[BIOINFO 2023]

Characterizing feature selection for single-cell expression analysis

Juok Cho[1†], Bukyung Baik[2†], Hai C. T. Nguyen[2], Daeui Park[3], Dougu Nam[2,4*]

[1]*Department of Biomedical Engineering, Ulsan National Institute of Science and Technology*

[2]*Department of Biological Sciences, Ulsan National Institute of Science and Technology*

[3]*Department of Predictive Toxicology, Korea Institute of Toxicology*

[4]*Department of Mathematical Sciences, Ulsan National Institute of Science and Technology*

†Juok Cho and Bukyung Baik contributed equally to this work.

**Corresponding author: dougnam@unist.ac.kr*

Unsupervised feature selection is a critical step for efficient and accurate analysis of single-cell RNA-seq data (scRNA-seq). Various feature selection methods for scRNA-seq data have been developed; however, systematic benchmarks that characterize good methods are rare.

Previous benchmarks used two different criteria to compare feature selection methods:
(1) proportion of ground-truth marker genes included in the selected features and (2) accuracy of cell clustering using ground-truth cell types. Here, we compare the performance of eleven feature selection methods for both criteria. We demonstrate the discordance between these criteria and suggest using the latter. In particular, we show the widely used highly variable gene selection includes more marker genes than deviation-based method; however, the latter performs better in clustering cells and data visualization. Notably, deviation-based method was able to clearly separate the same cell type from different tissues and showed its capability to delineate cell trajectories. We find that lowly expressed genes exhibit seriously high noise-to-signal ratios and are mostly excluded by high-performance methods.

In conclusion, highly and lowly expressed features do not make an equal contribution to scRNA-seq downstream analysis, as they exhibit very different noise-to-signal ratios. High-deviation or high-expression-based methods outperform the widely used highly variable gene selection by mostly selecting genes with high mean values.