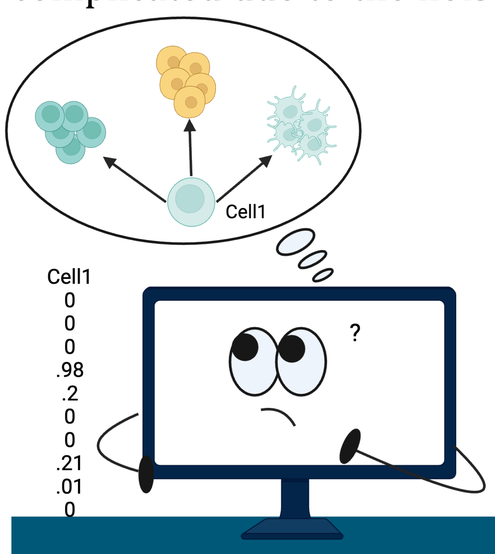
Integration of Marker Genes as Centroids in Single-Cell RNA sequencing Clustering: A Machine Learning-Driven Approach for Accurate Identification of Cell Types

Denisse Chacón Ramírez, Claudia Rangel Escareño
School of Engineering and Sciences, Tecnologico de Monterrey.

Motivation

scRNA-seq analysis plays a crucial role in precision medicine by enabling the identification of involved cell types in various diseases [1]. Clustering is an essential step in scRNA-seq analysis as it helps to group together cells with similar gene expression profiles, leading to the identification of cell populations. However, this step is complicated due to the noise and sparsity present in scRNA-seq data [2].



Decades of research has led to the identification of marker genes, those with expression profiles that define a particular cell type, useful for cellular characterization [3]. It is therefore valuable to analyze how different clustering algorithms behave and if marker genes truly make a difference identifying cell types.

Methods

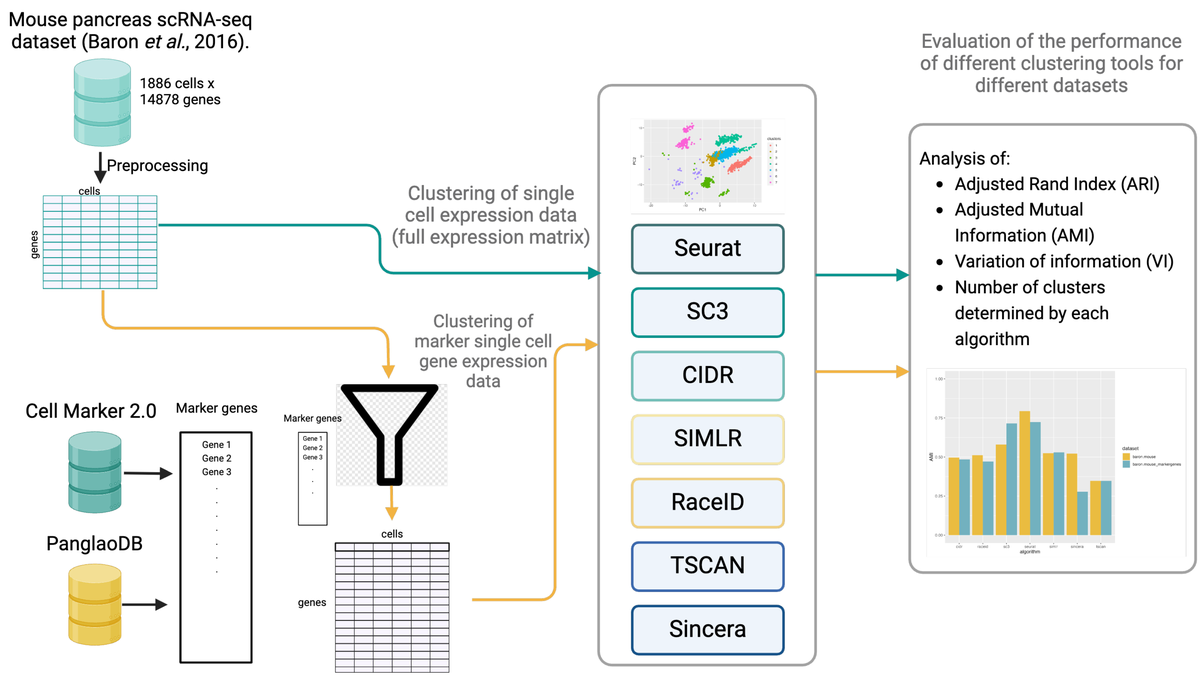


Figure 1. Methodology overview. The mouse pancreas scRNA-seq dataset [4] was downloaded and preprocessed to obtain the expression matrix. Marker genes for mouse pancreas were downloaded from Cell Marker and PanglaoDB databases. The marker genes were used to filter expression matrix rows to extract features with high cell type impact. Then, 7 clustering algorithms were applied to filtered and full data. Clustering results were evaluated.

Conclusions and Future Directions

- While marker genes provide valuable cell identity information, there are multiple biological and computational factors that can limit their direct utility for enhancing algorithm-based clustering performance.
- The use of only reported high-impact cell type information does not universally improve the performance of clustering algorithms, suggesting that genes not recognized as markers likely contribute to distinguishing cell types.

Acknowledgments

- We acknowledge Tecnológico de Monterrey, INMEGEN and CONAHCYT for their support.

Results

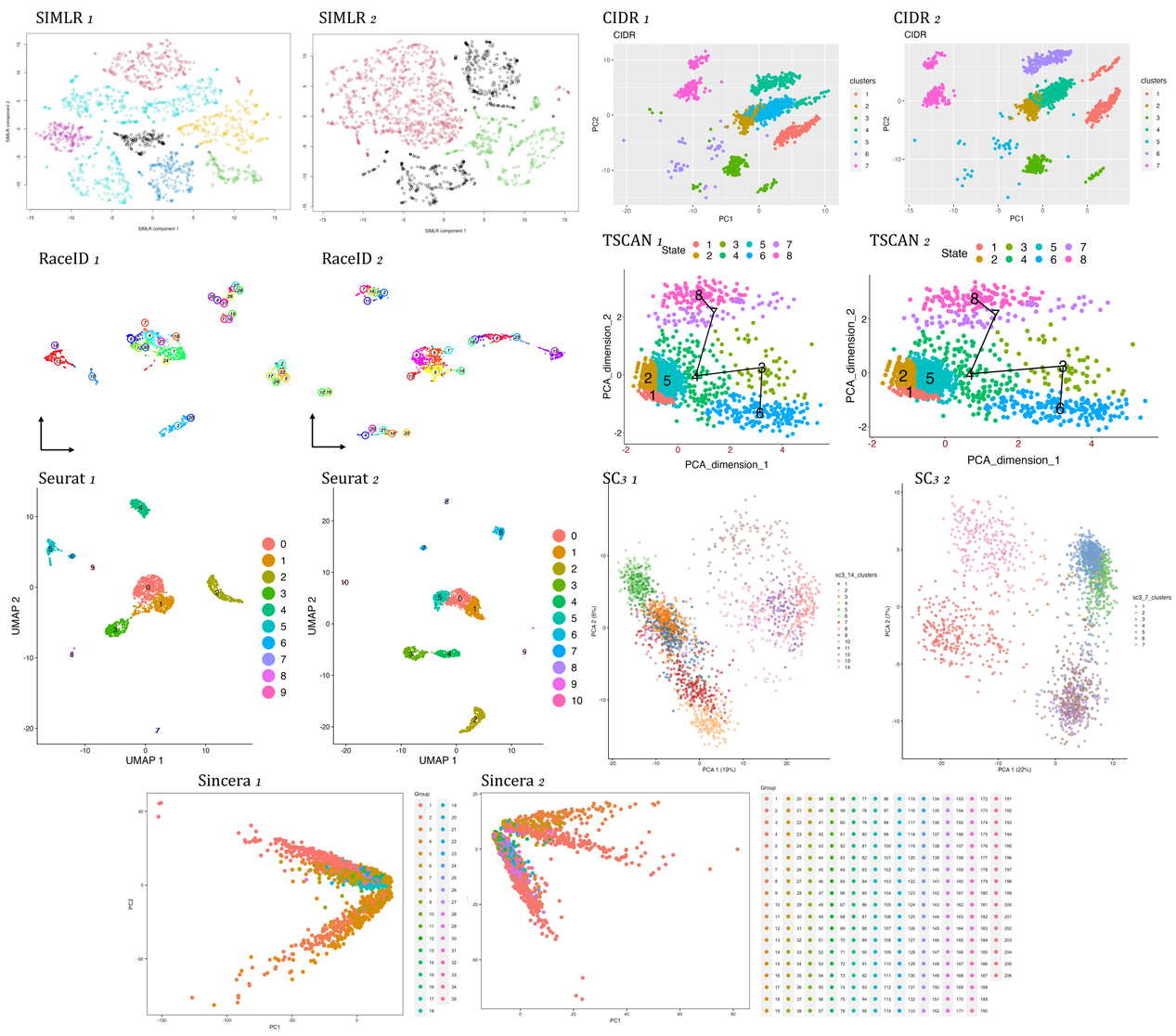


Figure 2. UMAP, t-SNE and PCA projections for Baron-mouse dataset colored according cluster assignation of each algorithm. The first set of projections (labeled as 1) corresponds to the full expression matrix, while the second set (labeled as 2) represents the expression matrix post-feature extraction.

Some algorithms enhanced their performance after feature extraction using marker genes, while others worsened or maintained similar performance. This variability may be because:

•The available gene lists provided as cell type markers could be limited or incomplete. Certain relevant genes may be missing for specific cell identities [5].

•Certain genes may not be detected in a given cell type [6]. Absence of expression can serve as a negative marker, yet dropout events can complicate the detection of these non expressed genes [7].

•Marker genes can be expressed in more than one cell type and across multiple clusters, complicating cluster designation [7, 8].

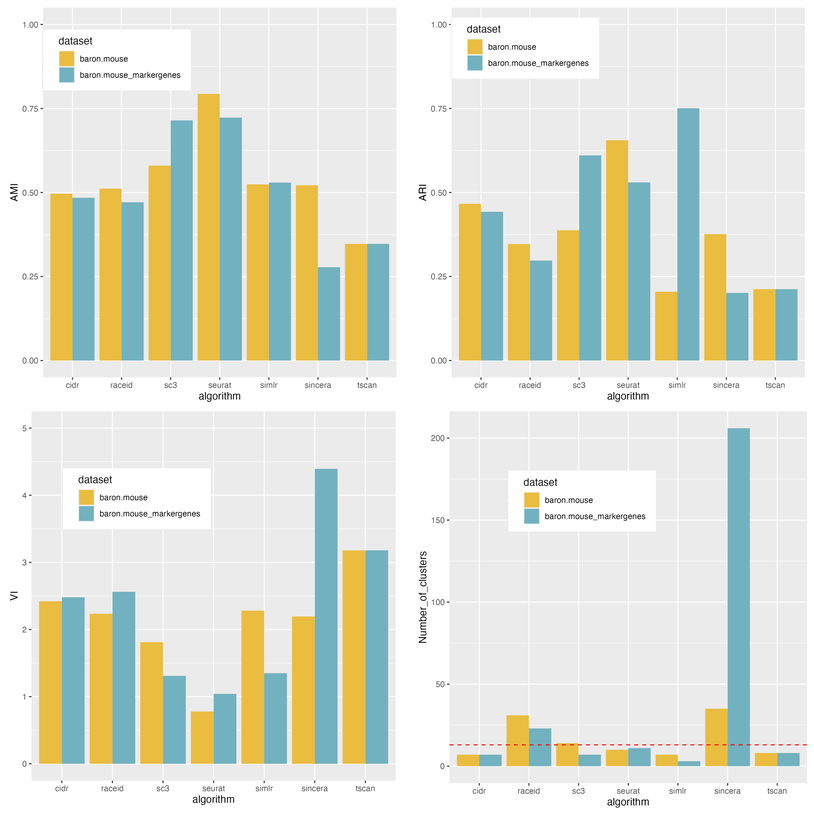


Figure 3. Concordance of clustering output and predefined cell type labels for the full expression matrix (yellow) and for the expression matrix post-feature extraction (blue). as quantified by three concordance measures. A) Results for Adjusted Mutual information (AMI). B) Results for Adjusted Rand Index (ARI). C) Results for Variation Information. D) Number of clusters determined by each algorithm, the red line represents the number of cell types in the dataset.

References

Code Availability