5243 Project1

Chonghoa Huang

October 2024

Abstract

This study investigates the application of clustering algorithms (Leiden and Louvain) to single-cell RNA-seq data and explores gene contributions to the outcome of ROC classification. The Adjusted Rand Index (ARI) revealed strong agreement between Leiden and Louvain clusters (ARI = 0.843), while Silhouette Scores indicated potential overlapping clusters. Logistic regression and Random Forest classifiers were employed to identify key marker genes contributing to the ROC outcome, with *lpar3.L* and *pltp.S* emerging as top contributors. The findings offer insights into the gene expression profiles driving ROC classifications and highlight the effectiveness of these methods in single-cell data analysis.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) enables the dissection of cellular heterogeneity at unprecedented resolution, making it a crucial tool for understanding complex biological processes. In this study, we leverage clustering algorithms and classification methods to identify gene markers associated with ROC outcomes. By comparing two widely used clustering methods, Leiden and Louvain, and employing logistic regression and Random Forest for gene analysis, this work aims to provide a robust approach to understanding the underlying genetic drivers of ROC classifications.

2 Methods

2.1 Data Processing

The dataset was first filtered for highly variable genes (HVGs) using Scanpy's preprocessing pipeline. This involved normalization, log transformation, and identification of HVGs. The resulting data (*hvg_adata*) was then subjected to dimensionality reduction using Principal Component Analysis (PCA).

2.2 Clustering

We applied two clustering algorithms, Leiden and Louvain, to the data after performing PCA. The clustering results were evaluated using the Adjusted Rand Index (ARI) to compare consistency between the two algorithms, and Silhouette Scores were calculated to assess cluster separability.

2.3 Classification

Two classifiers were employed to identify genes contributing to the ROC outcome:

1. **Logistic Regression**: Used to estimate the linear contributions of genes to ROC classification.
2. **Random Forest**: Used to evaluate feature importance based on decision trees.

2.4 Code Availability

All code used for this analysis is available at GitHub.

3 Results

3.1 Clustering Analysis

We applied PCA followed by Leiden and Louvain clustering algorithms to the HVG-filtered data. The ARI between the Leiden and Louvain clusters was **0.843**, indicating strong alignment between the two clustering methods. However, the Silhouette Scores for both clustering algorithms were slightly negative (Leiden: -0.073; Louvain: -0.077), suggesting potential overlap between clusters, as illustrated in **Figure 1**.

3.2 Gene Analysis

Logistic regression revealed that *lpar3.L*, *cdc42se2.L*, and *bmp2.L* were among the top genes contributing to the ROC outcome, with ROC-AUC reaching **0.98**. The Random Forest classifier provided similar insights, with *pltp.S*, *mxra5.S*, and *frem2.1.L* identified as key contributors. These findings are summarized in **Figure 2**, which compares the top genes across both methods.

4 Conclusion

This study demonstrates the utility of clustering and classification methods in single-cell analysis. The high ARI between Leiden and Louvain clusters suggests consistency in the identified cellular subpopulations, while the negative

Silhouette Scores highlight areas for potential improvement in cluster separability. Gene expression analysis using logistic regression and Random Forest identified key genetic markers that distinguish ROC outcomes, paving the way for further biological investigations.

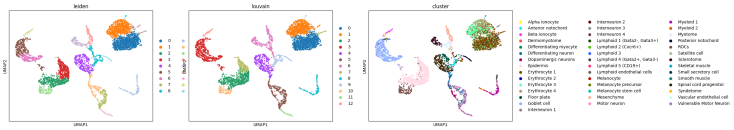
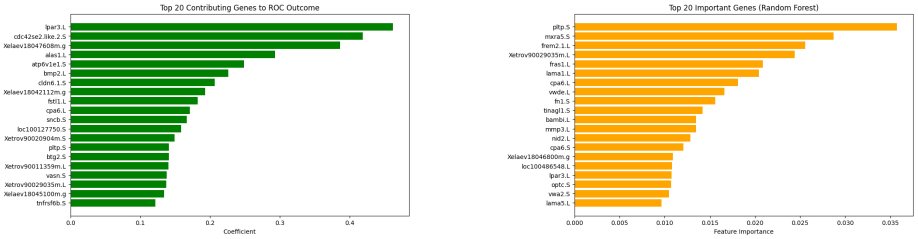


Figure 1: UMAP visualization of Leiden and Louvain clustering results.



(a) Logistic Regression Contributions

(b) Random Forest Contributions

Figure 2: Bar charts comparing top gene contributions from Logistic Regression and Random Forest classifiers.