ClusterDE: a post-clustering differential expression method

Carson Zhang

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

Cell-type annotation

Motivation

Understanding which types of cells are in a data sample allows an analyst to better make use of existing knowledge about those cells. "Cell annotation" is the process of labeling cells in a sample of data. In this paper, the focus is on annotating the "cell type" of each cell: a cellular phenotype that is robust across datasets (Heumos et al. 2023). For example, plasma B cells are one type of white blood cell that are involved in the human body's immune response by secreting antibodies (Heumos et al. 2023). T cells are another type of white blood cell that are also involved in the immune response (Green et al. 2024). They produce cytokines, which are signaling proteins that activate other parts of the human immune system. A scientist interested in a patient's immune response may be interested in the counts of B cells and T cells (and their subtypes): for example, in order to better understand the roles of each cell, or how they affect patient outcomes. Cell-type annotation is required in order to obtain this information from e.g. a blood sample.

Cell-type markers

(TODO: other methods of annotation)

Differential expression testing

Differential expression testing is the primary method by which scientists identify marker genes. If genes are

(define validity)

(define FDR)

(mention FDR control like Benjamini-Hochberg)

TODO: mention Scanpy, Seurat, and their default methods

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.

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.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.

The double-dipping issue

False discoveries

James et al. (2021)

ClusterDE

Summary of steps

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. Separately for each dataset, cluster the cells into two groups.

3. Separately for each dataset, perform differential expression testing between the two groups from step 2.

4. Combine the results to determine which genes to output as discoveries (DE genes). Rafi and Greenland (2020)

TODO: similarity to the S-value.

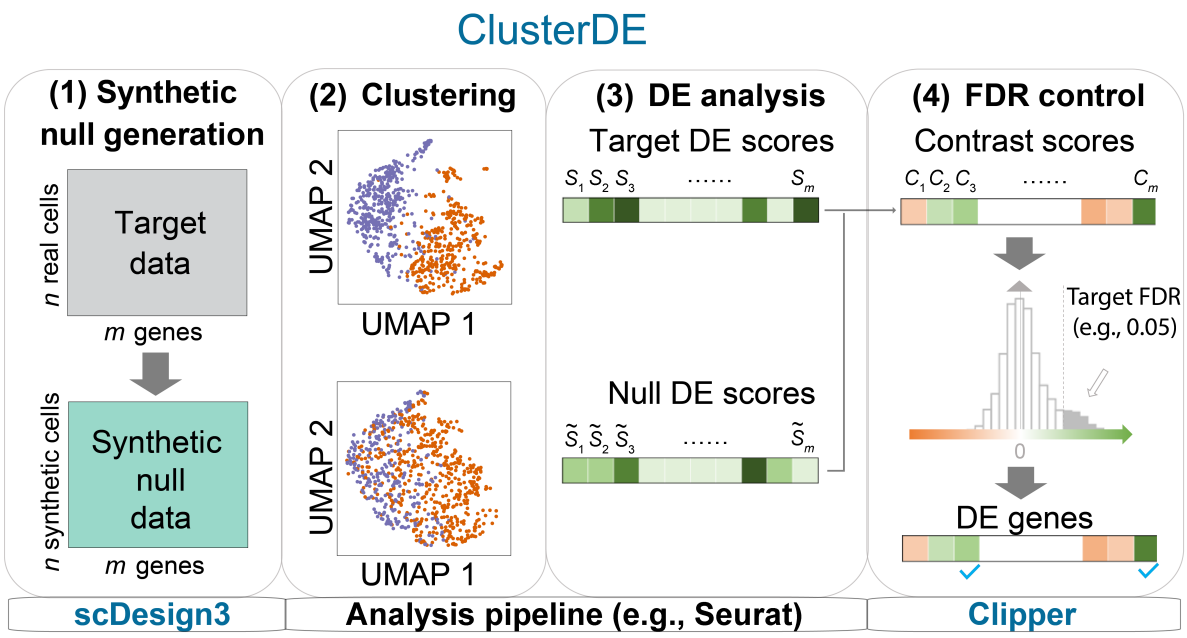

Figure 1: A visual overview of the ClusterDE method. In step 1, a negative control dataset is generated. In step 2, a clustering algorithm is applied to each dataset. In step 3, a differential expression test is performed for each gene, computing a DE score for each gene in each dataset. In step 4, the difference in results is computed as a contrast score, and Clipper is used to choose a minimum contrast score for the true DE genes outputted by ClusterDE.

Step 1

Idea: negative control

Idea: copulas

To actually generate this negative control data, (Song et al. 2023) use the copula approach. Special methods are required to simulate data from the desired multivariate negative binomial distribution, as statistical packages such as `R` do not come with samplers already implemented.

Thus, ClusterDE uses the copula generator implemented in scDesign3 (Song et al. 2024) for its *in silico* negative control data.

Theorem (Probability Integral Transform): $F_X(X) \sim \text{Uniform}(0, 1)$.

Intuition for the PIT

Takeaway

Theorem (Sklar's Theorem): Let $\mathbf{X}$ be a m-dimensional random vector with joint cumulative distribution function F and marginal distribution functions $F_j, j = 1, ..., m$. The joint CDF can be expressed as

$$F(x_1, ..., x_m) = C(F_1(y_1), ..., F_m(y_m))$$

with associated probability density (or mass) function

$$f(y_1, ..., y_m) = c(F_1(y_1), ..., F_m(y_m))f_1(y_1)...f_m(y_m)$$

for a m-dimensional copula C with copula density c.

The inverse also holds: the copula corresponding to a multivariate CDF F with marginal distribution functions $F_j, j = 1, ..., m$ can be expressed as

$$C(u_1, ..., u_m) = F(F_1^{-1}(u_1), ..., F_m^{-1}(u_m))$$

, and the copula density (or mass) function is

$$c(u_1, ..., u_m) = \frac{f(F_1^{-1}(u_1), ..., F_m^{-1}(u_m))}{f_1(F_1^{-1}(u_1))...f_m(F_m^{-1}(u_m))}$$

.

Takeaway

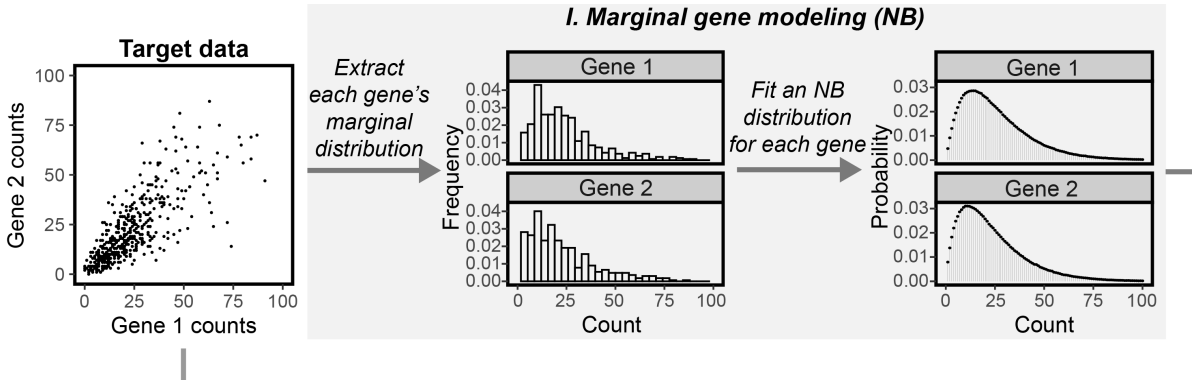

Figure 2: Fit a marginal distribution for each gene. The copula approach allows us to separately model the marginal distributions.

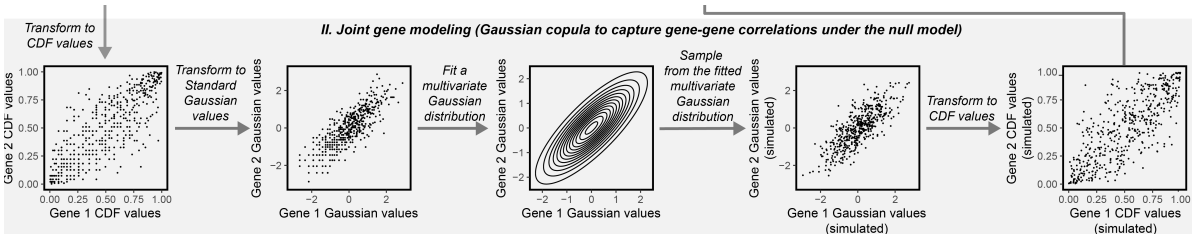

Figure 3: Estimate the covariance matrix for the m-dimensional gene distribution. The copula approach allows us to model the correlations between genes separately from their marginal distributions.

Note: compositional nature

Step 2

Scanpy, Seurat defaults

Step 3: choice of tests

Step 4: false discovery rate control using Clipper

In Step 4, ClusterDE uses the Clipper method to choose which discoveries from step 3 to output as true DE genes.

Intuition

Given that the negative control generated in step 1 accomplished its goal, the two datasets should be similar, and therefore the p-values (and DE scores) outputted by each test should be similar. This means that, when a test on a given gene has a very low p-value, but this p-value is similar across both datasets, it is reasonable to believe that this low p-value occurred due to noise. However, when a p-value is much lower in the real data than in the synthetic null data, this indicates that the gene is truly differentially expressed between the two clusters.

Clipper

Symmetry assumption for contrast scores

In practice, the symmetry assumption may be violated.

Differential expression methods that address double-dipping

Count splitting

TN Test

Practical notes on ClusterDE usage

How to handle multiple clusters

How to decide whether to merge clusters

Whether you should cluster once or twice

```
# library(flextable)
# library(tinytable)
# library(readr)
```

```
# table_e1 = read_csv("../seminar_paper-bacsc/python/reproduce_results/table_e1.csv")
# flextable(table_e1)
#
# table_e1_synthetic = read_csv("../seminar_paper-bacsc/python/synthetic_null_generation/tabl
# flextable(table_e1_synthetic)
```

Performance of ClusterDE

- Against other DE methods
- Against other null generation strategies

Data analysis

BacSC data

Synthetic null data generation

Schäfer-Strimmer

Results

Simulation study

Heumos et al. (2023) Benjamini and Hochberg (1995)

Appendix

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300. http://www.jstor.org/stable/2346101.

Green, Eric, Lawrence Brody, Sarah Bates, Mary Beth Gardiner, Jill Thomas, Britny Kish, Darryl Leja, et al. 2024. "Lymphocyte." https://www.genome.gov/genetics-glossary/Lymphocyte.

Heumos, Lukas, Anna C. Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D. Lücken, et al. 2023. "Best Practices for Single-Cell Analysis Across Modalities." *Nature Reviews Genetics* 24 (8): 550–72. https://doi.org/10.1038/s41576-023-00586-w.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning with Applications in r*. Second. Springer Texts in Statistics.

Rafi, Zad, and Sander Greenland. 2020. "Semantic and Cognitive Tools to Aid Statistical Science: Replace Confidence and Significance by Compatibility and Surprise." *BMC Medical Research Methodology* 20 (1): 244. https://doi.org/10.1186/s12874-020-01105-9.

Song, Dongyuan, Kexin Li, Xinzhou Ge, and Jingyi Jessica Li. 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.

Song, Dongyuan, Qingyang Wang, Guanao Yan, Tianyang Liu, Tianyi Sun, and Jingyi Jessica Li. 2024. "scDesign3 Generates Realistic in Silico Data for Multimodal Single-Cell and Spatial Omics." *Nature Biotechnology* 42 (2): 247–52. https://doi.org/10.1038/s41587-023-01772-1.