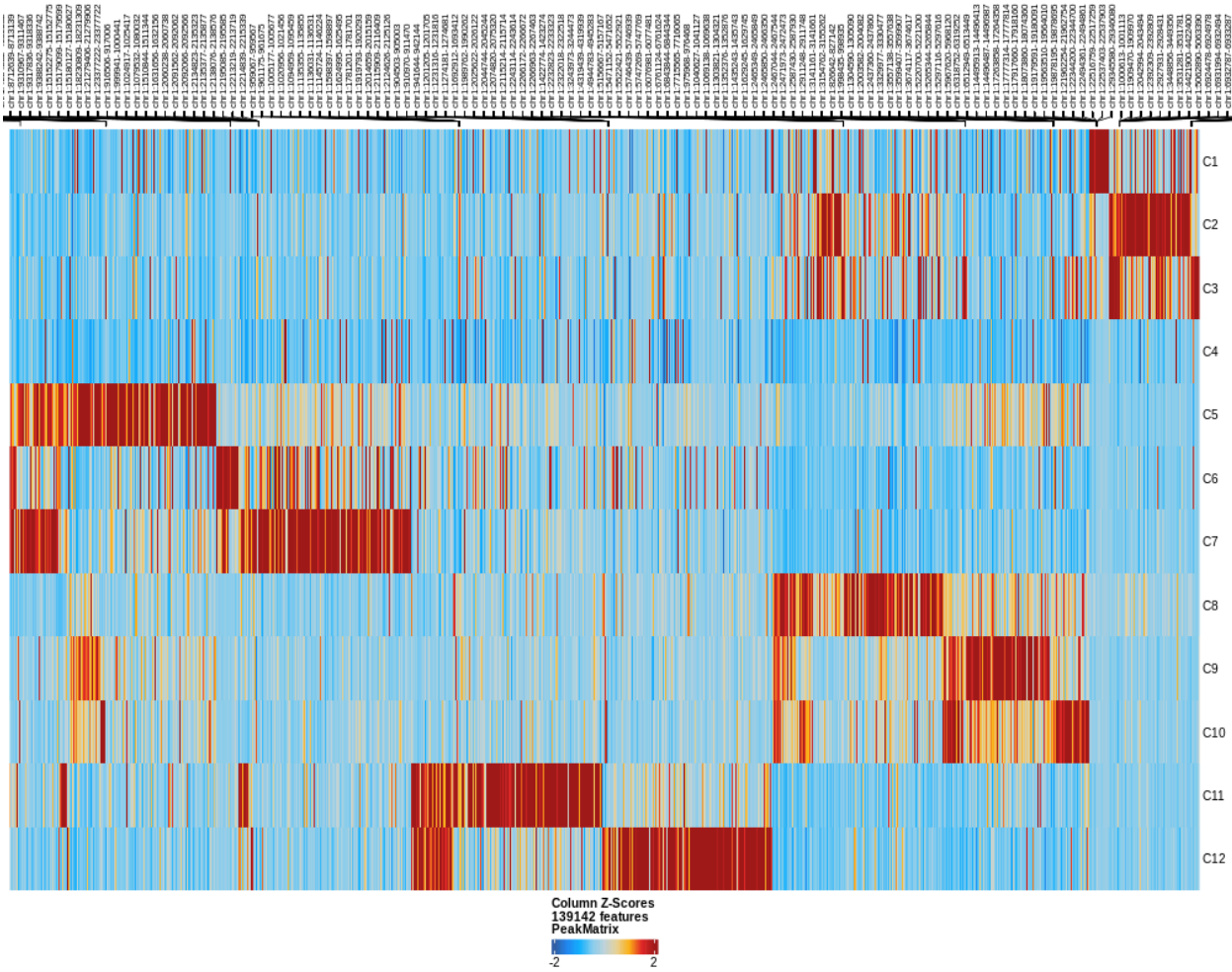
---

2. **Identify Marker Peaks**
   Marker peaks were identified for each cluster using the `wilcoxon` statistical test.
   - **Parameters**:
     - `useMatrix`: PeakMatrix.
     - `groupBy`: Clusters.
     - `bias`: TSSEnrichment and log-transformed fragment counts.
     - `cutOff`: FDR ≤ 0.1, Log2FC ≥ 0.5 (significant peaks).

---

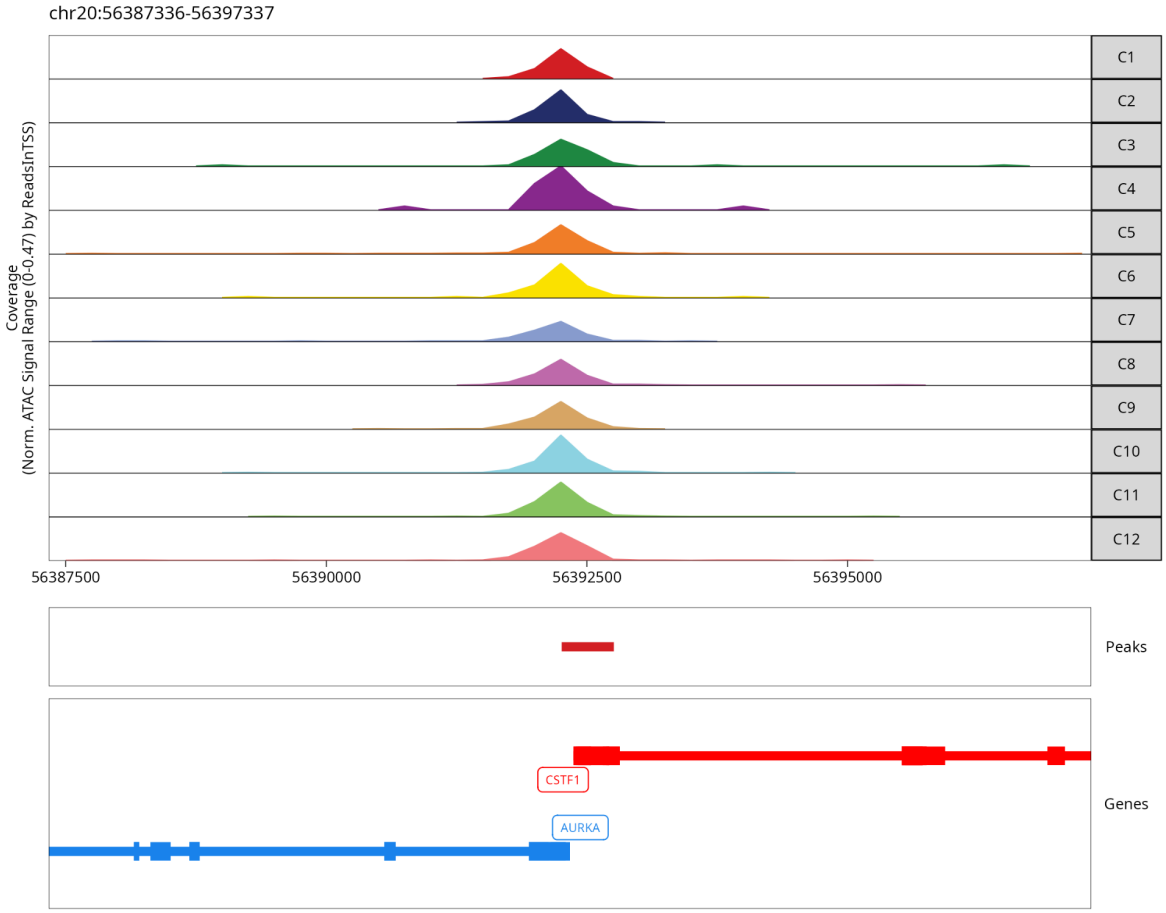
3. **Heatmap of Marker Peaks**
   A heatmap of cluster-specific peaks was generated and
   - 139,142 significant peaks were identified.
   - The heatmap highlights differential chromatin accessibility across clusters.
4. **Heatmap Observation**:

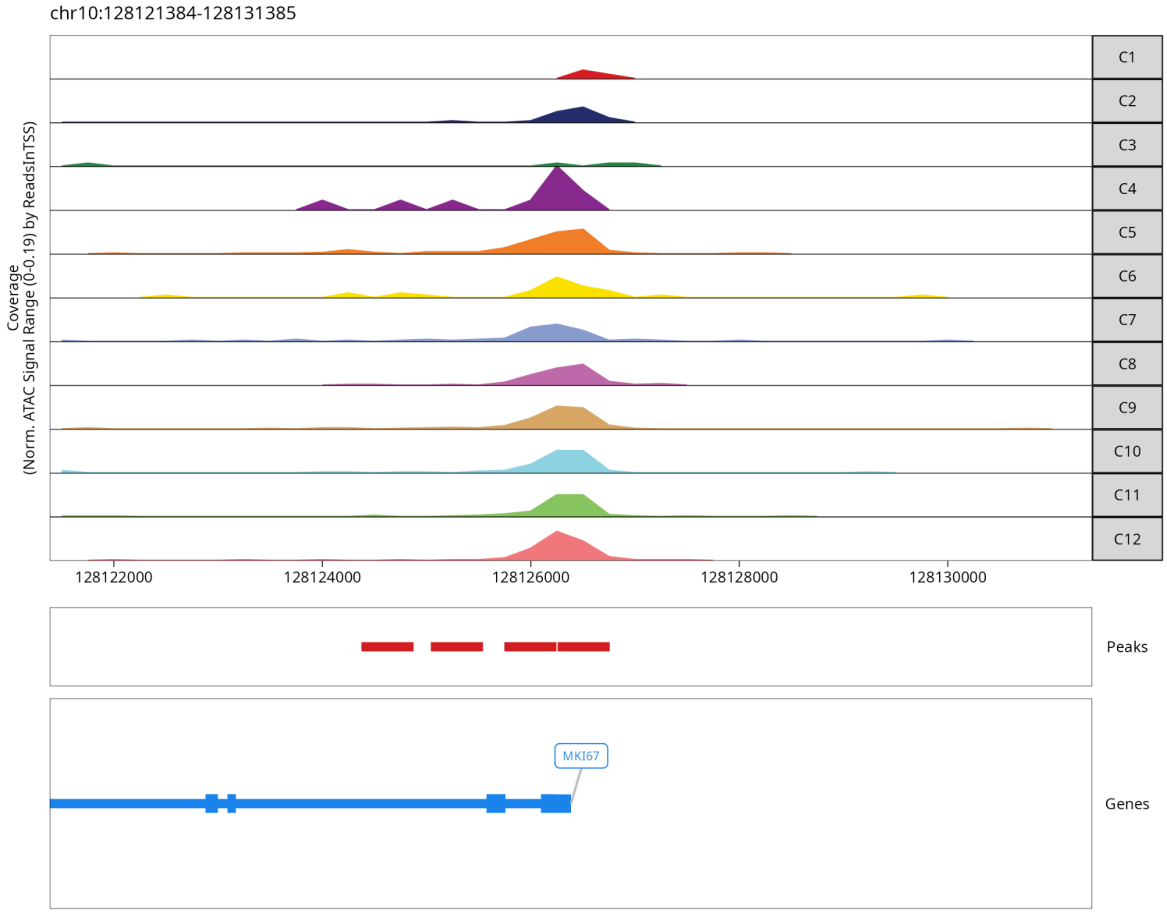
- Certain clusters (e.g., **C2**, **C5**, and **C9**) show higher chromatin accessibility for distinct peaks, indicating specific regulatory activity in these clusters.

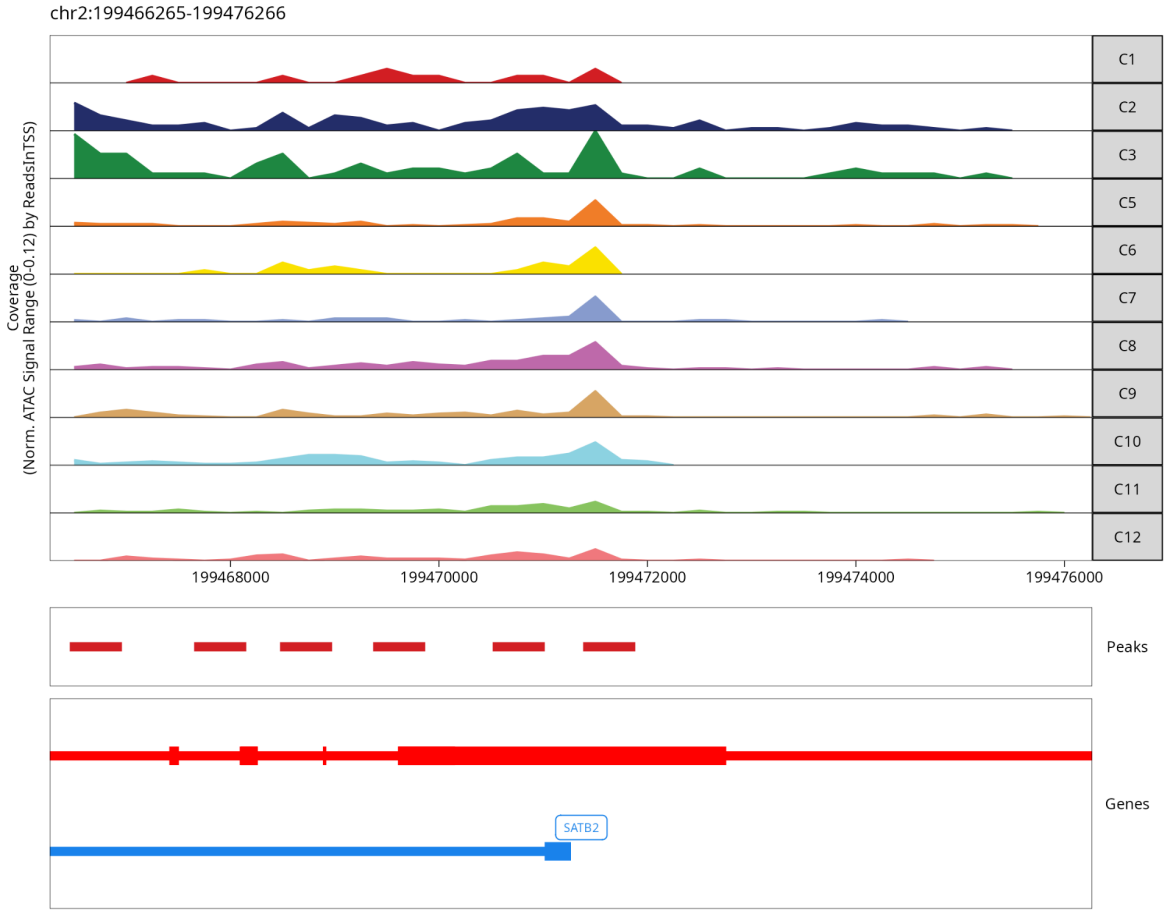


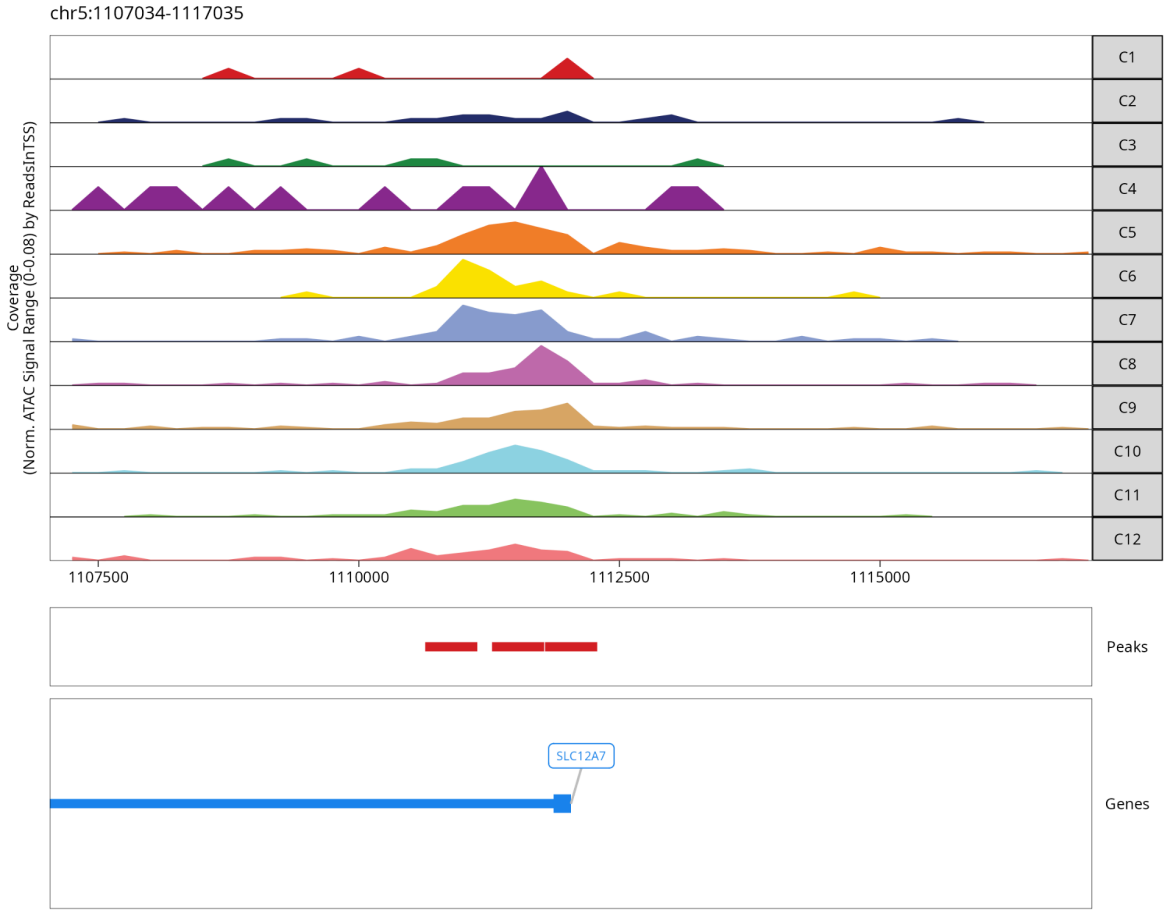
---

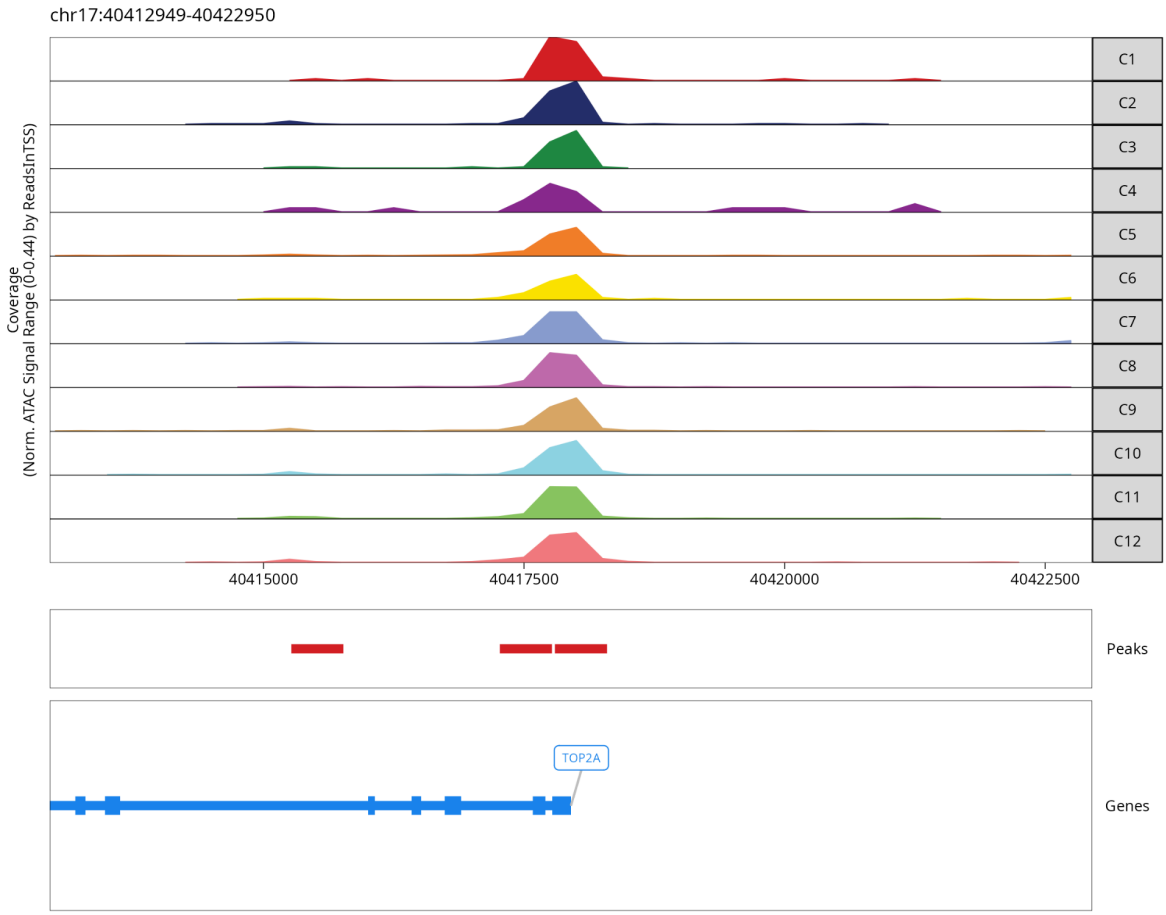
4. **Gene-Specific Accessibility Visualization**
   Plots were generated for genes of interest:


chr20:56387336-56397337
Coverage
(Norm. ATAC Signal Range (0-0.47) by ReadsInTSS)
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
56387500
56390000
56392500
56395000
Peaks
CSTF1
AURKA
Genes


chr10:128121384-128131385
Coverage
(Norm. ATAC Signal Range (0-0.19) by ReadsInTSS)
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
128122000
128124000
128126000
128128000
128130000
Peaks
MKI67
Genes


chr2:199466265-199476266
Coverage
(Norm. ATAC Signal Range (0-0.12) by ReadsInTSS)
C1
C2
C3
C5
C6
C7
C8
C9
C10
C11
C12
199468000
199470000
199472000
199474000
199476000
Peaks
Genes
SATB2


chr5:1107034-1117035
Coverage
(Norm. ATAC Signal Range (0-0.08) by ReadsInTSS)
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
1107500
1110000
1112500
1115000
Peaks
SLC12A7
Genes



---

5. Gene Activity Analysis

5.1 Compute Gene Activity Scores Using Chromatin Accessibility

- **Purpose**: Gene activity scores were computed to evaluate chromatin accessibility near gene regions, which serves as a proxy for gene expression.
- **Method**:
    - The `GeneScoreMatrix` was used, which calculates gene activity by integrating accessibility across the gene body and flanking regions.
    - The default matrix `"GeneScoreMatrix"` was used to compute these scores, and the output was saved in the project as `"GeneActivity"`.

5.2 Identify Marker Genes

- **Purpose**: To identify cluster-specific marker genes based on their chromatin accessibility.
- **Method**:
    - `getMarkerFeatures()` function was used with the following parameters:
        - **`useMatrix`**: "GeneScoreMatrix" – utilized chromatin accessibility scores.
        - **`groupBy`**: "Clusters" – grouped cells based on previously identified clusters.
        - **Bias**: `TSSEnrichment` and `log10(nFrags)` were provided to correct for technical biases.
        - **Statistical Test**: Wilcoxon Rank Sum Test was applied.
    - Marker genes with the following thresholds were retained:
        - **Cut-off**: FDR ≤ 0.1 and Log2FC ≥ 0.5.

Heatmap of Top Marker Genes

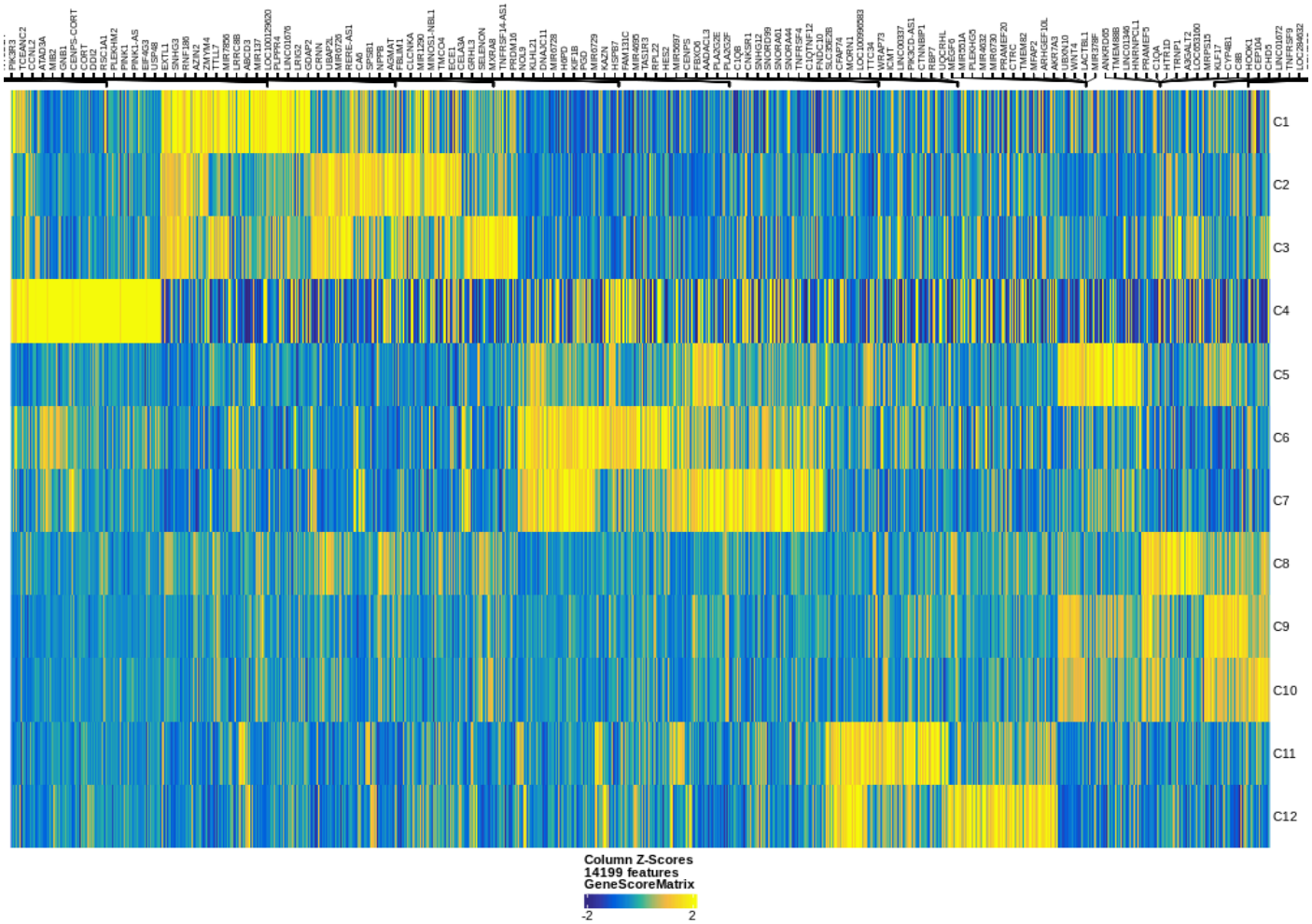
- **Purpose**: To visualize the accessibility patterns of marker genes across clusters.
- **Method**:
    - `plotMarkerHeatmap()` was used to generate the heatmap.
    - A z-score transformation was applied to normalize the accessibility scores across clusters.
    - The heatmap highlights gene accessibility (yellow = high accessibility, blue = low accessibility) across 12 clusters.

---

the heatmap visualizing gene activity scores for the top marker genes across clusters.

- **Patterns**:
    - Clusters **C5, C7, C9, and C11** show strong accessibility at certain marker genes (highlighted in yellow), indicating high chromatin accessibility and potential active transcription in these regions.

- Clusters like **C1** and **C4** show lower accessibility in most genes, suggesting less active transcription.



5.3 Using MAGIC: Gene Activity and Smoothing

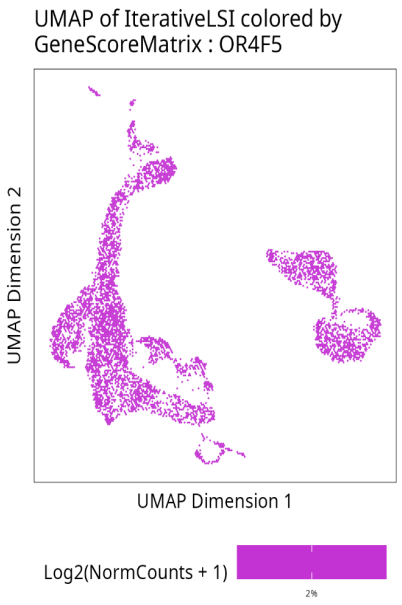
The following steps were implemented:

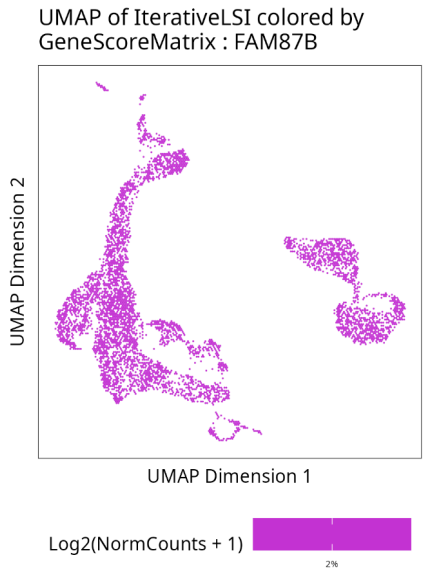
1. **Load the MAGIC Package**:
   `library(Rmagic)`
2. **Select Top 5 Marker Genes**: Marker genes were extracted from the `markerGenes` object:
   `topGenes <- rowData(markerGenes)$name[1:5] # Extract first 5 marker genes`

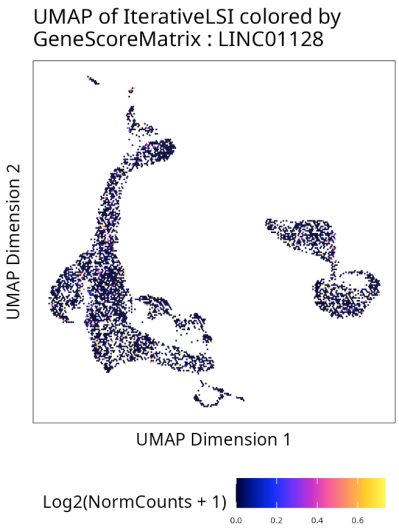
3. **Add MAGIC Smoothing**: MAGIC smoothing was applied using the `addImputeWeights()` function:
   `proj_filtered <- addImputeWeights(proj_filtered)`
4. **Plot UMAP with and without MAGIC**: For each gene, two UMAP plots were created:
   - **Without MAGIC Smoothing**: `imputeWeights = NULL`
   - **With MAGIC Smoothing**: `imputeWeights = getImputeWeights(proj_filtered)`

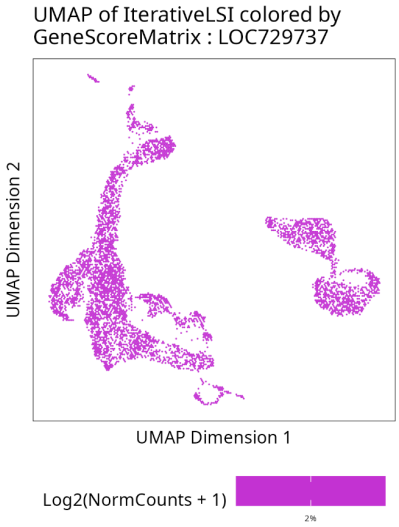
The UMAP visualizations display the expression pattern for the top 5 genes with and without MAGIC smoothing:

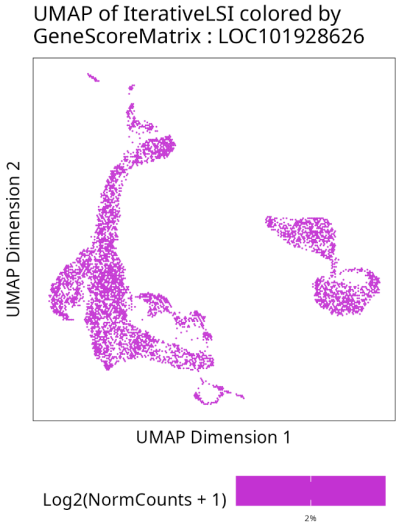
1. **Without MAGIC**:
   - The expression signal is sparse, highlighting individual cells with chromatin accessibility near marker genes.


UMAP of IterativeLSI colored by
GeneScoreMatrix : OR4F5
UMAP Dimension 2
UMAP Dimension 1
Log2(NormCounts + 1)
2%

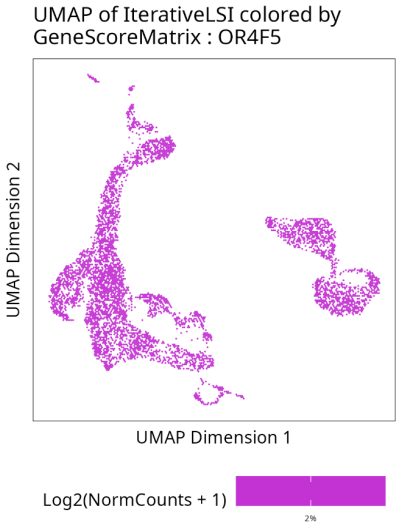

UMAP of IterativeLSI colored by
GeneScoreMatrix : FAM87B
UMAP Dimension 2
UMAP Dimension 1
Log2(NormCounts + 1)
2%

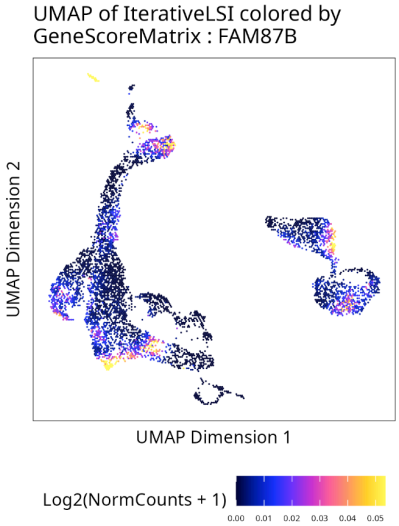

UMAP of IterativeLSI colored by
GeneScoreMatrix : LINC01128
UMAP Dimension 2
UMAP Dimension 1
Log2(NormCounts + 1)
0.0
0.2
0.4
0.6

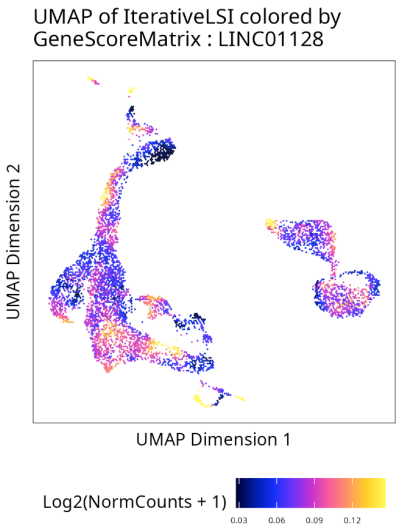


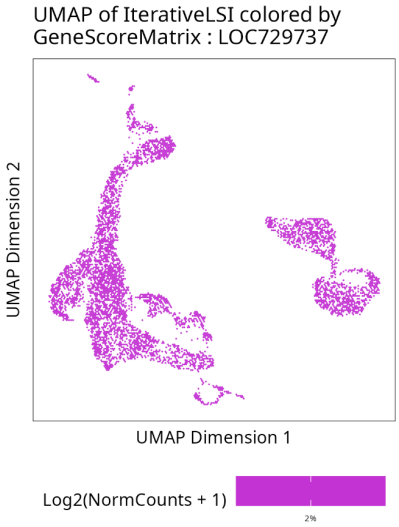


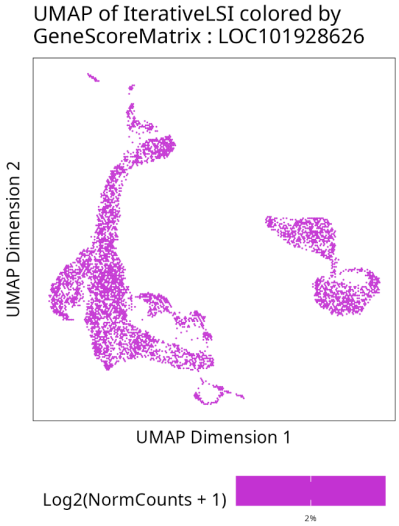
2. **With MAGIC**:
    - MAGIC smoothing enhances the signal by imputing nearby values, improving visualization of gene expression patterns across clusters.


UMAP of IterativeLSI colored by
GeneScoreMatrix : OR4F5
UMAP Dimension 2
UMAP Dimension 1
Log2(NormCounts + 1)
2%


UMAP of IterativeLSI colored by
GeneScoreMatrix : FAM87B
UMAP Dimension 2
UMAP Dimension 1
Log2(NormCounts + 1)
0.00 0.01 0.02 0.03 0.04 0.05


UMAP of IterativeLSI colored by
GeneScoreMatrix : LINC01128
UMAP Dimension 2
UMAP Dimension 1
Log2(NormCounts + 1)
0.03
0.06
0.09
0.12





Interpretation:

- **LINC01128**: Shows a clear spread of accessibility signal across certain regions, becoming more pronounced with MAGIC smoothing.
- **LOC729737 and LOC101928626**: Sparse signals without smoothing appear more continuous after applying MAGIC, revealing broader regions of gene activity.
- **OR4F5**: Accessibility signal becomes more uniform across the clusters when MAGIC is applied.

Main Idea Behind MAGIC:

MAGIC (Markov Affinity-based Graph Imputation of Cells) smooths sparse single-cell data by leveraging relationships between similar cells. It uses graph-based methods to impute missing values, improving the continuity and resolution of gene activity patterns for downstream analysis.

This allows for better visualization and clustering of gene activities, especially when signals are noisy or sparse in single-cell data.

6 Transcription Factor Motif Activity

6.1 Compute TF Motif Activity

- **Motif Annotation Chosen:** The **CIS-BP** motif annotation database was used.
- **How Was the Annotation Obtained?**
  The annotation was added to the `ArchRProject` using the `addMotifAnnotations()` function, specifying the CIS-BP motif set. The motifs were matched to peak regions within the chromatin accessibility data.

6.2 Plot UMAP Embeddings for Marker TFs

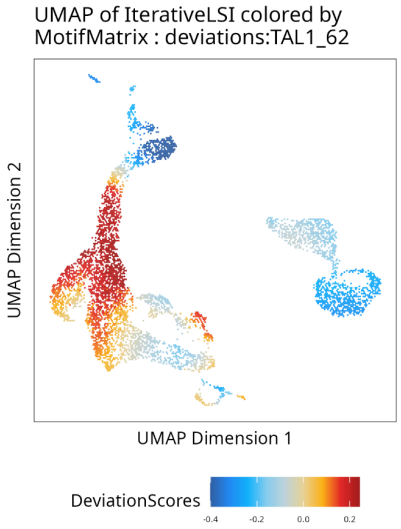
- **Top Variable TF Motifs Identified:**
  From the variability analysis, the most variable motifs identified were:
  - **TAL1_62**
  - **TAL2_822**

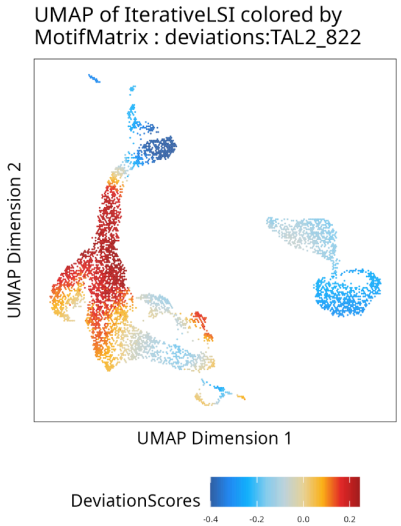
`

Results: The UMAP plots for the top 2 marker motifs (TAL1_62 and TAL2_822) are displayed.

- TAL1_62 is highly active in specific clusters with significant variability.

- TAL2_822 exhibits similar variability across certain regions in UMAP space.



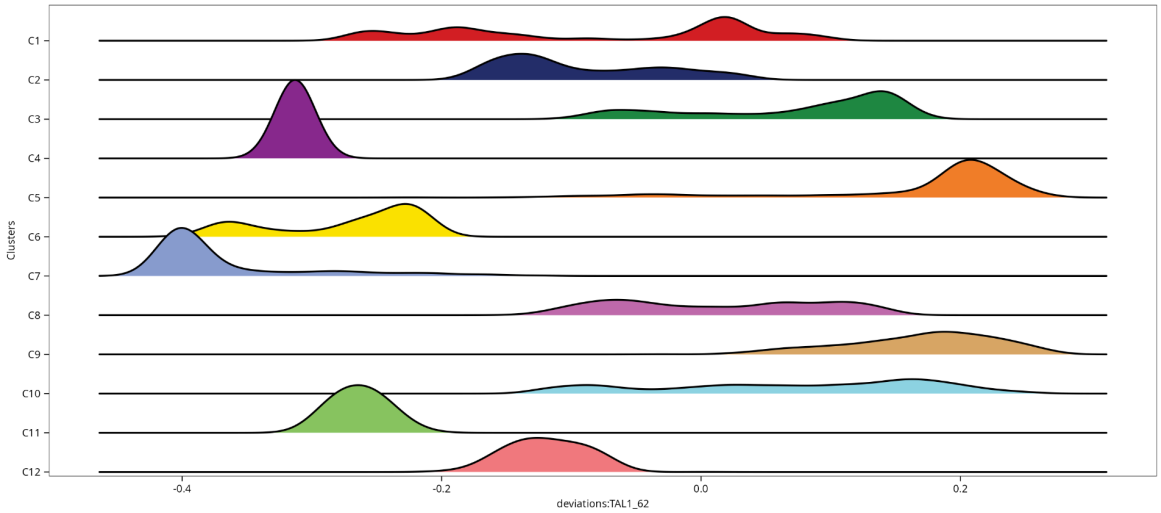


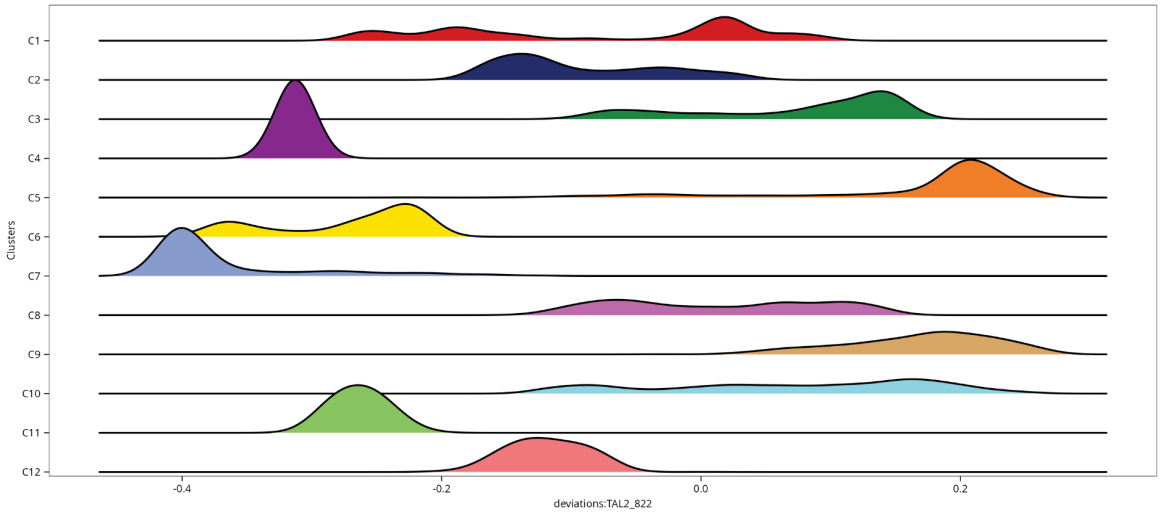
6.3 Motif Activity Distribution Across Clusters

The distribution of motif activity for the top 2 TF motifs (TAL1_62 and TAL2_822) across clusters was plotted.

Results:
The **distribution plots** of motif activity (deviation scores) across clusters for TAL1_62 and TAL2_822 show cluster-specific enrichments:

- **TAL1_62:** Active in clusters **C1, C2, and C6**.
- **TAL2_822:** Displays a similar pattern of activity, indicating a shared regulatory role in these clusters.





---

so in conclusion:

This step computed motif activity for transcription factors using the **CIS-BP database**. The most variable motifs (**TAL1_62** and **TAL2_822**) were identified, visualized on UMAP embeddings, and their activity distributions across clusters were plotted. These findings reveal cluster-specific regulatory motif activity patterns.