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

- Count splitting

(see presentations)

(define validity)

```
library(flextable)
library(readr)
```

```
table_e1 = read_csv("../seminar_paper-bacsc/python/reproduce_results/
table_e1.csv")
```

```
Rows: 13 Columns: 9
──                            Column                            specification
────────────────────────────────────────────────────────────
Delimiter: ","
dbl (9): Cells, Genes, Minimum seq. depth, Maximum seq. depth, Median seq. d...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
flextable(table_e1)
```

Cells	Genes	Minimum seq. depth	Maximum seq. depth	Median seq. depth	Zero counts (percentage)	Maximum count	95% quantile	99% quantile
1,544	5,553	413	5,704	794.5	0.862	136	1	3
1,255	5,540	360	4,464	647.0	0.881	80	1	2
3,386	3,968	103	495	163.0	0.963	14	0	1
2,784	2,952	141	1,289	325.0	0.911	45	1	2
13,801	2,959	268	1,839	555.0	0.861	110	1	3
6,703	2,937	136	948	267.0	0.940	105	1	2
19,638	2,500	14	275	21.0	0.992	13	0	0
48,511	2,500	12	728	21.0	0.991	30	0	0
9,168	2,500	15	371	45.0	0.990	26	0	0
315	1,265	9	196	19.0	0.978	10	0	1
983	1,301	10	556	21.0	0.981	35	0	1
103	628	8	137	18.0	0.953	7	0	1
2,113	1,606	12	289	22.0	0.985	19	0	1

```
table_e1_synthetic = read_csv("../seminar_paper-bacsc/python/
synthetic_null_generation/table_e1-synthetic.csv")
```

```
Rows: 3 Columns: 9
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
dbl (9): Cells, Genes, Minimum seq. depth, Maximum seq. depth, Median seq. d...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
flextable(table_e1_synthetic)
```

Cells	Genes	Minimum seq. depth	Maximum seq. depth	Median seq. depth	Zero counts (percentage)	Maximum count	95% quantile	99% quantile
2,784	2,952	141	1,289	325	0.911	45	1	2
5,568	2,951	83	1,354	379	0.907	129	1	2
5,568	2,952	80	1,422	340	0.911	81	1	2

Data analysis

BacSC data

Synthetic null data generation

Schäfer-Strimmer

Results

Simulation study

(Heumos et al., n.d.)

Appendix

Bibliography

Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M. D., Strobl, D. C., Henao, J., Curion, F., Aliee, H., Ansari, M., Badia-i-Mompel, P., Büttner, M., Dann, E., Dimitrov, D., Dony, L., Frishberg, A., He, D., … Consortium, S.-c. B. P. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, *24*(8), 550–572. https://doi.org/10.1038/s41576-023-00586-w

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