ClusterDE Seminar Paper

Carson Zhang

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

1 Introduction

1.1 Background

1.1.a Biology

TODO: give the basic biology background necessary to understand the paper.

1.1.a.a RNA

RNA carries the genetic information specific in DNA. There are two main types: - **non-coding RNA** performs some biological function - **messenger RNA** forms a template for protein production (it codes for a protein which performs some biological function).

TODO: explain the jump from RNA to the UMI count matrix.

TODO: define UMI.

1.1.b scRNA-seq

TODO: give an explanation of scRNA-seq data collection and analysis.

1.1.c Double-dipping

TODO: explain the double-dipping problem in differential expression analysis.

2 Method

In words, the ClusterDE method can be broken up into four steps:

1. Generate a synthetic null dataset that mimics the structure (in particular, the gene-gene correlation structure) of the original data.
2. Separately partition the synthetic null data and the target data (real data) into two clusters.
3. Separately for the null and target data, perform hypothesis tests for differentially expressed genes between the two clusters. For each gene, compute some sort of difference between the scores on the two datasets.
4. Output a subset of the significant results from step 3 as potential cell-type marker genes.

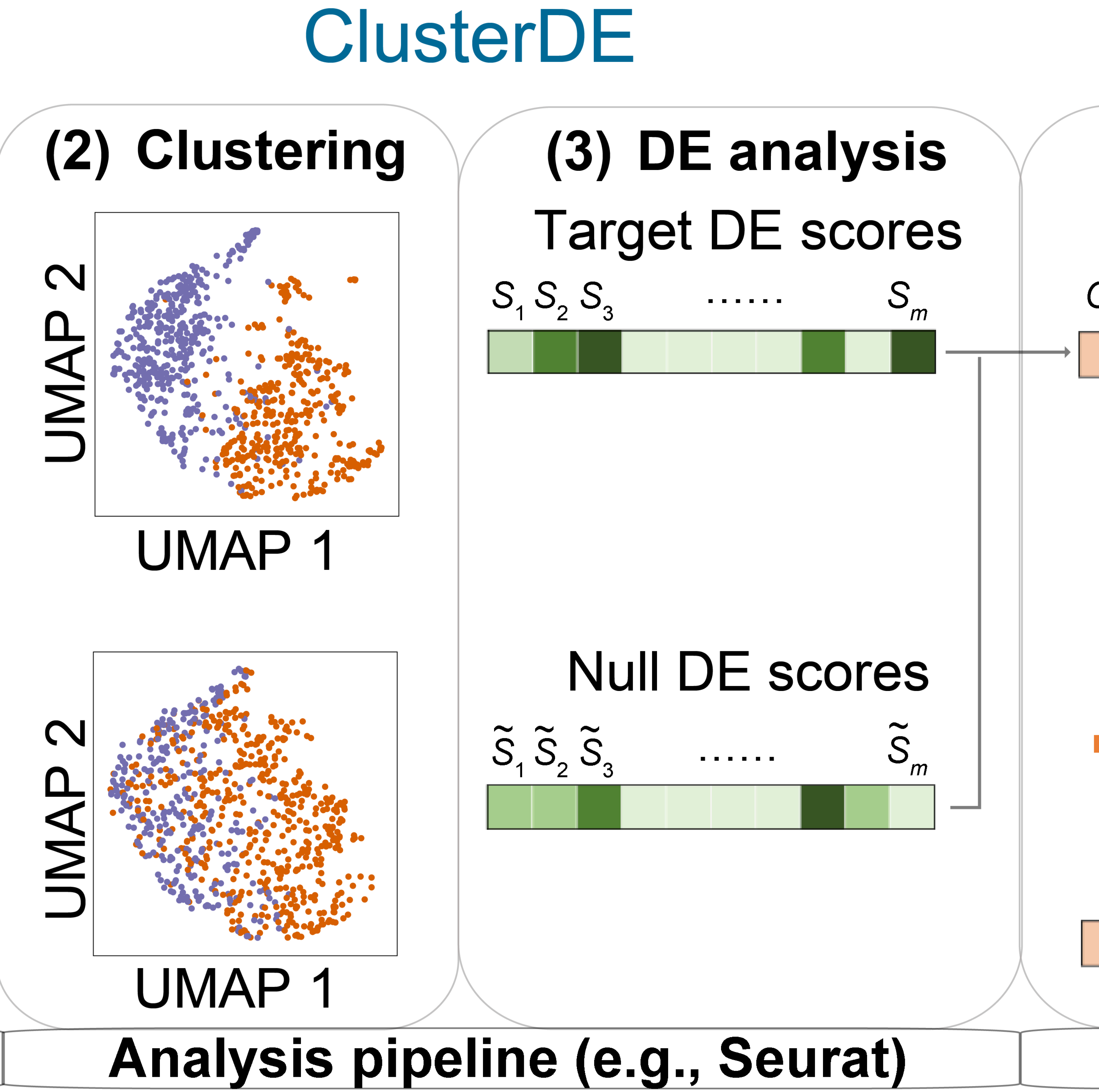
ClusterDE
(2) Clustering
UMAP 2
UMAP 1
UMAP 2
UMAP 1
(3) DE analysis
Target DE scores
S_1 S_2 S_3 ...... S_m
Null DE scores
$\tilde{S}_1$ $\tilde{S}_2$ $\tilde{S}_3$ ...... $\tilde{S}_m$
Analysis pipeline (e.g., Seurat)

Figure 1: An graphical illustration of ClusterDE.

It is important to note that ClusterDE “does not provide an automatic decision about whether two clusters should be merged”. Its outputs are potential DE genes, and therefore it does not directly measure the quality of a given clustering. These potential cell-type marker genes enable researchers to gain biological insights into the clusters, and they empower researchers to further explore the functional and molecular characteristics of the clusters.

2.1 Synthetic null generation

The synthetic null generation consists of three steps, as described in the following figure.

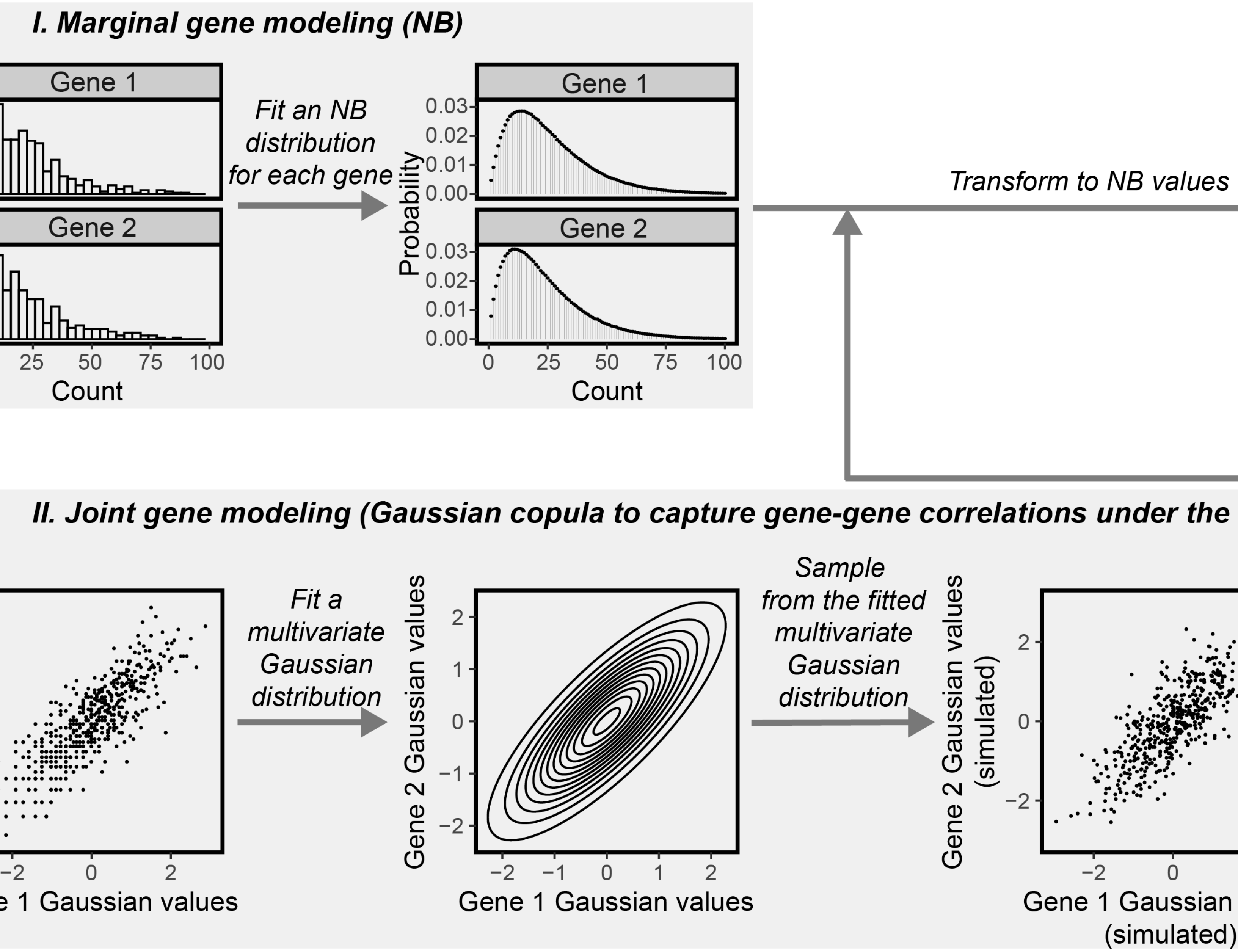
I. Marginal gene modeling (NB)
Gene 1
Gene 2
25
50
75
100
Count
Fit an NB distribution for each gene
Probability
0.03
0.02
0.01
0.00
Gene 1
Gene 2
0
25
50
75
100
Count
Transform to NB values
II. Joint gene modeling (Gaussian copula to capture gene-gene correlations under the
−2
0
2
e 1 Gaussian values
Fit a multivariate Gaussian distribution
Gene 2 Gaussian values
2
1
0
−1
−2
Gene 1 Gaussian values
Sample from the fitted multivariate Gaussian distribution
Gene 2 Gaussian values (simulated)
2
0
−2
Gene 1 Gaussian (simulated)

Figure 2: Null generation steps.

1. Model the null distribution in terms of the Gaussian copula.

Now, our goal is to estimate the parameters $\mu_j, \sigma_{j_{j=1}^m}$ and $\boldsymbol{R}$.

2. Fit the null model to the real data.
3. Sample from the fitted null model.

2.2 2. Clustering

ClusterDE allows any clustering algorithm. Note that it only handles the case of two clusters, so if you started out with more clusters, you should identify a particular pair of interest. In the **Practical guidelines for ClusterDE usage** subsection, steps 1 and 2 describe how an analyst should proceed.

1. Given ≥ 2 clusters, identify 2 clusters of interest. Generally, this will be a pair for which you suspect the clustering is spurious (i.e. you think the two clusters actually come from the same cell type, so they are strong candidates to be merged into a single cluster).
2. Filter the data so that you only consider the subset of cells that come from those two clusters.

TODO: describe the Seurat clustering pipeline.

2.2.a UMAP

UMAP is common.

TODO: summarize UMAP.

2.2.b Louvain

The example analyses in the presentation use the default Seurat clustering procedure, which uses the Louvain algorithm.

TODO: summarize the Louvain algorithm.

2.3 3. DE analysis (testing)

ClusterDE allows any DE test.

TODO: choose and summarize common DE tests.

Let $P_1, ..., P_m$ be the p-values computed by the m DE tests on the target data. Define the target DE score $S_j := -\log_{10} P_j$. Likewise for the synthetic null data.

The final outputs of step 3: m target DE scores $S_1, ..., S_m$; m null DE scores $\tilde{S}_1, ..., \tilde{S}_m$.

2.4 4. FDR control

Given the target and null DE scores, compute a contrast score for gene j as $C_j := S_j - \tilde{S}_j$.

3 Results

3.1 Simulation

3.2 Real data example

4 Appendix