ClusterDE: a post-clustering differential expression method

Carson Zhang
LMU Munich
carson.zhang@campus.lmu.de

Abstract In typical differential expression analysis, a clustering algorithm is applied to scRNA-seq data, and then a differential expression test is conducted in order to identify genes that are differentially expressed between the clusters. However, this procedure constitutes "double dipping", as it first clusters the data to identify cell types, and then uses those same clusters to identify cell-type marker genes. This leads to an inflated FDR for DE genes. (Song et al., 2023) propose ClusterDE, a post-clustering DE method that controls the FDR of DE genes. ClusterDE generates a synthetic null dataset that preserves the structure of the real data, computes differences between this null dataset and the real data, then performs FDR control on the results. Simulations and real data analysis demonstrate that ClusterDE controls the FDR and identifies cell-type marker genes as top DE genes, successfully distinguishing them from housekeeping genes.

Table of contents

Introduction

Biologists like to identify the cell types in their scRNA-seq samples.

To identify these cell types, they identify a set of cell-type marker genes.

To identify these cell-type marker genes, they perform differential expression testing.

Naive differential expression testing is susceptible to false discoveries caused by double-dipping.

Paper overview

Overview of differential expression methods

ClusterDE

The ClusterDE method consists of four basic steps.

1. Generate a synthetic null dataset that consists of a single cluster but otherwise mimics the real data.
2. Perform clustering on both datasets.
3. Perform differential expression testing on both datasets.
4. Combine the results to determine which genes to output as discoveries (DE genes).

Other differential expression methods

Data analysis

BacSC data

Synthetic null data generation

Schäfer-Strimmer

Results

Simulation study

(Song et al., 2023)

(Ostner et al., 2024)

(Badri et al., 2020)

(Schäfer & Strimmer, 2005)

Appendix

Bibliography

Badri, M., Kurtz, Z. D., Bonneau, R., & Müller, C. L. (2020). Shrinkage improves estimation of microbial associations under different normalization methods. *NAR Genomics and Bioinformatics*, *2*(4), lqaa100. https://doi.org/10.1093/nargab/lqaa100

Ostner, J., Kirk, T., Olayo-Alarcon, R., Thöming, J. G., Rosenthal, A. Z., Häussler, S., & Müller, C. L. (2024). BacSC: A general workflow for bacterial single-cell RNA sequencing data analysis. *Biorxiv*. https://doi.org/10.1101/2024.06.22.600071

Schäfer, J., & Strimmer, K. (2005). *Statistical Applications in Genetics and Molecular Biology*, *4*(1). https://doi.org/doi:10.2202/1544-6115.1175

Song, D., Li, K., Ge, X., & Li, J. J. (2023). ClusterDE: a post-clustering differential expression (DE) method robust to false-positive inflation caused by double dipping. *Biorxiv*. https://doi.org/10.1101/2023.07.21.550107