ClusterDE: a post-clustering differential expression method

Carson Zhang

Contributing authors: 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. Furthermore, investigation of the covariance

Table of contents

Introduction 2
- Cell-type annotation 2
- Differential expression testing 3

The double-dipping issue 3
- Toy example illustrating double dipping 3
- False discoveries 3

ClusterDE 3
- Summary of steps 3
- Notation 4

Step 1: synthetic null generation . . . 4
Step 2: clustering . . . 6
Step 3: DE testing . . . 6
Step 4: false discovery rate control using Clipper . . . 7
Differential expression methods that address double-dipping . . . 7
Count splitting . . . 7
TN Test . . . 7
"Traditional" FDR control methods . . . 7
Considerations for using ClusterDE in practice . . . 7
Symmetry assumption for contrast scores . . . 7
How to handle multiple clusters . . . 8
How to decide whether to merge clusters . . . 8
Whether you should cluster once or twice . . . 8
Performance of ClusterDE . . . 8
Data analysis . . . 8
BacSC data . . . 8
Synthetic null data generation . . . 8
Schäfer-Strimmer . . . 8
Results . . . 8
Simulation study . . . 8
Discussion . . . 8
Appendix . . . 9
References . . . 9

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 immune response. They produce "cytokines, which are signaling proteins that activate other parts of the human immune system" (Green et al. 2024). 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.

TODO: distinction between different types of T cells requires better annotation methods.

Cell-type markers

https://academic.oup.com/nar/article/51/D1/D870/6775381

http://bio-bigdata.hrbmu.edu.cn/CellMarker/

https://www.nature.com/articles/s41467-022-28803-w

(TODO: other methods of annotation) (see book)

Differential expression testing

Differential expression testing is the primary method by which scientists identify marker genes. If gene A is differentially expressed across two conditions,

(define validity)

(define FDR)

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.

The double-dipping issue

Mention Scanpy, Seurat defaults

Mention warnings but lack of solutions

Toy example illustrating double dipping

False discoveries

James et al. (2021)

ClusterDE

Summary of steps

The ClusterDE method consists of four basic steps, summarized in Figure 1.

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).

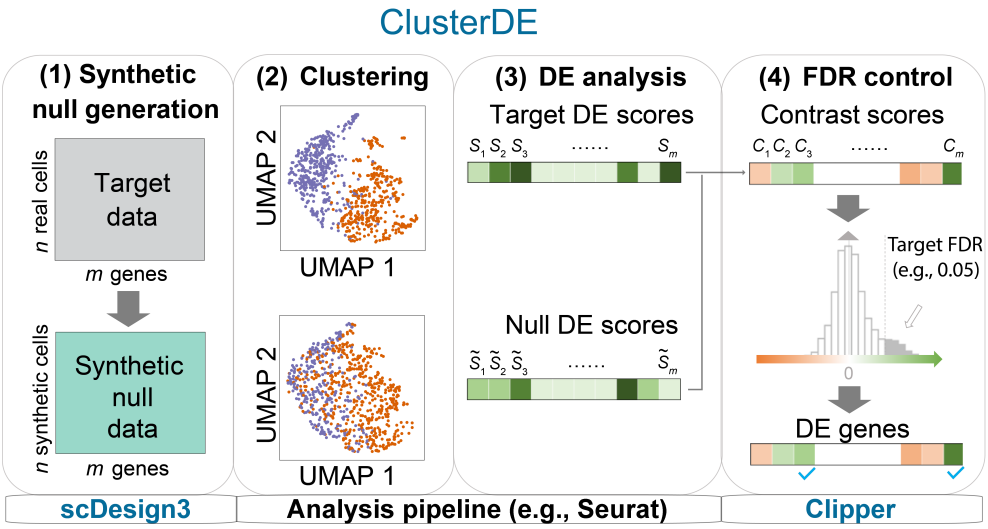

Figure 1: A visual overview of the ClusterDE method. In step 1, a negative control dataset is generated. In step 2, a clustering algorithm is applied to each dataset. In step 3, a differential expression test is performed for each gene, computing a DE score for each gene in each dataset. In step 4, the 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.

Notation

$$\begin{bmatrix} Y_{11} & \cdots & Y_{1m} \\ \vdots & \ddots & \\ Y_{n1} & & Y_{nm} \end{bmatrix}$$

$$\begin{bmatrix} Y_{11} & \cdots & Y_{1m} & Z_1 \\ \vdots & \ddots & & \vdots \\ Y_{n1} & & Y_{nm} & Z_n \end{bmatrix}$$

Step 1: synthetic null generation

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-based sampler implemented in scDesign3 (Song et al. 2024) for its *in silico* negative control data: that is, data that was created by a computer (Ekins, Mestres, and Testa 2007).

Probability Integral Transform.

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

(A more rigorous proof and discussion can be found in Theorem 2.1.10 in Casella and Berger (Casella and Berger 2001).)

Intuition for the PIT.

Takeaway: if we can compute F^{-1}, we can move freely between a standard uniform random variable and a random variable with distribution F. Sklar's Theorem, and therefore the copula approach to modeling multivariate distributions, relies on this result.

Sklar's Theorem.

(The theorem statement is adapted from Czado (Czado 2019).)

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

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

.

Proof: See Nelsen (Nelsen 2006).

Sklar's Theorem allows statisticians to use the copula approach to model the joint distribution: the goal is now to find a copula C that yields a good approximation of F.

ClusterDE makes the popular choice of the Gaussian copula to model the multivariate gene distribution, which is convenient because it has existing software implementations [see mvtnorm, numpy].

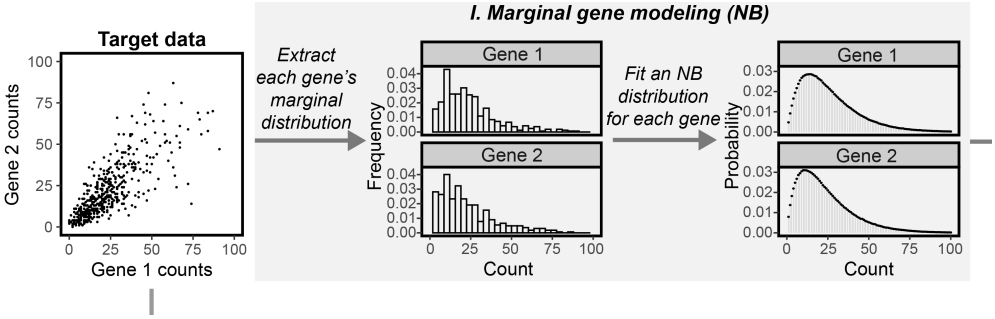

Figure 2: Fit a marginal distribution for each gene. The copula approach allows us to model the marginal distributions separately from the covariance structure of the variables (see Figure 3) (Song, Li, and Chen 2024).

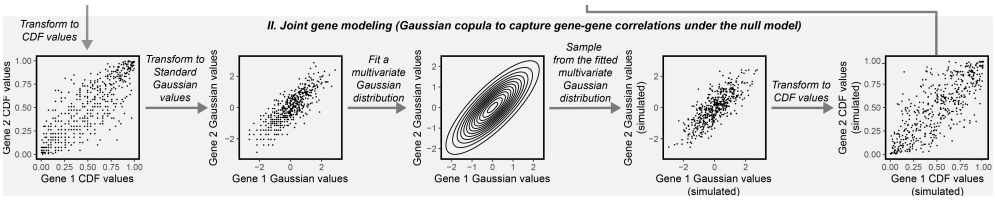

Figure 3: Estimate the covariance matrix for the m-dimensional gene distribution. The copula approach allows us to model the correlations between genes separately from their marginal distributions (see Figure 2) (Song, Li, and Chen 2024).

Note: compositional nature

Step 2: clustering

This is similar to the usual clustering step in differential expression testing.

Scanpy, Seurat defaults

Step 3: DE testing

DE tests are performed as usual on the two clusters.

Define Wilcoxon rank-sum test

Step 4: false discovery rate control using Clipper

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

TODO: similarity to the S-value.

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

Differential expression methods that address double-dipping

Count splitting

TN Test

"Traditional" FDR control methods

(mention FDR control like Benjamini-Hochberg)

Considerations for using ClusterDE in practice

Symmetry assumption for contrast scores

In step 4, the Clipper method for FDR control assumes that the contrast score distribution is symmetric. In practice, this symmetry assumption may be violated. ClusterDE tests the symmetry of the contrast score distribution using Yuen's trimmed mean test: if the test statistic has p-value < 0.001, reject the null hypothesis of symmetry, and perform a contrast score adjustment. It uses a one-sided "greater than" hypothesis for this test: that is, it only adjusts the contrast scores when too few contrast scores are negative. This is because the authors wanted to be conservative with their adjustment strategy, only transforming the contrast scores when they know that there would have been too many false discoveries. When there are too many negative contrast scores, these will not lead to an inflated false discovery rate, since only positive contrast scores become discoveries.

The software implementation comes from the `PairedData` `R` package.

How to handle multiple clusters

How to decide whether to merge clusters

Whether you should cluster once or twice

Performance of ClusterDE

- Against other DE methods
- Against other null generation strategies

Data analysis

Change

BacSC data

We chose to investigate the *Bsub_minmed_PB* dataset. This is a dataset that was generated by ProBac sequencing (ProBac-seq), in order to validate the performance of this method. ProBac-seq uses messenger RNA-specific probes, and multiple probes per organism, to sequence bacterial samples (McNulty et al. 2023), (Samanta et al. 2024). The *Bsub_minmed_PB* dataset contains the *B. subtilis 168* strain (Ostner et al. 2024), "grown to late exponential phase in M9 minimal media supplemented with malate" (McNulty et al. 2023).

Synthetic null data generation

Schäfer-Strimmer

Results

```
# 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/t
# flextable(table_e1_synthetic)
```

Simulation study

Discussion

Different copula generation strategies.

Appendix

References

Casella, George, and Roger Berger. 2001. *Statistical Inference*. Textbook Binding; Duxbury Resource Center.

Czado, Claudia. 2019. *Analyzing Dependent Data with Vine Copulas: A Practical Guide with r*. First. Springer Cham.

Ekins, S, J Mestres, and B Testa. 2007. "In Silico Pharmacology for Drug Discovery: Methods for Virtual Ligand Screening and Profiling." *British Journal of Pharmacology* 152 (1): 9–20. https://doi.org/https://doi.org/10.1038/sj.bjp.0707305.

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.

McNulty, Ryan, Duluxan Sritharan, Seong Ho Pahng, Jeffrey P. Meisch, Shichen Liu, Melanie A. Brennan, Gerda Saxer, Sahand Hormoz, and Adam Z. Rosenthal. 2023. "Probe-Based Bacterial Single-Cell RNA Sequencing Predicts Toxin Regulation." *Nature Microbiology* 8 (5): 934–45. https://doi.org/10.1038/s41564-023-01348-4.

Nelsen, Roger B. 2006. *An Introduction to Copulas*. First. Springer New York, NY.

Ostner, Johannes, Tim Kirk, Roberto Olayo-Alarcon, Janne Gesine Thöming, Adam Z. Rosenthal, Susanne Häussler, and Christian L. Müller. 2024. "BacSC: A General Workflow for Bacterial Single-Cell RNA Sequencing Data Analysis." *bioRxiv*. https://doi.org/10.1101/2024.06.22.600071.

Samanta, Prosenjit, Samuel F. Cooke, Ryan McNulty, Sahand Hormoz, and Adam Rosenthal. 2024. "ProBac-Seq, a Bacterial Single-Cell RNA Sequencing Methodology Using Droplet Microfluidics and Large Oligonucleotide Probe Sets." *Nature Protocols*. https://doi.org/10.1038/s41596-024-01002-1.

Song, Dongyuan, Kexin Li, and Siqi Chen. 2024. *ClusterDE: A Post-Clustering Differential Expression (DE) Method for Solving Double Dipping*. https://github.com/SONGDONGYUAN1994/ClusterDE.

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.