Synthetic Control Removes Spurious Discoveries from Double Dipping in Single-Cell and Spatial Transcriptomics Data Analyses

Dongyuan Song[1,2], Siqi Chen[3], Christy Lee[3], Kexin Li[3], Xinzhou Ge[3,4], and Jingyi Jessica Li[2,3(✉)]

[1] Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, CT 06032, USA
[2] Interdepartmental Program of Bioinformatics, University of California, Los Angeles, Los Angeles, CA 90095-7246, USA
[3] Department of Statistics and Data Science, University of California, Los Angeles, Los Angeles, CA 90095-1554, USA
[4] Department of Statistics, Oregon State University, Corvallis, OR 97331-4606, USA
jli@stat.ucla.edu

Abstract. Double dipping is a common analytical pitfall in single-cell and spatial transcriptomics data analysis: after a clustering algorithm finds clusters as putative cell types or spatial domains, statistical tests are applied to the same data to identify differentially expressed (DE) genes as potential cell-type or spatial-domain markers. Because the genes that contribute to clustering are inherently likely to be identified as DE genes, double dipping can result in false-positive markers, especially when clusters are spurious, leading to ambiguously defined cell types or spatial domains. To address this challenge, we propose ClusterDE, a statistical method designed to identify post-clustering DE genes as reliable markers of cell types and spatial domains, while controlling the false discovery rate (FDR) regardless of clustering quality.

Keywords: Single-cell RNA-seq · Spatial Transcriptomics · Clustering · Differential Expression Analysis · Marker Gene Identification

1 Introduction

A key task in single-cell RNA-seq (scRNA-seq) and spatially resolved transcriptomics (SRT) data analysis is the annotation of cell types or spatial domains using marker genes. This process typically involves first clustering cells or spatial spots into putative cell types or spatial domains, followed by differential expression (DE) analysis to identify genes that are highly expressed in each

D. Song, S. Chen and C. Lee—These authors contributed equally to this work.

S. Sankararaman (Ed.): RECOMB 2025, LNBI 15647, pp. 400–404, 2025.
https://doi.org/10.1007/978-3-031-90252-9_48

cluster. However, it has been realized that this post-clustering DE procedure is conceptually flawed, with one important issue being "double dipping"—using the same gene expression data twice: first to find clusters and then to identify DE genes. This leads to an inflated false discovery rate (FDR) when identifying post-clustering DE genes as putative cell-type or spatial-domain marker genes.

To explain this double-dipping issue in scRNA-seq cell-type annotation, we discuss two extreme scenarios. If two cell types are distinct and the inferred clusters are accurate, double dipping has no impact on post-clustering DE analysis, and cell-type marker genes can be successfully identified (Fig. 1a, top). In contrast, if a single cell type is over-clustered, post-clustering DE analysis will falsely highlight genes as DE, as the same expression variations used for clustering also drive DE detection (Fig. 1a, bottom). While methods like the Truncated Normal (TN) test [4] and Countsplit [2] attempt to address this issue, they do not work well on real scRNA-seq data where genes are correlated.

The double-dipping issue in post-clustering DE analysis also affects SRT data analysis, particularly in spatial domain detection, where proximal spatial spots are clustered, aiming to identify functionally distinct tissue structures. This issue can lead to unreliable marker genes for "indistinct" domains lacking sharp gene expression changes across boundaries. To tackle this, we define spatial domain marker genes as those with sharp expression changes at domain boundaries, while non-marker genes may still show spatial variation, but in a smooth manner (Fig. 1b). Similar to scRNA-seq data analysis, our goal is to enable post-clustering DE analysis to identify reliable marker genes for annotating distinct spatial domains.

Here, we introduce ClusterDE, a post-clustering DE method designed to identify potential cell-type or spatial-domain marker genes while avoiding the inflated FDR caused by double dipping. Additionally, ClusterDE offers a practical advantage by allowing users to interpret an abstract statistical null hypothesis through concrete synthetic null data (i.e., *in silico* negative controls).

2 Methods

ClusterDE introduces a novel synthetic control and contrastive approach to identify reliable cell-type or spatial-domain genes that are robust to double dipping. The approach centers on establishing an *in silico* negative-control data (referred to as the "synthetic null data") to be analyzed in parallel with the real data (referred to as the "target data"). The contrastive approach identifies reliable cell-type or spatial-domain marker genes by comparing DE analysis results from the target data to those from the synthetic null data. To generate the synthetic null data, ClusterDE includes two null models: a scRNA-seq null model, assuming a homogeneous cell type with unimodal gene expression, and an SRT null model, assuming a homogeneous spatial domain where genes show smooth expression variation. Under each null model, no marker genes are expected to be detected. Leveraging these null models, ClusterDE employs four main steps to identify potential cell-type or spatial-domain marker genes (Fig. 1c).

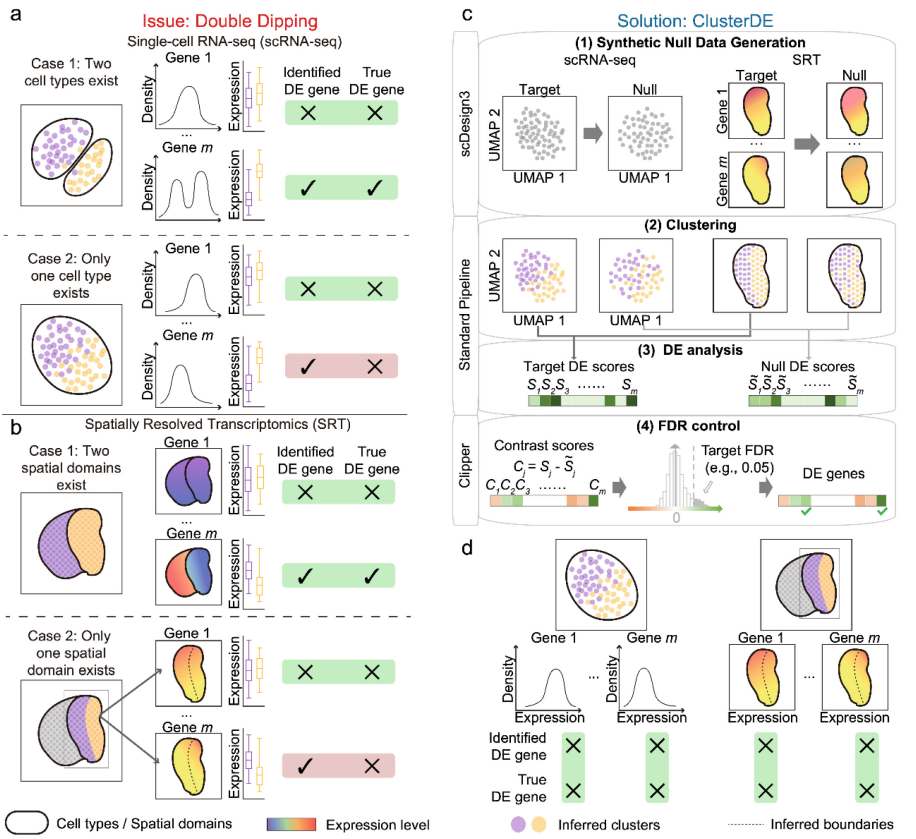

Fig. 1. ClusterDE mitigates double dipping in post-clustering DE analysis for identifying cell-type and spatial-domain marker genes. a–b, An illustration of the double-dipping issue in (**a**) scRNA-seq and (**b**) SRT post-clustering DE analysis. **c**, An overview of ClusterDE steps. **d**, ClusterDE mitigates the false discoveries caused by double dipping in Case 2 shown in **a** and **b**.

Step 1 of ClusterDE is synthetic null data generation, where the statistical simulator scDesign3 [3] is used to generate the synthetic null data that represents a hypothetical homogeneous cell type or spatial domain. The synthetic null data preserves per-gene means, variances, and gene-gene correlations of the target data, while maintaining the same number of cells or spots and the same genes.

Steps 2 and 3 of ClusterDE comprise a user-defined pipeline for clustering and subsequent DE analysis, allowing flexibility in choosing the pipeline to analyze the target data and the synthetic null data in parallel. These two steps yield a "target DE score" and a "null DE score" for each gene, which are summary statistics that quantify the significance of the gene's expression difference between two clusters in the target data and the synthetic null data, respectively.

By default, the DE score is defined as the negative logarithm of the P value from a DE test that compares a gene's expression levels between two clusters.

Step 4 of ClusterDE implements a contrastive strategy to compare each gene's target and null DE scores, identifying a gene as a reliable marker only if its target DE score significantly exceeds its null DE score. Specifically, ClusterDE computes a "contrast score" for each gene by subtracting its null DE score from its target DE score. True non-marker genes are expected to have contrast scores symmetrically distributed around zero. ClusterDE uses Clipper [1] to determine a contrast score cutoff based on a target FDR (e.g., 0.05). Genes with contrast scores equal to or exceeding the cutoff are identified as DE genes.

Through these four steps, ClusterDE effectively eliminates false-positive marker genes caused by double dipping, particularly when the target data consists of a single cell type or spatial domain (Fig. 1d).

3 Results and Conclusion

ClusterDE demonstrates strong performance in both scRNA-seq and SRT post-clustering DE analysis. In scRNA-seq applications, ClusterDE was benchmarked against the Seurat pipeline (which includes double dipping), the TN test, and Countsplit, and was shown to be the only method that effectively controls the FDR when the target data consists of a single cell type. ClusterDE effectively avoids false-positive cell-type markers, including housekeeping genes, and prioritizes relevant canonical markers in datasets from five cell lines and peripheral blood mononuclear cells (PBMC). Additionally, ClusterDE outperforms Seurat in distinguishing cell types in an adult *Drosophila* dataset. When extended to SRT post-clustering DE analysis, ClusterDE also outperforms the double-dipping approach (BayesSpace [5] for spatial clustering and Seurat for DE tests), demonstrating effective FDR control. ClusterDE successfully identifies no DE genes between spurious spatial clusters in a human brain tissue SRT dataset and a human pancreas cancer tissue SRT dataset. Moreover, the top marker genes identified by ClusterDE for spatial domains, which align well with annotated cancer regions, exhibit distinctly higher expression in these regions compared to the genes identified by the double-dipping approach.

In conclusion, ClusterDE effectively addresses the double-dipping issue in post-clustering DE analysis of scRNA-seq and SRT data. ClusterDE adapts well to a wide range of clustering algorithms and DE tests, effectively avoiding false discoveries caused by double dipping and identifying biologically meaningful marker genes. The R package ClusterDE is available at https://github.com/SONGDONGYUAN1994/ClusterDE. The source code and data for reproducing the results are available at: http://doi.org/10.5281/zenodo.8161964.

Full Paper: A preprint of the full paper is available on bioRxiv at https://www.biorxiv.org/content/10.1101/2023.07.21.550107v2.

References

1. Ge, X., et al.: Clipper: p-value-free FDR control on high-throughput data from two conditions. Genome Biol. **22**(1), 1–29 (2021)
2. Neufeld, A., Gao, L.L., Popp, J., Battle, A., Witten, D.: Inference after latent variable estimation for single-cell RNA sequencing data. Biostatistics **25**(1), 270–287 (2024)
3. Song, D., Wang, Q., Yan, G., Liu, T., Sun, T., Li, J.J.: scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics. Nat. Biotechnol. **42**(2), 247–252 (2024)
4. Zhang, J.M., Kamath, G.M., David, N.T.: Valid post-clustering differential analysis for single-cell RNA-seq. Cell Syst. **9**(4), 383–392 (2019)
5. Zhao, E., et al.: Spatial transcriptomics at subspot resolution with BayesSpace. Nat. Biotechnol. **39**(11), 1375–1384 (2021)