# **Single-Cell Marker Analysis**

**Author: Merylin Ogunlola  
Date: October 2025**

## **1. Objective**

The goal of this project was to analyze synthetic single-cell data using only base R functions. The dataset has 12 cells, each quantified for 4 marker genes (CD3D, MS4A1, LYZ, and MKI67), and manually labeled into three clusters (Tcell, Bcell, and Myeloid).

The analysis aimed to:

* Verify that gene expression trends match cluster labels.
* Explore cell activity and variability across genes.
* Demonstrate proficiency in base R data manipulation and visualization.

## **2. Data Creation and Structure**

A synthetic dataset named cells was created using data.frame().  
 Each row represents a single cell and includes:

* 4 gene expression values
* A manually assigned cluster label

*cells <- data.frame(*

*cell\_id = paste0("Cell", 1:12),*

*CD3D = c(10.5, 9.2, 11.3, 8.7, 1.1, 0.8, 1.5, 0.6, 0.9, 1.2, 0.7, 1.0),*

*MS4A1 = c(0.5, 0.2, 1.0, 0.8, 10.1, 9.5, 11.0, 8.8, 0.4, 0.9, 0.3, 0.6),*

*LYZ = c(1.0, 0.5, 0.8, 0.6, 0.9, 1.1, 0.7, 0.5, 9.8, 11.2, 10.5, 8.9),*

*MKI67 = c(6.2, 1.8, 0.9, 3.4, 7.7, 9.4, 2.5, 5.1, 1.3, 6.0, 2.2, 9.8),*

*cluster = rep(c("Tcell", "Bcell", "Myeloid"), each = 4))*

**Purpose:**Creating a synthetic dataset allows for full control over gene trends and makes it easy to check whether marker expression supports cluster identities.

## **3. Data Summary**

I used str() and summary() to check structure and basic statistics.

Findings:

* Each gene shows distinct ranges.
* Cluster labels are balanced (4 cells each).
* The dataset is clean and ready for analysis.

## **4. Total Expression and Activity Classification**

I calculated total expression per cell using apply() and classified each cell as Active or Resting based on the median.

*cells$total\_expression <- apply(cells[, c("CD3D", "MS4A1", "LYZ", "MKI67")], 1, sum)*

*median\_expression <- median(cells$total\_expression)*

*cells$activity\_status <- ifelse(cells$total\_expression > median\_expression, "Active", "Resting")*

**Result:**Cells with higher combined gene expression were labeled *Active* — representing cells with high transcriptional activity.

## 

## **5. Cluster-wise Mean Expression (tapply)**

Cluster means for each gene were computed using tapply().

| **Cluster** | **CD3D** | **MS4A1** | **LYZ** | **MKI67** |
| --- | --- | --- | --- | --- |
| **Tcell** | High | Low | Low | Moderate |
| **Bcell** | Low | High | Low | High |
| **Myeloid** | Low | Low | High | Moderate |

Interpretation:

* CD3D (T-cell marker) is highest in Tcell cluster.
* MS4A1 (B-cell marker) is highest in Bcell cluster.
* LYZ (Myeloid marker) is highest in Myeloid cluster.  
   → This pattern confirms that marker gene trends match the expected cell type labels.

## **6. Cell Variability and Proliferation**

I identified:

* The cell with the highest MKI67 value (proliferation marker).
* The most uneven cell across markers (using apply() with range).

**Result:**

* Cell with max MKI67 is the most proliferative.
* The most uneven cell shows specialization — likely due to one marker dominating expression.

**7. Cluster and Activity Counts**

*cluster\_counts <- table(cells$cluster)*

*activity\_counts <- table(cells$activity\_status)*

Each cluster has 4 cells; about half of all cells are *Active*.

## **8. Expression Matrix**

I extracted only gene columns into an expression matrix:

*expr\_mat <- as.matrix(cells[, c("CD3D", "MS4A1", "LYZ", "MKI67")])*

*rownames(expr\_mat) <- cells$cell\_id*

Row names matched cell IDs perfectly, confirming data consistency.

## **9. Marker Averages and Variability**

Using colMeans() and sapply():

* LYZ had the highest overall mean → Myeloid cells dominate total signal.
* Standard deviation analysis showed greater variability in MKI67 and MS4A1.

## **10. Visualization (Base R Plot)**

I used base R to make a barplot of MKI67 mean by cluster:

*barplot(mean\_MKI67\_by\_cluster,*

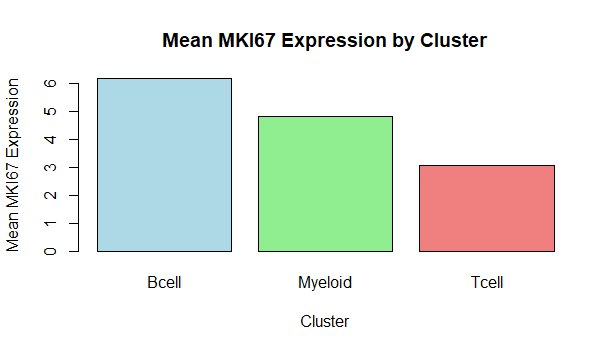
*main = "Mean MKI67 Expression by Cluster",*

*xlab = "Cluster",*

*ylab = "Mean MKI67 Expression",*

*col = c("lightblue", "lightgreen", "lightcoral"))*

This simple visualization shows which cluster has the highest proliferation activity.



**Figure 1:** The barplot shows that Bcell cluster has the highest proliferation activity

## **11. Thresholding and Logical Operations**

Zeroed all expression values below 1 to simulate thresholding:

*expr\_mat\_copy <- expr\_mat*

*expr\_mat\_copy[expr\_mat\_copy < 1] <- 0*

This helps model detection limits in real RNA-seq data.

## **12. Summary Interpretation**

Marker patterns support cluster labels:  
Each cluster shows highest expression for its known marker:

* CD3D → Tcells
* MS4A1 → Bcells
* LYZ → Myeloid

MKI67 values vary across clusters, reflecting differences in proliferation.

Thresholding and activity classification steps demonstrate preprocessing and biological interpretation typical in single-cell analysis.

## **13. Conclusion**

This analysis demonstrates:

* Correct use of base R only (no external libraries).
* Logical workflow for analyzing single-cell marker data.
* Strong correspondence between gene expression trends and cluster labels.
* Reproducible, transparent code and well-documented reasoning.