Synthetic control removes spurious discoveries from double dipping in single-cell and spatial transcriptomics data analyses

Dongyuan Song[1,2,†][0000−0003−1114−1215], Siqi Chen[3,†][0000−0002−4955−3832],
Christy Lee[3,†][0009−0002−6803−8948], Kexin Li[3][0000−0002−4955−3832],
Xinzhou Ge[3,4][0000−0002−8229−5716], and Jingyi Jessica Li[2,3,5,6,7][0000−0002−9288−5648]

[1] Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, CT 06030
[2] Interdepartmental Program of Bioinformatics, University of California, Los Angeles, CA 90095-7246
[3] Department of Statistics and Data Science, University of California, Los Angeles, CA 90095-1554
[4] Department of Statistics, Oregon State University, Corvallis, OR 97331-4606
[5] Department of Human Genetics, University of California, Los Angeles, CA 90095-7088
[6] Department of Computational Medicine, University of California, Los Angeles, CA 90095-1766
[7] Department of Biostatistics, University of California, Los Angeles, CA 90095-1772
[8] †: These authors contributed equally to this work

Abstract. Double dipping is a well-known 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 cell-type or spatial-domain 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. The core of ClusterDE involves generating synthetic null data as an *in silico* negative control that contains only one cell type or spatial domain, allowing for the detection and removal of spurious discoveries caused by double dipping. We demonstrate that ClusterDE controls the FDR and identifies canonical cell-type and spatial-domain markers as top DE genes, distinguishing them from housekeeping genes. ClusterDE's ability to discover reliable markers, or the absence of such markers, can be used to determine whether two ambiguous clusters should be merged. Additionally, ClusterDE is compatible with state-of-the-art analysis pipelines like Seurat and Scanpy.

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

1 Introduction

Single-cell RNA-seq (scRNA-seq) and spatially resolved transcriptomics (SRT) technologies have revolutionized transcriptomic studies by providing unprecedented snapshots of gene expression within individual cells and in a spatial context, respectively. A major task in scRNA-seq and spatial transcriptomics data analysis is to annotate cell types and spatial domains with marker genes. For example, in scRNA-seq data analysis, a standard annotation procedure includes two steps: (1) partitioning cells into clusters, and (2) finding differentially expressed (DE) genes between cell clusters as marker genes for annotating clusters as potential cell types [1, 2, 3, 4], as in state-of-the-art analysis pipelines such as the R package Seurat [5] and the Python module Scanpy [6]. However, it has been realized that this post-clustering differential expression (DE) procedure is conceptually flawed. For instance, Seurat contains the warning message that "P values should be interpreted cautiously, as the genes used for clustering are the same genes tested for differential expression." This issue is commonly referred to as "double dipping," meaning that the same gene expression data are used twice: first to find cell clusters and then to identify DE genes. The double dipping issue leads to an inflated false discovery rate (FDR) when identifying post-clustering DE genes as putative cell-type marker genes. The FDR inflation is more pronounced when the cell clusters are more ambiguous.

To explain this double-dipping issue in scRNA-seq cell-type annotation, we discuss two extreme scenarios. In one scenario, if two cell types are considered "distinct" (i.e., cell-type marker genes exhibit bimodal expression distributions, with the high-expression mode corresponding to the cell type in which they are highly expressed) and the inferred clusters are accurate, then double-dipping has no impact on post-clustering DE analysis. In this case, DE analysis can successfully identify cell-type marker genes that are highly expressed in a cluster corresponding to a distinct cell type (Fig. 1a, top). In the opposite scenario, when a clustering algorithm over-clusters a single cell type into two clusters, post-clustering DE analysis will identify genes highly expressed in each cluster due to double-dipping: the clustering is based on gene expression data, and variations in the expression of certain genes drive the clustering, resulting in these genes being highly expressed in one of the clusters (Fig. 1a, bottom). However, since only one true cell type exists, these post-clustering DE genes cannot be used to justify the two clusters as distinct cell types and should not be regarded as cell-type marker genes.

Some biologists might argue that even if the two clusters are “indistinct” (i.e., all genes exhibit unimodal expression distributions across all cells), they could still reflect meaningful biological variation and thus represent potential cell sub-types. While this argument may hold in certain scenarios, we argue that if the two clusters are indistinct, it is impossible to determine whether they should be classified as two distinct cell types or merely arbitrary divisions along a continuum of cell state changes, without external knowledge. A more critical issue is that if clusters are not required to be distinct, cells could be continuously divided into finer and finer clusters, which would make it unreasonable to consider any such clusters as potential cell types. This would also hinder the transfer of cell type labels from one dataset to another. Given this issue and the need for a clear definition of cell types in cell atlases—reflected by the ongoing debate about what constitutes a cell type [7, 8]—we focus, in this work, on the practice of using post-clustering DE analysis to annotate distinct cell clusters as potential cell types.

Methods aimed at addressing the double-dipping issue in scRNA-seq post-clustering DE analysis include the Truncated Normal (TN) test [9] and the Countsplit method [10]. The TN test method relies on splitting cells and consists of five steps: (1) dividing the cells into two sets—training and test cells; (2) applying a clustering algorithm to the training cells to identify two clusters; (3) training a support vector machine (SVM) classifier on the training cells to predict cluster labels based on gene expression; (4) using the trained classifier to predict the cluster labels of the test cells; and (5) identifying differentially expressed (DE) genes between the two clusters in the test cells using the TN test. Since the TN test method requires gene expression levels to follow normal distributions, it applies the log-transformation to the scRNA-seq count data before the five steps. Instead of splitting cells, the Countsplit method splits the scRNA-seq cell-by-gene count matrix into training and test matrices of the same dimensions (cells and genes) using a procedure called data thinning [11], which splits each count into two counts, whose sum equals the original count, through binomial sampling. Countsplit identifies cell clusters by applying a clustering algorithm to the training matrix. Since the test and training matrices contain the same cells, the cell cluster assignments can be directly applied to the test matrix. Countsplit then identifies DE genes between the clusters based on the test matrix, allowing the use of any suitable statistical test for count data, beyond the TN test. Although both methods claimed to provide well-calibrated P values (i.e., uniformly distributed between 0 and 1 under the null hypotheses), our findings indicate that their P values are anti-conservative (i.e., enriched near 0) in the presence of gene-gene correlations (see **Results**). The reason behind this issue is that the validity check of P values in the TN test and Countsplit papers relied on simulation studies that implicitly assumed genes to be independent [10], an assumption that, however, does not hold in real scRNA-seq data. As a result of their anti-conservative P values, both methods still lead to inflated FDRs when applied to real scRNA-seq data.

Several cluster-free DE tests attempt to circumvent the double-dipping issue by bypassing the cell clustering step [12, 13, 14, 15, 16, 17]. However, it is important to note that these cluster-free methods are not designed to identify potential cell types. Consequently, the DE genes identified by these methods cannot be interpreted as marker genes for specific cell types, unlike the DE genes identified post-clustering. In other words, cluster-free DE genes and post-clustering DE genes serve different purposes and are not conceptually comparable. Another class of methods aim to assess the quality of clustering results, such as evaluating the “purity” of a cluster or determining if two clusters should be merged [18, 19, 20, 21, 22, 23, 24, 25]. However, these methods lack a direct statistical test for identifying DE genes and rely on setting a threshold for clustering quality—an arbitrary decision to mitigate concerns about double dipping. In this study, we focus on identifying reliable marker genes for cell clusters during the cell-type annotation process. Hence, we do not consider cluster-free DE tests and clustering assessment methods as competing alternatives in our investigation.

The double-dipping issue in post-clustering DE analysis is not unique to scRNA-seq data; it is also relevant for the more recent spatially resolved transcriptomics (SRT) data, which captures tissue context information of cells or cell neighborhoods, depending on the spatial resolution of the SRT technology used [26]. A common analysis task is spatial domain detection, which involves clustering proximal spatial spots—representing cells or cell neighborhoods—based on their gene expression profiles, with the goal of identifying spatial domains that correspond to functionally distinct tissue structures [27]. Although not widely discussed, the double-dipping issue can lead to unreliable marker genes for "indistinct" spatial domains, where no genes show sharp expression changes at the domain boundaries [28]. However, we argue that reliable spatial domains should be "distinct" from nearby domains, meaning that the domain's marker genes should exhibit sharp expression changes at the boundaries. Otherwise, similar to the challenge of defining cell types in scRNA-seq data, indistinct spatial domains cannot have their labels reliably transferred across datasets. Despite this challenge, no methods have been developed to circumvent the double-dipping issue in spatial domain annotation. Motivated by the need for annotating distinct spatial domains, we define spatial domain marker genes as those exhibiting discontinuity or sharp changes in expression at domain boundaries (Fig. 1b, top). In contrast, genes that are not defined as spatial domain markers may still show spatial variation, but the variation is smooth (Fig. 1b, bottom). 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. ClusterDE effectively controls the FDR when identifying marker genes, even in cases where cell or spatial clusters are spurious. The ultimate goal is to use the identified marker genes to reliably annotate distinct cell types or spatial domains. It is important to note that ClusterDE is not intended to replace existing pipelines that perform cell or spatial clustering followed by DE analysis (e.g., Seurat). Instead, it functions as an add-on to any current pipeline, enabling more reliable discoveries. 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). This feature enables users to assess whether the synthetic null data accurately reflects the negative control scenario they envision and, if necessary, adjust the generation of the synthetic null data.

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. 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 for spatial clustering and Seurat for DE), 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. Hence, ClusterDE is an effective tool for scRNA-seq and SRT data analysis, enabling the reliable identification of cell-type or spatial-domain marker genes that can effectively distinguish cell types or spatial domains.

2 Results

2.1 ClusterDE: a synthetic control and contrastive approach to address double-dipping

ClusterDE introduces a novel synthetic control and contrastive approach to identify reliable cell-type and spatial-domain genes that are robust to double dipping (clustering followed by DE analysis). 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") through a computational pipeline, which includes cell or spatial clustering followed by DE analysis. The contrastive approach identifies reliable cell-type and spatial-domain marker genes by comparing DE analysis results from the target data to those from the synthetic null data. The target data consists of cells or spatial spots divided into two potentially ambiguous clusters, which are then analyzed by DE analysis to identify cell-type or spatial-domain marker genes. To generate the synthetic null data, ClusterDE includes two null models: an scRNA-seq null model, which assumes that the cells in the target data belong to a homogeneous cell type, with gene expression exhibiting unimodal distributions across the cells; and an SRT null model, which assumes that the spatial spots in the target data belong to a homogeneous domain, with gene expression varying smoothly within the domain. Under the null models, no cell-type or spatial-domain marker genes are expected to be detected. To identify potential cell-type or spatial-domain marker genes, ClusterDE involves four main steps (Fig. 1c).

Step 1 of ClusterDE is synthetic null data generation, where the statistical simulator scDesign3 [29] is used to generate the synthetic null data that represents a hypothetical homogeneous cell type or spatial domain. The synthetic null data preserves key statistics of the target data, including gene means, variances, and, importantly, gene-gene correlations, while maintaining the same number of cells or spatial spots and the same set of genes (Figs. 2a and S2 for scRNA-seq; Figs. S3 and S4 for SRT). Details of the generation process for scRNA-seq and SRT data are provided in "Supplementary Methods" in the Supplementary Material.

Steps 2 and 3 of ClusterDE comprise a user-defined pipeline for clustering and subsequent DE analysis, with examples including Seurat in R or Scanpy in Python. ClusterDE offers flexibility, allowing users to choose their preferred clustering algorithms and DE tests to analyze the target data and synthetic null data in parallel. These two steps yield a "target DE score" and a "null DE score" for each gene. Specifically, we define a gene's DE score as a summary statistic that quantifies the significance of the gene's expression difference between two clusters in the target data or the synthetic null data; a higher DE score indicates that the gene is more likely DE. In the default setting of ClusterDE, the DE score is defined as the negative logarithm of the P value obtained from a DE test (e.g., the Wilcoxon rank-sum test) that compares a gene's expression levels between two clusters.

Finally, Step 4 of ClusterDE implements a contrastive strategy to compare each gene's target DE score and null DE score. A gene is identified as a reliable cell-type or spatial-domain marker only if its target DE score significantly exceeds its null DE score. To implement this identification strategy, ClusterDE computes "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 the FDR control method Clipper [30] 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. Regarding why ClusterDE can control the FDR when the number of genes is large, a mathematical proof is provided in "Theoretical justification of ClusterDE" in the Supplementary Material.

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).

2.2 Simulations verify ClusterDE's reliable FDR control and statistical power

We conducted extensive simulation studies to validate ClusterDE as a post-clustering DE method with reliable FDR control under double dipping. We also compared ClusterDE with Seurat, the most widely used analysis pipeline that involves double dipping, and two methods designed to address double dipping—the TN test [9] and Countsplit [10]. For all four methods, we used the default Louvain clustering algorithm in Seurat, as it is commonly employed in scRNA-seq data analyses. In the DE analysis step for ClusterDE, Seurat, and Countsplit, we evaluated five DE tests available in Seurat: the Wilcoxon rank-sum test (Wilcoxon; the default in Seurat), the two-sample t-test (t-test), the negative binomial generalized linear model (NB-GLM), logistic regression (LR), and the likelihood-ratio test (bimod). The TN test was an exception, as it requires its own DE test. Since Seurat, Countsplit, and the TN test all produce a P value for each gene, we applied the Benjamini-Hochberg (BH) procedure to determine a P value cutoff for a given target FDR (e.g., 0.05). Genes with P values at or below the cutoff were identified as DE genes.

In the first simulation setting, we simulated target data from a single cell type, mimicking naïve cytotoxic T cells in a PBMC scRNA-seq dataset [31] (Fig. 2a, top left; see "Simulation designs" in the Supplementary Material), representing the most severe double-dipping scenario. Since the data consist of only a single cell type, any identified DE genes are false discoveries. At a target FDR of 0.05, all three existing methods—Seurat, Countsplit, and the TN test—failed to control the actual FDR below 0.05 (Fig. 2b). Seurat, which employs a double-dipping approach, performed the worst, with all five DE tests yielding actual FDRs of 1, indicating that Seurat consistently identified false DE genes across 200 simulation runs. Although

Countsplit and the TN test were designed to mitigate double dipping, their actual FDRs still far exceeded 0.05 due to anti-conservative P values in the presence of gene-gene correlations (Fig. S5, middle to right), despite their validity in idealized settings with independent genes, as demonstrated in their own simulation studies [10]. In contrast, ClusterDE successfully controlled the FDR below 0.05 for three of the five DE tests: Wilcoxon, t-test, and LR (Fig. 2b), supported by contrast scores that satisfied the required symmetry around zero (Fig. S5, left; "Theoretical justification of ClusterDE" section of Supplementary Material). While ClusterDE did not control the FDR below 0.05 for the NB-GLM and bimod tests, likely due to violations of their parametric modeling assumptions, the FDR inflation for these tests was substantially less severe than that observed with Countsplit. Specifically, ClusterDE's actual FDRs were 0.28 and 0.16 for NB-GLM and bimod, respectively, compared to Countsplit's actual FDRs of 0.68 and 1 for the same tests (Fig. 2b).

In the second simulation setting, we generated datasets with varying degrees of double-dipping, again mimicking naïve cytotoxic T cells from the same PBMC scRNA-seq dataset [31] (Fig. 2c, top; see "Simulation designs" in the Supplementary Material). Each simulated dataset consisted of two cell types with equal numbers of cells and 200 pre-specified true DE genes with varying expression level differences between the cell types, summarized as the log fold change (logFC), where larger logFC values indicate greater distinctions between cell types. After applying Seurat's default clustering algorithm to each dataset to identify two clusters, the agreement between the identified clusters and the true cell types was assessed using the adjusted Rand index (ARI), with lower ARI values corresponding to more severe double-dipping scenarios, as visualized by UMAP (Fig. 2c, top row). For DE analysis, we used Wilcoxon, the default DE test in Seurat, as it provided the best FDR control for both ClusterDE and Countsplit, while the TN test employed its own DE test. Results show that ClusterDE consistently controlled FDRs across a range of target thresholds under varying degrees of double dipping, whereas Seurat, Countsplit, and the TN test failed to control FDRs under the target thresholds and exhibited greater FDR inflation as the severity of double dipping increased (Fig. 2c, middle row). Notably, ClusterDE achieved comparable or superior statistical power relative to Seurat, Countsplit, and the TN test at equivalent actual FDR levels (Fig. 2c, bottom). These findings held across four other DE tests (t-test, NB-GLM, LR, and bimod) when used with ClusterDE, Seurat, and Countsplit (Fig. S6). To further evaluate performance in realistic scenarios with unbalanced cell type sizes, we simulated datasets with cell type size ratios of 1:4 and 1:9, where ClusterDE still outperformed the other three methods in FDR control across target thresholds and varying degrees of double-dipping. Specifically, ClusterDE consistently demonstrated robust FDR control and comparable or superior statistical power to the other methods when using Wilcoxon as the DE test (Figs. S6–S8).

Technically, ClusterDE and knockoff methods share the concept of controlling the FDR by generating *in silico* negative control data [32], but their applications differ significantly. Knockoff methods are designed to identify important features in high-dimensional predictive models within a supervised-learning framework, in contrast to the unsupervised-learning setting addressed by ClusterDE. Broadly, knockoff methods generate features uncorrelated with the outcome variable given the other features while preserving feature-feature correlations. To assess their applicability to our unsupervised learning setting, we applied the default model-X knockoffs method [33] to the target data from the aforementioned simulation, treating genes as features and the cell cluster label as the outcome variable. While the method controlled the FDR, it consistently yielded zero statistical power, making it impractical for DE gene identification. Furthermore, we compared three alternative strategies for synthetic null data generation in Step 1 of ClusterDE: the model-X knockoffs method, a permutation-based approach (where genes are independently permuted across all cells to create a homogeneous cell population), and a variant of ClusterDE using scDesign3 without copula modeling [29] (assuming gene independence). Fig. S9 shows that ClusterDE's synthetic null data effectively preserved per-gene mean and variance statistics, as well as gene-gene correlations. In contrast, the model-X knockoffs method failed to maintain gene mean and variance statistics, while the permutation-based approach and the ClusterDE variant (labeled as "scDesign3 Ind"), as expected, did not preserve gene-gene correlations. As a result, only ClusterDE's synthetic null data effectively mimics the target data while creating a single hypothetical cell type (Fig. S9). Overall, results on simulated datasets demonstrate that ClusterDE achieved the most robust FDR control and the best statistical power among the four strategies for synthetic null generation (Fig. S10).

To address concerns about the randomness involved in generating synthetic null data—a process that relies on random sampling from the null model fitted to the target data—we conducted an analysis to evaluate the stability of DE genes identified by ClusterDE. The results demonstrate that the DE genes identified by

ClusterDE are relatively stable and robust to this randomness (Fig. S11). In summary, these simulation studies confirm that ClusterDE is an effective method that controls FDR under varying degrees of double dipping while maintaining strong statistical power.

2.3 ClusterDE distinguishes cell-type marker genes from housekeeping genes

We applied ClusterDE to multiple scRNA-seq datasets to demonstrate its ability to enhance the reliability of cell-type marker genes identified through post-clustering DE analysis.

In the first application, we aimed to demonstrate that ClusterDE can avoid false-positive cell-type marker gene discovery when the cells belong to only one type, not more than one. To this end, we analyzed five datasets of pure cell lines commonly used as gold standards of a "single cell type" in benchmarking studies [34, 35] (Fig. 3a, left). While biological heterogeneity within a cell line may include variations in cell cycle, differentiation, or subclonal populations, a cell line is generally not considered to be composed of distinct cell types [36, 37]. Accordingly, any post-clustering DE genes identified from a cell line should not be interpreted as between-cell-type DE genes. Hence, these cell-line datasets served as negative cases to demonstrate the inflated FDRs of existing methods and the ability of ClusterDE to eliminate this inflation. As a sanity check, we verified that the synthetic null data generated by ClusterDE resembled the target data (Fig. 3a, right). Applying ClusterDE, Seurat, Countsplit, and the TN test to the five datasets, we found that all methods except ClusterDE identified thousands of DE genes, often exceeding 50% of all genes, indicating severe FDR inflation at the target threshold of 5%. In contrast, ClusterDE identified zero DE genes in 22 out of 25 combinations of the five datasets and five DE tests (Wilcoxon, t-test, NB-GLM, LR, and bimod). Notably, ClusterDE consistently identified zero DE genes in all five datasets when using Wilcoxon, leading us to designate Wilcoxon as the default DE test for ClusterDE.

In the second application, we analyzed eight PBMC datasets of CD14$^+$/CD16$^+$ monocytes [38] to evaluate ClusterDE's ability to detect known or potential marker genes for these two monocyte subtypes. The datasets were generated from two technical replicates using four UMI-based scRNA-seq protocols (10X Genomics Versions 2 and 3, Drop-seq, and inDrop). After applying the default Seurat clustering to identify two clusters in each dataset, we found that four datasets exhibited relatively accurate clustering results (ARI > 0.5; Fig. 3c, left; Fig. S12, top), while the other four showed poor agreement between clusters and the two monocyte subtypes (ARI < 0.2; Fig. S12, bottom). We anticipated that an effective post-clustering DE method would be able to detect meaningful marker genes for monocyte subtypes in the first four datasets with high-quality clusters, but not necessarily in the latter four datasets with low-quality clusters. The reason is that low-quality clusters do not accurately reflect the true distinction between the two cell subtypes, making it unreasonable to expect reliable subtype marker genes to be identified by comparing these clusters. When applied to the latter four datasets with low-quality clusters, Seurat, Countsplit, and the TN test often identified hundreds to thousands of DE genes (Fig. S13). In contrast, ClusterDE did not identify any DE genes in most cases (Fig. S13), yielding a more reasonable result.

On the first four datasets with high-quality clusters (Fig. S12, top), which align well with the two monocyte subtypes, we expected post-clustering DE genes to reveal cell-subtype markers. As a sanity check, we verified that ClusterDE's synthetic null data resembled the target data but filled the "gap" between CD14$^+$ and CD16$^+$ monocytes, representing a single "hypothetical" cell type in each dataset (Fig. 3c, right). When applied to these four datasets, ClusterDE with Wilcoxon identified between 55 and 173 DE genes (corresponding to 1–4% of all genes) at the target FDR of 5%. In contrast, Seurat and Countsplit identified between 1 and 1,288 DE genes (0–26% of all genes), while the TN test consistently identified at least 1,187 DE genes (25% of all genes) (Fig. 3d). Given the knowledge that the two monocyte subtypes are not drastically different, identifying thousands of potential marker genes seemed implausible. Thus, the number of DE genes identified by ClusterDE appeared more reasonable and aligned with biological expectations.

Examining the post-clustering DE genes identified by ClusterDE and Seurat across the five DE tests on the four datasets (resulting in 20 DE gene lists for each method), we found that ClusterDE more effectively distinguished known subtype marker genes from housekeeping genes compared to Seurat. For each dataset, both methods performed post-clustering DE analysis on the same two clusters identified by Seurat's default clustering, with ClusterDE acting as an add-on to address the double-dipping issue in Seurat's post-clustering DE results. The distinction in the two methods' results was evident in the ranking of specific genes within the DE gene lists. For instance, we focused on *FCGR3A* (*CD16*), a canonical marker for distinguishing CD14$^+$ monocytes from CD16$^+$ monocytes, and *B2M*, a widely recognized housekeeping gene expressed across various cell types [39]. Notably, ClusterDE consistently ranked *FCGR3A* among its top DE genes (typically ranked between 1 and 10) while placing *B2M* consistently low in its DE gene lists (generally ranked below 1,000) (Fig. 3e, top). In contrast, Seurat ranked the two genes similarly (with ranks between 10 and 100), making it difficult to discern which gene was more likely to be a subtype marker without prior knowledge (Fig. 3e, bottom). Using the dataset "Rep2_10x(V2)" as an example, we examined the five most frequently identified post-clustering DE genes (determined based on the top 50 DE genes identified by each of the five DE tests) by ClusterDE and Seurat (Fig. 3f). Our analysis revealed that the five genes identified by ClusterDE displayed more distinct distributions of normalized expression levels between the two clusters compared to the five genes identified by Seurat, suggesting that the genes identified by ClusterDE are more likely to be subtype markers (Fig. 3f).

Further, we performed gene set enrichment analysis (GSEA) to evaluate the enrichment of two gene sets—known CD14$^+$/CD16$^+$ monocyte markers and housekeeping genes—in the post-clustering DE gene lists identified by ClusterDE and Seurat. GSEA revealed a strong enrichment of monocyte subtype markers among the top-ranked DE genes identified by ClusterDE, with a clear distinction from housekeeping genes (Fig. 3g, top). In contrast, monocyte subtype markers exhibited a weaker enrichment in Seurat's top-ranked DE genes and showed a similar enrichment pattern to that of housekeeping genes, making it difficult to distinguish the two gene sets in Seurat's ranked DE gene list (Fig. 3g, bottom). GSEA results from the other three datasets further confirmed that ClusterDE consistently outperformed Seurat in distinguishing monocyte subtype markers from housekeeping genes (Fig. S14).

Considering the common practice of using only the top k DE genes (e.g., $k = 100$) for further investigation, we summarized the numbers of monocyte subtype markers and housekeeping genes among the top $k = 1$ to 100 DE genes identified by ClusterDE and Seurat across the five DE tests on the four datasets. Fig. S15 shows that ClusterDE consistently identified more monocyte markers and fewer housekeeping genes among the top DE genes compared to Seurat. To further explain why ClusterDE better distinguished monocyte markers from housekeeping genes, we used the minus-average (MA) plots [40] to illustrate the role of the synthetic null data as a contrast to exclude housekeeping genes from the top DE genes. In the MA plots (Fig. S16), we observed that four exemplary housekeeping genes (*ACTB*, *ACTG1*, *B2M*, and *GAPDH*, marked in blue in Fig. S16) exhibited both large target DE scores and large null DE scores, resulting in close-to-zero contrast scores. Consequently, these genes were excluded from the top DE genes by ClusterDE. In contrast, Seurat identified these housekeeping genes as top DE genes solely based on their large target DE scores. On the other hand, we examined four exemplary monocyte markers (*CD14*, *FCGR3A*, *MS4A7*, and *LYZ*, marked in red in Fig. S16) and found that they had large target DE scores but small null DE scores, allowing ClusterDE to correctly identify them as top DE genes.

In the third application, we used a scRNA-seq atlas of the adult *Drosophila* neuronal visual system [41] to demonstrate ClusterDE's usage for guiding the merging of spurious clusters caused by over-clustering. The original study clustered the data using the Louvain algorithm at a resolution of 10, resulting in 208 clusters, and identified spurious neuron clusters using a random forest-based method in Seurat. We applied ClusterDE to three pairs of spurious neuron clusters flagged as over-clustered and subsequently merged in the original study (Fig.4a). For all three pairs, ClusterDE identified zero DE genes, supporting the original merging decisions, while Seurat (using Wilcoxon as the DE test) identified hundreds of DE genes at the target FDR of 5% (Fig.4b). The gating plots, commonly used in single-cell literature to assess the separation of clusters based on marker gene expression, also indicate that each of the three pairs of neuron clusters should be merged (Fig. 4c), consistent with the conclusion that these pairs of clusters were spurious. Additionally, we examined three specific genes—*alrm*, *nrv2*, and *Gs2*—that Seurat identified as DE in at least two of the three pairs of neuron clusters. These genes are glial cell markers [42, 43], unlikely to be expressed in neurons, and more plausibly attributed to contamination from ambient glial RNA. Hence, ClusterDE successfully identified cases of over-clustering and avoided false-positive marker gene identifications between spurious clusters.

We further analyzed the Tm1 and Tm2 cell types in the same dataset to assess the reliability of cell-type marker genes identified by ClusterDE and Seurat. ClusterDE identified 105 DE genes, while Seurat identified 768, including the 105 genes found by ClusterDE and an additional 663 genes. To evaluate the ability of these genes to distinguish the two cell types, we trained binary classifiers using random forest and support vector machine with three gene sets as features: the 105 DE genes identified by ClusterDE, the top 105 DE genes uniquely identified by Seurat (ranked by p-value), and a random selection of 105 genes as a baseline. The binary classifiers were tested on held-out cells (20% of the dataset), revealing significantly higher classification accuracy with genes identified by ClusterDE compared to those from Seurat (P value $< 2.2 \times 10^{-16}$, two-sided t-test, $n = 100$ train-test splits). Moreover, the accuracy improvement over the random baseline was substantially greater for ClusterDE's DE genes. Gene expression heatmaps further supported this finding, showing that DE genes identified by ClusterDE exhibited more distinct expression differences between Tm1 and Tm2 cell types compared to those identified solely by Seurat. GSEA revealed that the DE genes identified by ClusterDE were significantly enriched in pathways specific to the visual system, including G protein-coupled receptor (GPCR) activity and eye development pathways, consistent with the known roles of Tm1 and Tm2 cell types. Furthermore, the enrichment of "positive regulation of response to stimulus" in Tm2 aligns with the knowledge that Tm1 exhibits a delayed response compared to Tm2 [44]. In contrast, DE genes uniquely identified by Seurat were enriched in more general pathways applicable to many cell types (Fig. S17). These analyses demonstrate that ClusterDE identifies more biologically relevant and reliable cell-type marker genes.

2.4 ClusterDE reveals reliable spatial-domain marker genes

In spatial domain detection and annotation, spatial clustering (e.g., using BayesSpace [45]) is initially applied to SRT data to infer putative spatial domains (referred to as "spatial clusters"). Then, DE tests are conducted to identify potential domain marker genes, leading to a double-dipping issue similar to that in scRNA-seq data analysis. We define "spatial-domain marker genes" as those exhibiting discontinuous expression changes between adjacent domains while being more highly expressed in one domain compared to the others (Fig.S18). In contrast, other genes may still display spatial variation but change smoothly across adjacent domain boundaries (Fig.S18).

Both scRNA-seq and spatial clustering rely on the gene expression profiles of cells or spots; however, spatial clustering algorithms also incorporate the spatial coordinates of spots. This incorporation creates spatial dependency among spots within clusters, which ClusterDE accounts for in the generation of synthetic null data (see "Supplementary Methods" in the Supplementary Material). Using both simulated and real data, we demonstrated how ClusterDE effectively resolves the double-dipping problem in post-spatial clustering DE analysis, facilitating the identification of reliable and interpretable spatial-domain marker genes.

We first simulated an SRT dataset with a single spatial domain (see "Supplementary Methods" in the Supplementary Material) and applied ClusterDE alongside a common SRT post-clustering DE analysis pipeline, which included the popular spatial clustering algorithm BayesSpace [45] to infer spatial domains, followed by two commonly used DE tests—the Wilcoxon rank-sum test and t-test in Seurat—to identify spatial-domain marker genes (referred to as BayesSpace+Seurat). Since the dataset contained only one domain, no true domain marker genes should be detected by this post-clustering DE analysis. However, only ClusterDE met this expectation, whereas BayesSpace+Seurat identified hundreds of false-positive spatial-domain marker genes, failing to control the target FDR of 0.05 (Fig. S19a). Further examination of the P values obtained from BayesSpace+Seurat revealed an anti-conservative bias, whereas the corresponding contrast scores from ClusterDE exhibited symmetry around zero, satisfying the symmetry requirement (Fig. S19b), which explained the observed differences in FDR control. In summary, in this single spatial domain scenario, only ClusterDE successfully controlled the FDR and avoided reporting spurious spatial-domain marker genes.

We next simulated two datasets with multiple domains: (1) a two-domain dataset containing 200 pre-specified true domain marker genes and (2) a three-domain dataset containing 200 and 168 pre-specified true domain marker genes for each pair of adjacent domains. These true marker genes defined distinct domains with clear boundaries in the target datasets; these domains were approximately captured by spatial clusters (Fig. S19c for two domains; Fig. 5a for three domains). Since ClusterDE compares two spatial clusters at a time, we analyzed each pair of adjacent clusters separately in the three-domain case—specifically, Layer 5 vs. Layer 6 and Layer 6 vs. WM. In both the two- and three-domain scenarios, BayesSpace+Seurat failed to control the target FDR of 0.05 (Fig. S19d for two domains; Fig. S20 for three domains). In contrast, ClusterDE successfully controlled the FDR while maintaining high power to detect most true domain marker genes in both datasets (Fig. S19d for two domains; Fig. S20 for three domains). To further assess the reliability of the identified spatial-domain marker genes, we visualized the top three genes identified by ClusterDE and

BayesSpace+Seurat, respectively. The top genes detected by ClusterDE exhibited sharp expression changes at the spatial domain boundaries, whereas those identified by BayesSpace+Seurat showed no such distinct changes (Fig. 5b). In conclusion, these simulation experiments demonstrated that ClusterDE effectively controlled FDR while reliably identifying spatial-domain marker genes.

We applied ClusterDE to real spatially resolved transcriptomics (SRT) datasets to verify its performance in identifying spatial-domain marker genes, analyzing slices from the human dorsolateral prefrontal cortex (DLPFC) dataset [46] and the human primary pancreatic cancer (PDAC) dataset [47].

In the DLPFC dataset, we first applied ClusterDE and BayesSpace+Seurat to Layers 1, 4, and 5. Based on manual annotations, each of these layers represents a homogeneous domain; therefore, any detected DE genes within a layer would be false-positive domain marker genes. Using BayesSpace, we identified two clusters within each layer and applied both methods to compare these clusters. ClusterDE identified nearly zero DE genes, while BayesSpace+Seurat detected hundreds to thousands of DE genes (Fig. S21). This result highlighted the double-dipping issue and demonstrated ClusterDE's effectiveness in addressing it.

Next, we used BayesSpace (in its default setting, without constraining the number of clusters per layer) to perform spatial clustering across the entire DLPFC slice. Several layers (annotated spatial domains) were divided into smaller clusters. Focusing on Layers 3 and WM, we applied ClusterDE and BayesSpace+Seurat to pairs of clusters within each domain. In both cases, ClusterDE did not identify DE genes, supporting the merging of these clusters and demonstrating its ability to correct over-clustering (Fig. S22). In contrast, BayesSpace+Seurat identified DE marker genes between most pairs of clusters, potentially leading to the misidentification of spurious spatial domains due to over-clustering. These findings demonstrate that ClusterDE effectively avoids identifying spurious domain marker genes and prevents the annotation of false spatial domains, whereas BayesSpace+Seurat, due to double-dipping issues, cannot mitigate these problems.

In the PDAC dataset (Fig. 5c), we examined pairs of adjacent spatial domains (duct epithelium vs. interstitium, and interstitium vs. cancer regions) to evaluate the quality of spatial-domain marker genes identified by ClusterDE and BayesSpace+Seurat. The top marker genes detected by ClusterDE exhibited clearer domain-specific patterns compared to those identified by BayesSpace+Seurat (Fig. 5d). For instance, the gene *CRP*, identified by ClusterDE, was exclusively expressed in the duct epithelium region. In contrast, the top marker genes identified by BayesSpace+Seurat did not align with specific domains, instead showing smooth expression changes across the entire slice. We also examined the DE ranks of four known marker genes and two housekeeping genes. ClusterDE ranked the known marker genes *VMA1*, *POSTN*, *SFRP2*, and *F3* higher and assigned lower ranks to the housekeeping genes *TMSB10* and *S100A6* compared to BayesSpace+Seurat (Fig. S23). These results underscore ClusterDE's effectiveness in detecting spatial-domain marker genes that represent spatial domains by exhibiting sharp expression changes at domain boundaries.

3 Discussion

In conclusion, ClusterDE effectively addresses the double-dipping issue in post-clustering DE analysis of scRNA-seq and SRT data. One practical consideration is the scalability of ClusterDE, which primarily depends on the speed of synthetic null data generation. Our results show that this process can be completed within a reasonable timeframe (Fig. S24). Notably, the generation process is fully parallelized, enabling significant reductions in computational time with the use of additional CPUs (Fig. S24).

Another consideration is the power of ClusterDE, which is influenced by the conservative nature of the synthetic null data. The null model assumes a single cell type or spatial domain, which is necessary for controlling over-clustering when the target data contains only one homogeneous cell type or spatial domain. However, when the target data consists of two distinct cell types or spatial domains, this conservative null can lead to power loss. Despite this limitation, the goal of ClusterDE is to annotate distinct cell types or spatial domains, and the top DE genes identified by ClusterDE are typically sufficient for this purpose. Improving the power of ClusterDE while maintaining control of false discoveries is an important area for future research.

ClusterDE adapts well to a wide range of clustering algorithms and DE tests. Through extensive simulation studies and real data analyses, we demonstrated that ClusterDE effectively avoids false discoveries caused by double-dipping and identifies biologically meaningful cell-type or spatial-domain marker genes. For post-clustering DE analysis involving more than two clusters, we recommend using ClusterDE in a stepwise manner, potentially following a hierarchical organization of clusters based on their similarities (Fig. S25). In this approach, users compare pairs of ambiguous clusters at each step, using identified DE genes to determine whether the clusters are biologically meaningful and should remain distinct. For SRT data with multiple spatial clusters, this stepwise method would involve comparisons between spatially adjacent clusters.

Finally, while ClusterDE addresses the double-dipping problem specifically in post-clustering DE analysis, the concept of synthetic null data (*in silico* negative controls) has broader potential applications. It could be extended to other analyses affected by double-dipping, such as post-pseudotime DE analysis [48] and data integration analyses. Since double-dipping is an almost inevitable challenge in single-cell and spatial omics data analysis, we propose a general strategy to reduce false discoveries by implementing synthetic null data and applying a contrastive approach. This strategy can help achieve more reliable findings across various analyses.

Data Availability

All datasets used in the study are publicly available. The pre-processed datasets are available at https://figshare.com/articles/dataset/ClusterDE_datasets/23596764.

Code Availability

The R package ClusterDE is available at https://github.com/SONGDONGYUAN1994/ClusterDE. The tutorials of ClusterDE are available at https://songdongyuan1994.github.io/ClusterDE/index.html. The source code and data for reproducing the results are available at: http://doi.org/10.5281/zenodo.8161964 [49].

Competing interests

The authors declare no competing interests.

Acknowledgements

The authors appreciate the comments and feedback from the Junction of Statistics and Biology members at UCLA (http://jsb.ucla.edu).

Funding

This work was supported by the following grants: National Science Foundation DBI-1846216 and DMS-2113754, NIH/NIGMS R01GM120507 and R35GM140888, Johnson & Johnson WiSTEM2D Award, Sloan Research Fellowship, UCLA David Geffen School of Medicine W. M. Keck Foundation Junior Faculty Award, and the Chan-Zuckerberg Initiative Single-Cell Biology Data Insights Grant (to J.J.L.). J.J.L. was a fellow at the Radcliffe Institute for Advanced Study at Harvard University in 2022-2023 while she was writing this paper.

4 Figures

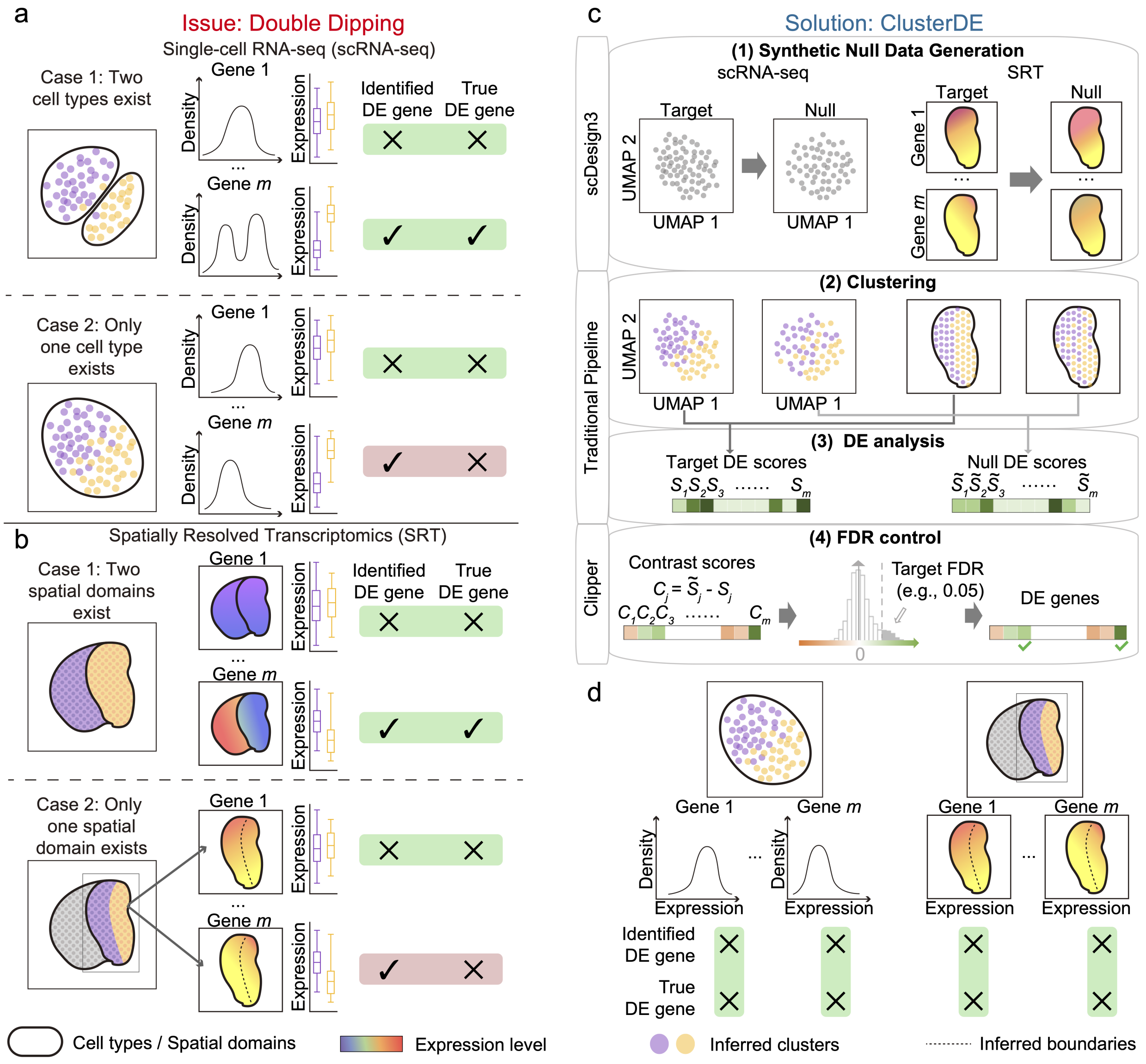


Fig. 1: **ClusterDE mitigates double dipping in post-clustering DE analysis for identifying cell-type and spatial-domain marker genes. a-b**, An illustration of the double-dipping issue in (**a**) scRNA-seq and (**b**) SRT post-clustering DE analysis. In case 1, where two cell types or spatial domains exist and are accurately identified through clustering, double dipping has minimal effect, and the identified post-clustering DE genes align well with the true cell-type or spatial-domain marker genes (top row in panels **a** and **b**). However, in case 2, where a single cell type or spatial domain is incorrectly split into two clusters, double dipping results in post-clustering DE genes that are false-positive cell-type or spatial-domain marker genes (bottom row in panels **a** and **b**). **c**, An overview of ClusterDE, consisting of four steps. Step (1): Given the target real data ("Target"), ClusterDE generates synthetic null data ("Null") using scDesign3 [29] to represent the null hypothesis of a single "hypothetical" cell type (scRNA-seq) or spatial domain (SRT). Steps (2)–(3): A clustering algorithm is used to divide cells or spatial spots into two clusters, followed by a DE test to compare each gene between the clusters. This process is applied to both the target data and the synthetic null data, producing two DE scores per gene—one from each dataset. Step (4): ClusterDE calculates a contrast score for each gene by subtracting the synthetic null DE score from the target DE score. It then applies the FDR-control method Clipper [30] to determine a contrast score cutoff based on all genes' contrast scores and a target FDR (default: 0.05). DE genes are identified as those with contrast scores exceeding this cutoff, as indicated by the vertical dashed line in the contrast score distribution across genes. **d**, ClusterDE mitigates the false discoveries caused by double dipping in case 2 shown in **a** and **b** (e.g., gene m is no longer identified as a false-positive marker gene).

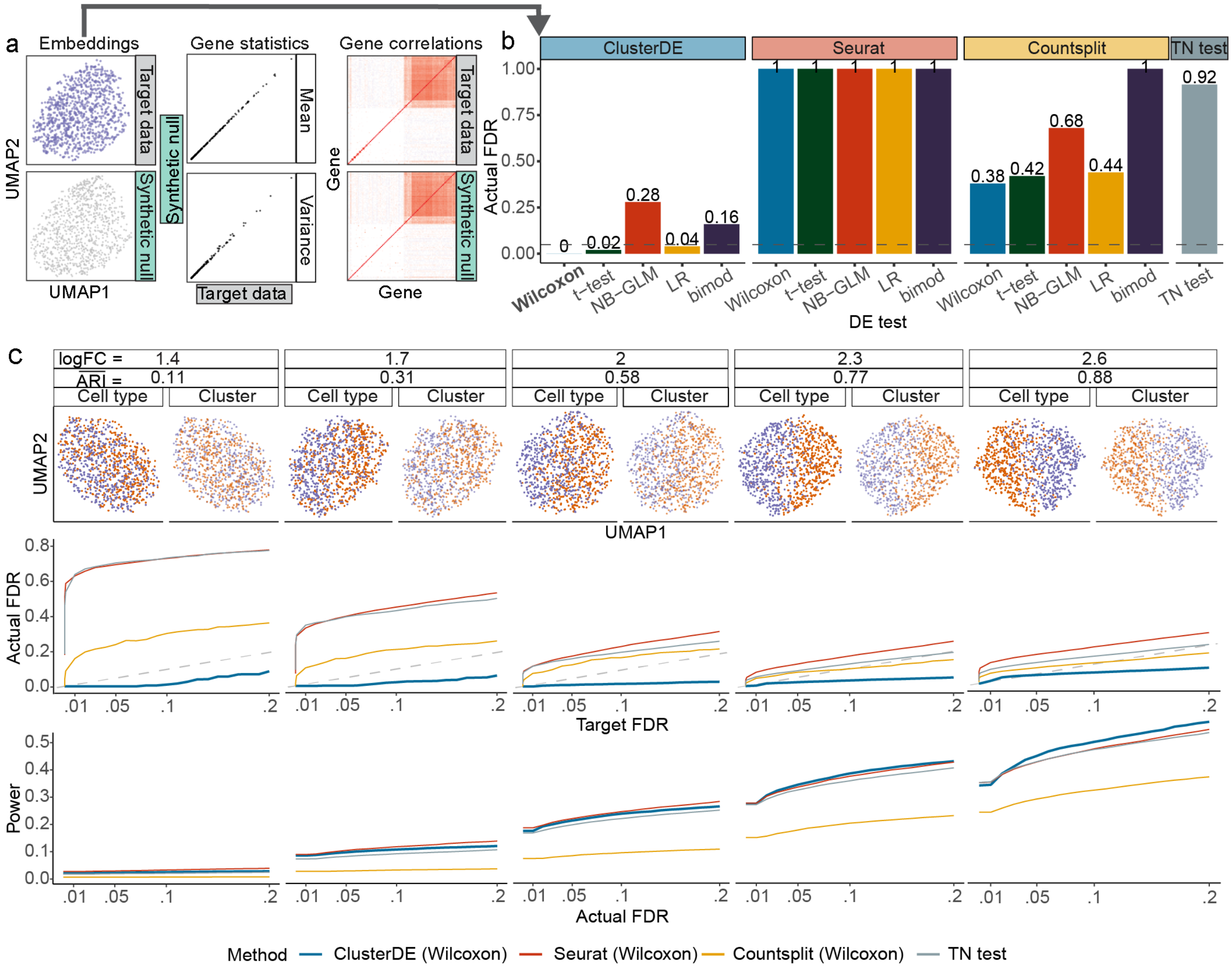


Fig. 2: **ClusterDE achieved reliable FDR control and good power in simulation studies. a**, When the target data contained cells from a single type (simulation; see Supplementary Methods "Simulation setting with one cell type" in the Supplementary Material), the synthetic null data generated by ClusterDE resembled the target data well in terms of UMAP cell embeddings (left), per-gene expression mean and variance statistics (middle), and gene-gene rank correlations (right). **b**, On the target data in **a**, ClusterDE (with five DE tests) outperformed three other methods—including Seurat (which involves double dipping), Countsplit (which aims to address double dipping and works with any DE test), and TN test (which aims to address double dipping and has its own DE test)—in FDR control. The horizontal dashed line indicates the target FDR of 0.05. The five DE tests are the Wilcoxon rank-sum test (Wilcoxon), t-test, negative binomial generalized linear model (NB-GLM), logistic regression model predicting cluster membership with likelihood-ratio test (LR), and likelihood-ratio test for single-cell gene expression (bimod). **c**, The FDRs and power of ClusterDE and the three other methods under various severity levels of double dipping. The log fold change (logFC) summarizes the average gene expression difference between two cell types in simulation (see Supplementary Methods "Simulation setting with two cell types and 200 true DE genes" in the Supplementary Material). Corresponding to a small logFC, a small adjusted Rand index (ARI) represents a bad agreement between cell clusters and cell types, representing a more severe double-dipping issue. Across various severity levels of double dipping, ClusterDE controlled the FDRs under the target FDR thresholds (diagonal dashed line) and achieved comparable or higher power compared to the three other methods at the same actual FDRs.

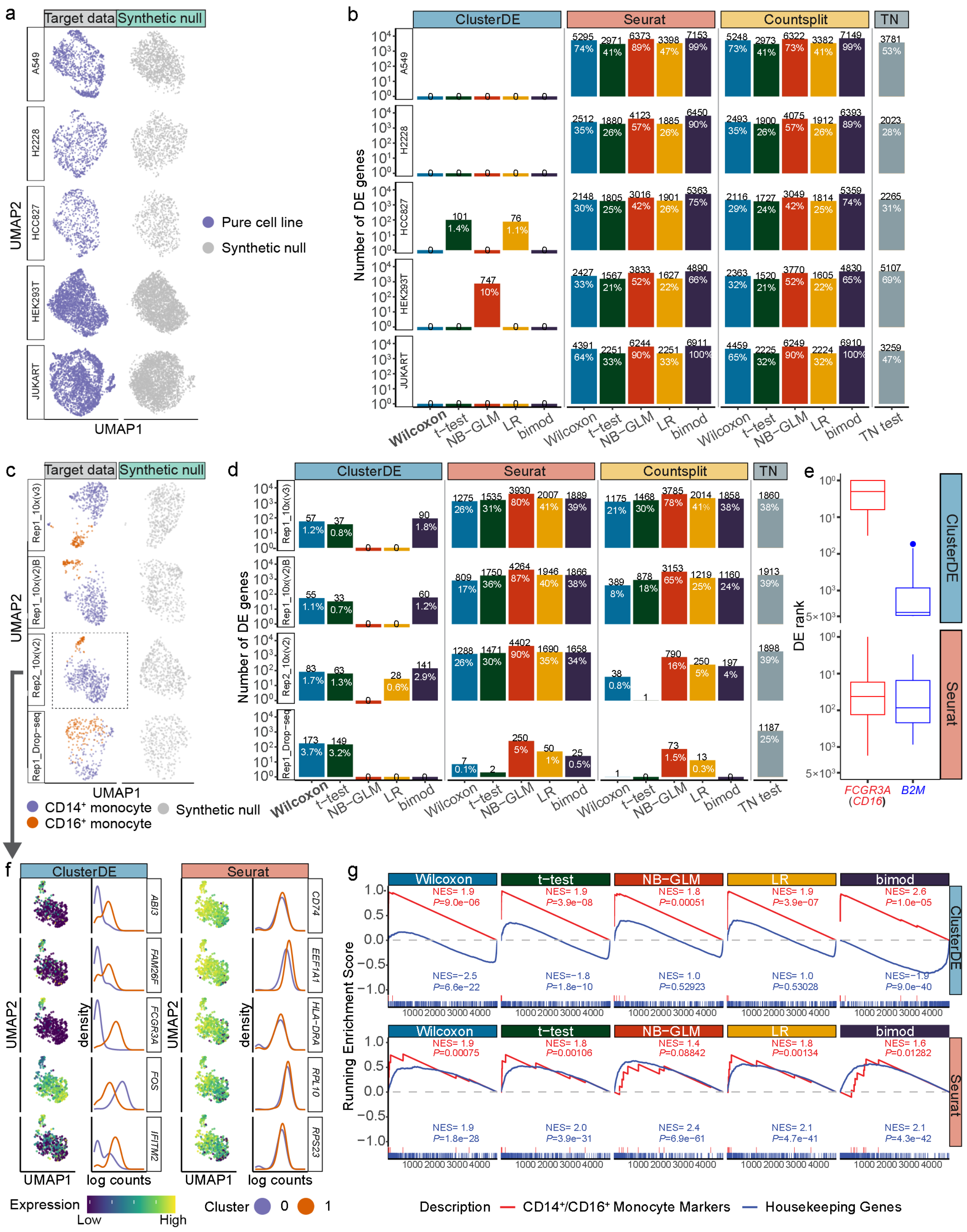
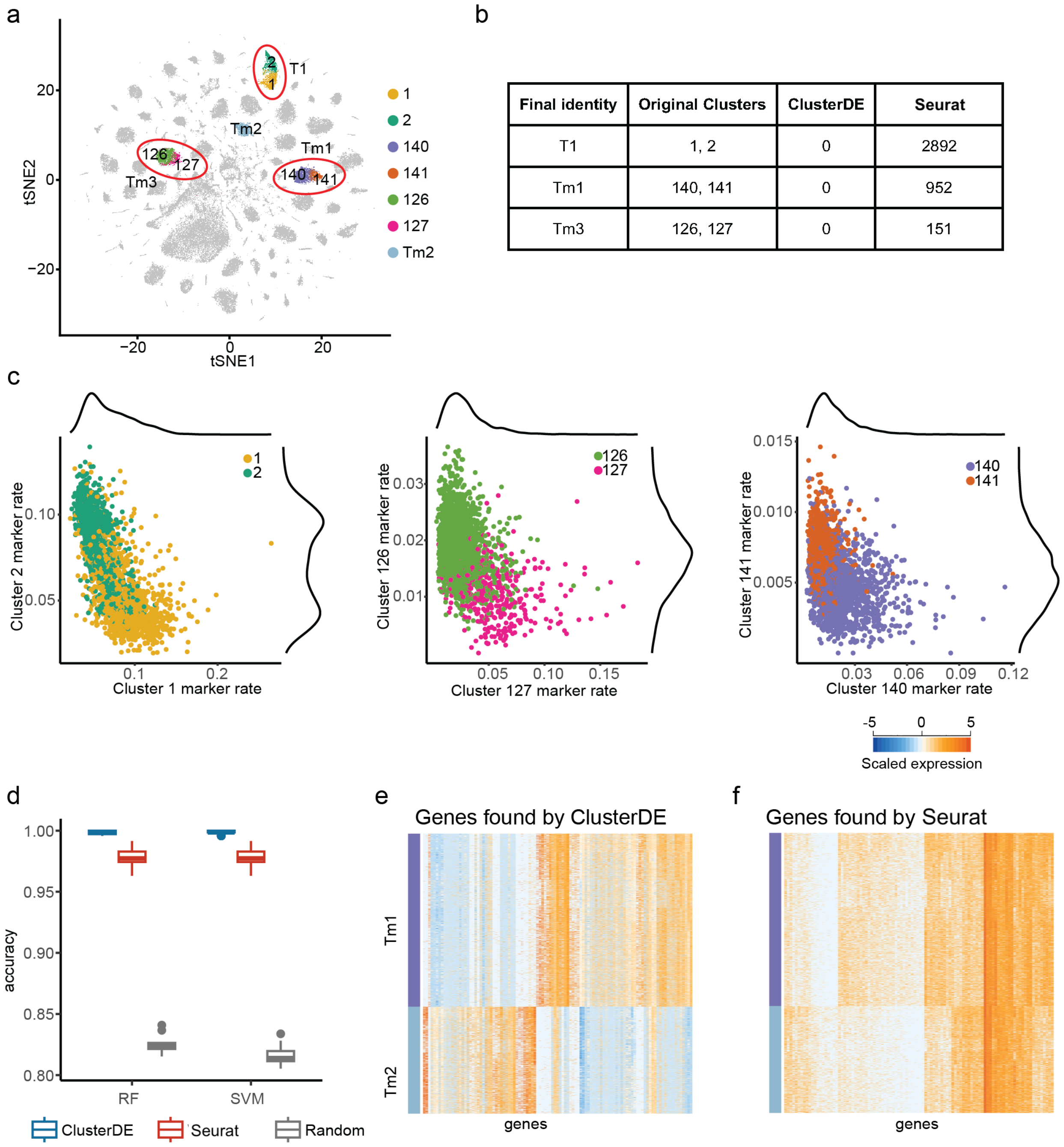

a
Target data
Synthetic null
A549
H2228
HCC827
HEK293T
JUKART
UMAP2
UMAP1
● Pure cell line
● Synthetic null
b
ClusterDE
Seurat
Countsplit
TN
Number of DE genes
A549: ClusterDE 0 0 0 0 0; Seurat 5295 74%, 2971 41%, 6373 89%, 3398 47%, 7153 99%; Countsplit 5248 73%, 2973 41%, 6322 73%, 3382 41%, 7149 99%; TN 3781 53%
H2228: ClusterDE 0 0 0 0 0; Seurat 2512 35%, 1880 26%, 4123 57%, 1885 26%, 6450 90%; Countsplit 2493 35%, 1900 26%, 4075 57%, 1912 26%, 6393 89%; TN 2023 28%
HCC827: ClusterDE 0, 101 1.4%, 0, 76 1.1%, 0; Seurat 2148 30%, 1805 25%, 3016 42%, 1901 26%, 5363 75%; Countsplit 2116 29%, 1727 24%, 3049 42%, 1814 25%, 5359 74%; TN 2265 31%
HEK293T: ClusterDE 0, 0, 747 10%, 0, 0; Seurat 2427 33%, 1567 21%, 3833 52%, 1627 22%, 4890 66%; Countsplit 2363 32%, 1520 21%, 3770 52%, 1605 22%, 4830 65%; TN 5107 69%
JUKART: ClusterDE 0 0 0 0 0; Seurat 4391 64%, 2251 33%, 6244 90%, 2251 33%, 6911 100%; Countsplit 4459 65%, 2225 32%, 6249 90%, 2224 32%, 6910 100%; TN 3259 47%
Wilcoxon t-test NB-GLM LR bimod Wilcoxon t-test NB-GLM LR bimod Wilcoxon t-test NB-GLM LR bimod TN test
c
Target data
Synthetic null
Rep1_10x(v3)
Rep1_10x(v2)B
Rep2_10x(v2)
Rep1_Drop-seq
UMAP2
UMAP1
● CD14⁺ monocyte
● CD16⁺ monocyte
● Synthetic null
d
ClusterDE
Seurat
Countsplit
TN
Number of DE genes
Rep1_10x(v3): ClusterDE 57 1.2%, 37 0.8%, 0, 0, 90 1.8%; Seurat 1275 26%, 1535 31%, 3930 80%, 2007 41%, 1889 39%; Countsplit 1175 21%, 1468 30%, 3785 78%, 2014 41%, 1858 38%; TN 1860 38%
Rep1_10x(v2)B: ClusterDE 55 1.1%, 33 0.7%, 0, 0, 60 1.2%; Seurat 809 17%, 1750 36%, 4264 87%, 1946 40%, 1866 38%; Countsplit 389 8%, 878 18%, 3153 65%, 1219 25%, 1160 24%; TN 1913 39%
Rep2_10x(v2): ClusterDE 83 1.7%, 63 1.3%, 0, 28 0.6%, 141 2.9%; Seurat 1288 26%, 1471 30%, 4402 90%, 1690 35%, 1658 34%; Countsplit 38 0.8%, 1, 790 16%, 250 5%, 197 4%; TN 1898 39%
Rep1_Drop-seq: ClusterDE 173 3.7%, 149 3.2%, 0, 0, 0; Seurat 7 0.1%, 2, 250 5%, 50 1%, 25 0.5%; Countsplit 1, 0, 73 1.5%, 13 0.3%, 0; TN 1187 25%
Wilcoxon t-test NB-GLM LR bimod Wilcoxon t-test NB-GLM LR bimod Wilcoxon t-test NB-GLM LR bimod TN test
e
DE rank
ClusterDE
Seurat
FCGR3A (CD16)
B2M
f
ClusterDE
Seurat
ABI3
FAM26F
FCGR3A
FOS
IFITM2
CD74
EEF1A1
HLA-DRA
RPL10
RPS23
UMAP2
UMAP1
density
log counts
Expression
Low
High
Cluster ● 0 ● 1
g
Wilcoxon
t-test
NB-GLM
LR
bimod
ClusterDE
Seurat
Running Enrichment Score
ClusterDE: Wilcoxon NES= 1.9 P=9.0e-06, NES=-2.5 P=6.6e-22; t-test NES= 1.9 P=3.9e-08, NES=-1.8 P=1.8e-10; NB-GLM NES= 1.8 P=0.00051, NES= 1.0 P=0.52923; LR NES= 1.9 P=3.9e-07, NES= 1.0 P=0.53028; bimod NES= 2.6 P=1.0e-05, NES=-1.9 P=9.0e-40
Seurat: Wilcoxon NES= 1.9 P=0.00075, NES= 1.9 P=1.8e-28; t-test NES= 1.8 P=0.00106, NES= 2.0 P=3.9e-31; NB-GLM NES= 1.4 P=0.08842, NES= 2.4 P=6.9e-61; LR NES= 1.8 P=0.00134, NES= 2.1 P=4.7e-41; bimod NES= 1.6 P=0.01282, NES= 2.1 P=4.3e-42
1000 2000 3000 4000
Description — CD14⁺/CD16⁺ Monocyte Markers — Housekeeping Genes

Fig. 3: **ClusterDE achieved reliable FDR control and good power in identifying cell-type marker genes from real scRNA-seq data. a**, UMAP visualizations of target data (left) and synthetic null data (right) of five cell lines. **b**, Numbers of DE genes (at the target FDR of 0.05) identified by ClusterDE and the three other methods (Seurat, Countsplit, and TN test). While the three other methods found thousands of DE genes—false positives when interpreted as cell-type marker genes—within a single cell line, ClusterDE made no false discoveries when used with most DE tests. The numbers in black and white indicate the number of DE genes and the proportions of DE genes among all genes, respectively. The five DE tests are the Wilcoxon rank-sum test (Wilcoxon), t-test, negative binomial generalized linear model (NB-GLM), logistic regression model predicting cluster membership with likelihood-ratio test (LR), and likelihood-ratio test for single-cell gene expression (bimod). **c**, UMAP visualizations of target data (left) and synthetic null data (right) for four datasets containing two monocyte subtypes: CD14$^+$ monocytes and CD16$^+$ monocytes. The synthetic null data captured the global topology of cells in the target data while filling the gap between the two cell subtypes. The dashed box labels the dataset used in panels **f** and **g**. **d**, ClusterDE identified DE genes between the two cell subtypes. The numbers in black and white indicate the number of DE genes and the proportions of DE genes among all genes, respectively. **e**, The ranks of two exemplary genes (a monocyte subtype marker *FCGR3A* in red and a well-known housekeeping gene *B2M* in blue) in the DE gene lists of ClusterDE and Seurat across the five DE tests and the four datasets in **c**. In each boxplot representing the distribution of 20 ranks, the center horizontal line represents the median, and the box limits represent the upper and lower quartiles. **f**, The top DE genes identified by ClusterDE exhibited distinct expression patterns in the two cell clusters identified by Seurat clustering, a phenomenon not observed for the top DE genes identified by Seurat. For ClusterDE and Seurat, the top DE genes are defined as the common DE genes found by the five DE tests in **d** at the target FDR of 0.05. The UMAP plots show each top DE gene's normalized expression levels in the dataset "Rep2_10x(V2)" (marked by the dashed box in c; see"Dimensionality reduction and visualization" in the Supplementary Material). The density plots depict each top DE gene's normalized expression distributions in the two cell clusters. **g**, Gene set enrichment analysis (GSEA) of the ranked DE gene lists identified by ClusterDE and Seurat with five DE tests from the dataset Rep2_10x(V2). The red lines represent the enrichment of the CD14$^+$/CD16$^+$ monocyte marker gene set, and the blue lines represent the enrichment of the housekeeping gene set. The normalized enrichment score (NES) reflects the direction and magnitude of enrichment, and the P value indicates the significance of enrichment.



Final identity	Original Clusters	ClusterDE	Seurat
T1	1, 2	0	2892
Tm1	140, 141	0	952
Tm3	126, 127	0	151

Fig. 4: **Application of ClusterDE to the adult *Drosophila* neuronal atlas. a**, t-SNE visualization of the data with highlighted cell types: T1 (Clusters 1 and 2), Tm3 (Clusters 126 and 127), Tm1 (Clusters 140 and 141), and Tm2. **b**, Numbers of DE genes (at the target FDR of 0.05) identified by ClusterDE and Seurat (with Wilcoxon as the DE test). **c**, Gating plots comparing Clusters 1 vs. 2 (left), Clusters 126 vs. 127 (middle), and Clusters 140 vs. 141 (right). **d**, Boxplot showing accuracy achieved by RF and SVM models using the 105 DE genes found by ClusterDE, the top 105 genes (ordered by increasing P value) found by Seurat only, and 105 random genes. A two-sided t-test comparing the accuracy obtained using the DE genes from ClusterDE vs the DE genes from Seurat only shows a significant difference (P value $< 2.2 \times 10^{-16}, n = 100$ samples each) for both RF and SVM models. **e**, Heatmap of the 105 DE genes found by ClusterDE. **f**, Heatmap of the top 105 DE genes found by Seurat only. In both **e** and **f**, genes are ordered by hierarchical clustering as implemented by R function `heatmap.2`.

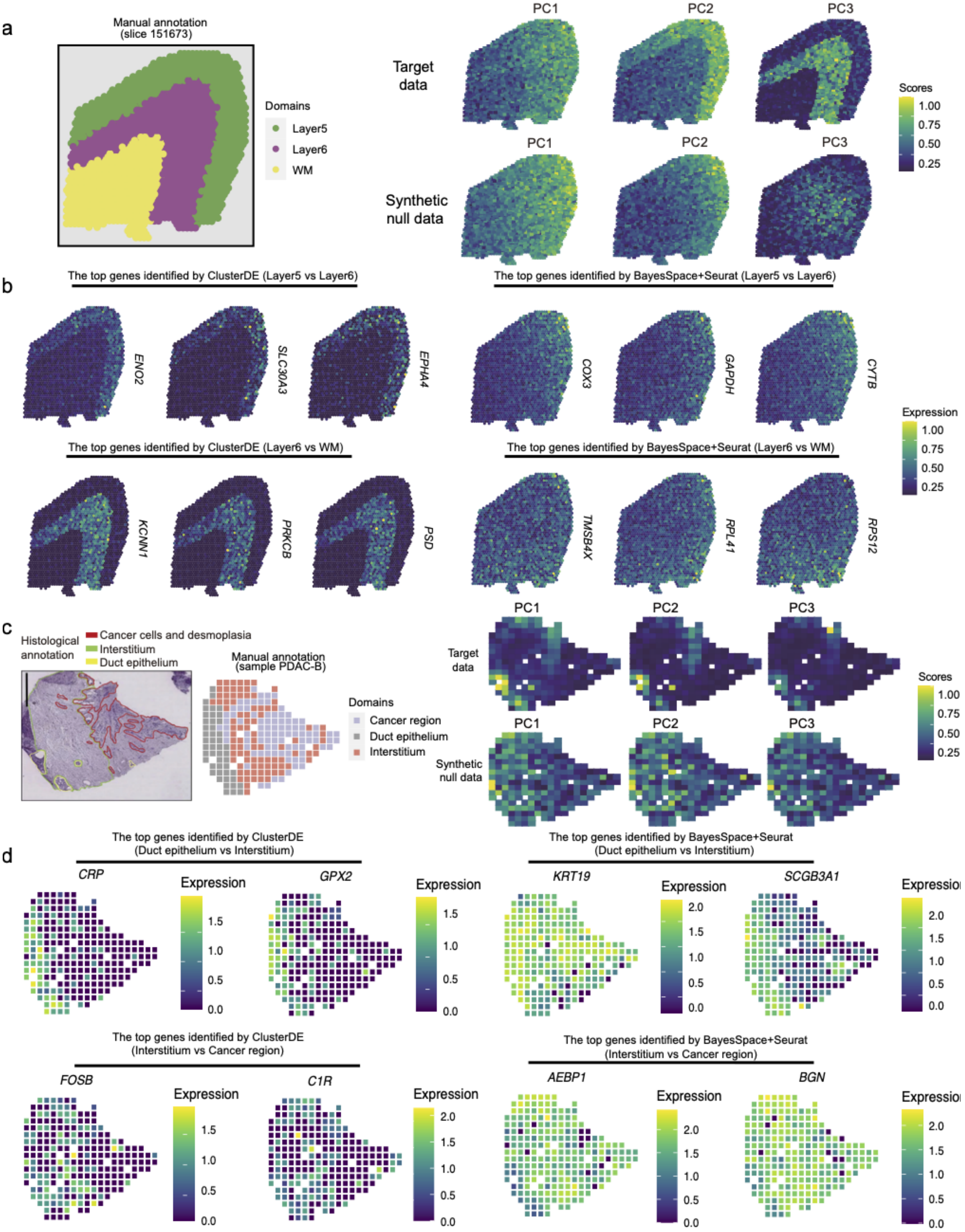

a
Manual annotation
(slice 151673)
Domains
Layer5
Layer6
WM
PC1
PC2
PC3
Target data
Synthetic null data
Scores
1.00
0.75
0.50
0.25
b
The top genes identified by ClusterDE (Layer5 vs Layer6)
ENO2
SLC30A3
EPHA4
The top genes identified by BayesSpace+Seurat (Layer5 vs Layer6)
COX3
GAPDH
CYTB
The top genes identified by ClusterDE (Layer6 vs WM)
KCNN1
PRKCB
PSD
The top genes identified by BayesSpace+Seurat (Layer6 vs WM)
TMSB4X
RPL41
RPS12
Expression
1.00
0.75
0.50
0.25
c
Histological annotation
Cancer cells and desmoplasia
Interstitium
Duct epithelium
Manual annotation
(sample PDAC-B)
Domains
Cancer region
Duct epithelium
Interstitium
PC1
PC2
PC3
Target data
Synthetic null data
Scores
1.00
0.75
0.50
0.25
d
The top genes identified by ClusterDE
(Duct epithelium vs Interstitium)
CRP
GPX2
The top genes identified by BayesSpace+Seurat
(Duct epithelium vs Interstitium)
KRT19
SCGB3A1
The top genes identified by ClusterDE
(Interstitium vs Cancer region)
FOSB
C1R
The top genes identified by BayesSpace+Seurat
(Interstitium vs Cancer region)
AEBP1
BGN
Expression
2.0
1.5
1.0
0.5
0.0

Fig. 5: **ClusterDE mitigates double dipping and identifies reliable spatial-domain marker genes in SRT post-clustering DE analysis.** ClusterDE demonstrates robust statistical power in identifying spatial-domain marker genes from both a simulated SRT dataset containing three domains and the PDAC-B cancer slice from the human primary pancreatic cancer (PDAC) SRT dataset. **a**, Visualization of annotated spatial domains alongside the spatial expression patterns of three representative principal components (PCs) among the top five PCs in the target data (simulated) and synthetic null data. **b**, Spatial expression patterns of top domain marker genes. The first and second rows display the expression of top domain marker genes of Layer 5 and Layer 6, respectively, detected by ClusterDE and BayesSpace+Seurat (BayesSpace for clustering and Seurat for DE). **c**, Visualization of annotated domains alongside the spatial expression patterns of three representative PCs among the top five PCs in the target data (PDAC-B) and synthetic null data. **d**, Spatial expression patterns of top domain marker genes. The first and second rows display the expression of top domain marker genes of the Duct Epithelium and Interstitium, respectively, detected by ClusterDE and BayesSpace+Seurat.

References

[1] Ashraful Haque et al. "A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications". In: *Genome medicine* 9.1 (2017), pp. 1–12.

[2] Vladimir Yu Kiselev et al. "SC3: consensus clustering of single-cell RNA-seq data". In: *Nature methods* 14.5 (2017), pp. 483–486.

[3] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. "Challenges in unsupervised clustering of single-cell RNA-seq data". In: *Nature Reviews Genetics* 20.5 (2019), pp. 273–282.

[4] Tallulah S Andrews et al. "Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data". In: *Nature protocols* 16.1 (2021), pp. 1–9.

[5] Yuhan Hao et al. "Integrated analysis of multimodal single-cell data". In: *Cell* 184.13 (2021), pp. 3573–3587.

[6] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. "SCANPY: large-scale single-cell gene expression data analysis". In: *Genome biology* 19 (2018), pp. 1–5.

[7] Jonas Simon Fleck, J. Gray Camp, and Barbara Treutlein. "What is a cell type?" In: *Science* 381.6659 (2023), pp. 733–734. DOI: 10.1126/science.adf6162. eprint: https://www.science.org/doi/pdf/10.1126/science.adf6162. URL: https://www.science.org/doi/abs/10.1126/science.adf6162.

[8] Amber Dance. "What is a cell type, really? The quest to categorize life's myriad forms". In: *Nature* 633.8031 (2024), pp. 754–756. DOI: 10.1038/d41586-024-03073-2. URL: https://www.nature.com/articles/d41586-024-03073-2.

[9] Jesse M Zhang, Govinda M Kamath, and N Tse David. "Valid post-clustering differential analysis for single-cell RNA-Seq". In: *Cell systems* 9.4 (2019), pp. 383–392.

[10] Anna Neufeld et al. "Inference after latent variable estimation for single-cell RNA sequencing data". In: *Biostatistics* 25.1 (2024), pp. 270–287.

[11] Anna Neufeld et al. "Data thinning for convolution-closed distributions". In: *Journal of Machine Learning Research* 25.57 (2024), pp. 1–35.

[12] Alexis Vandenbon and Diego Diez. "A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data". In: *Nature communications* 11.1 (2020), p. 4318.

[13] Anna Hendrika Cornelia Vlot, Setareh Maghsudi, and Uwe Ohler. "Cluster-independent marker feature identification from single-cell omics data using SEMITONES". In: *Nucleic Acids Research* 50.18 (2022), e107–e107.

[14] Chanwoo Kim et al. "MarcoPolo: a method to discover differentially expressed genes in single-cell RNA-seq data without depending on prior clustering". In: *Nucleic acids research* 50.12 (2022), e71–e71.

[15] Jiadi Zhu and Youlong Yang. "scMEB: a fast and clustering-independent method for detecting differentially expressed genes in single-cell RNA-seq data". In: *BMC genomics* 24.1 (2023), pp. 1–15.

[16] Alsu Missarova et al. "Sensitive cluster-free differential expression testing." In: *bioRxiv* (2023), pp. 2023–03.

[17] Huidong Chen et al. "SIMBA: SIngle-cell eMBedding Along with features". In: *Nature Methods* (2023), pp. 1–11.

[18] Baolin Liu et al. "An entropy-based metric for assessing the purity of single cell populations". In: *Nature communications* 11.1 (2020), p. 3155.

[19] Jiyuan Fang et al. "Clustering Deviation Index (CDI): a robust and accurate internal measure for evaluating scRNA-seq data clustering". In: *Genome Biology* 23.1 (2022), p. 269.

[20] Wei Vivian Li. "Phitest for analyzing the homogeneity of single-cell populations". In: *Bioinformatics* 38.9 (2022), pp. 2639–2641.

[21] Maria Mircea et al. "Phiclust: a clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations". In: *Genome Biology* 23.1 (2022), pp. 1–24.

[22] I.N. Grabski, K. Street, and R.A. Irizarry. "Significance analysis for clustering with single-cell RNA-sequencing data". In: *Nat Methods* 1.1 (2023), p. 1.

[23] Xinwei He et al. "scAce: an adaptive embedding and clustering method for single-cell gene expression data". In: *Bioinformatics* 39.9 (2023), btad546.

[24] Cathrine Petersen, Lennart Mucke, and M Ryan Corces. "CHOIR improves significance-based detection of cell types and states from single-cell data". In: *Biorxiv* (2024).

[25] Alan DenAdel et al. “A knockoff calibration method to avoid over-clustering in single-cell RNA-sequencing”. In: *bioRxiv* (2024), pp. 2024–03.

[26] Vivien Marx. “Method of the Year: spatially resolved transcriptomics”. In: *Nature methods* 18.1 (2021), pp. 9–14.

[27] Lulu Shang and Xiang Zhou. “Spatially aware dimension reduction for spatial transcriptomics”. In: *Nature communications* 13.1 (2022), p. 7203.

[28] Peiying Cai, Mark D Robinson, and Simone Tiberi. “DESpace: spatially variable gene detection via differential expression testing of spatial clusters”. In: *Bioinformatics* 40.2 (2024), btae027.

[29] Dongyuan Song et al. “scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics”. In: *Nature Biotechnology* (2023), pp. 1–6.

[30] Xinzhou Ge et al. “Clipper: p-value-free FDR control on high-throughput data from two conditions”. In: *Genome biology* 22.1 (2021), pp. 1–29.

[31] Angelo Duò, Mark D Robinson, and Charlotte Soneson. “A systematic performance evaluation of clustering methods for single-cell RNA-seq data”. In: *F1000Research* 7 (2018).

[32] Rina Foygel Barber and Emmanuel J Candès. “Controlling the false discovery rate via knockoffs”. In: *Annals of Statistics* (2015).

[33] Emmanuel Candes et al. “Panning for gold:‘model-X’knockoffs for high dimensional controlled variable selection”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.3 (2018), pp. 551–577.

[34] Grace XY Zheng et al. “Massively parallel digital transcriptional profiling of single cells”. In: *Nature communications* 8.1 (2017), p. 14049.

[35] Luyi Tian et al. “Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments”. In: *Nature methods* 16.6 (2019), pp. 479–487.

[36] R Ian Freshney. *Culture of animal cells: a manual of basic technique and specialized applications*. John Wiley & Sons, 2015.

[37] Andrew Butler et al. “Integrating single-cell transcriptomic data across different conditions, technologies, and species”. In: *Nature biotechnology* 36.5 (2018), pp. 411–420.

[38] Jiarui Ding et al. “Systematic comparison of single-cell and single-nucleus RNA-sequencing methods”. In: *Nature biotechnology* 38.6 (2020), pp. 737–746.

[39] Yu Matsuzaki et al. “β2-Microglobulin is an appropriate reference gene for RT-PCR-based gene expression analysis of hematopoietic stem cells”. In: *Regenerative Therapy* 1 (2015), pp. 91–97.

[40] Gordon K Smyth and Terry Speed. “Normalization of cDNA microarray data”. In: *Methods* 31.4 (2003), pp. 265–273.

[41] Mehmet Neset Özel et al. “Neuronal diversity and convergence in a visual system developmental atlas”. In: *Nature* (2021), pp. 88–95.

[42] Ines Lago-Baldaia et al. “A *Drosophila* glial cell atlas reveals a mismatch between transcriptional and morphological diversity”. In: *PLoS Biology* 21 (2023).

[43] Rita Kottmeier et al. “Wrapping glia regulates neuronal signaling speed and precision in the peripheral nervous system of *Drosophila*”. In: *Nature communications* (2020).

[44] Rudy Behnia et al. “Processing properties of ON and OFF pathways for Drosophila motion detection”. In: *Nature* (2014), pp. 427–30.

[45] Edward Zhao et al. “Spatial transcriptomics at subspot resolution with BayesSpace”. In: *Nature biotechnology* 39.11 (2021), pp. 1375–1384.

[46] Kristen R Maynard et al. “Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex”. In: *Nature neuroscience* 24.3 (2021), pp. 425–436.

[47] Reuben Moncada et al. “Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas”. In: *Nature biotechnology* 38.3 (2020), pp. 333–342.

[48] Dongyuan Song and Jingyi Jessica Li. “PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data”. In: *Genome biology* 22.1 (2021), p. 124.

[49] Dongyuan Song, Kexin Li, and Jingyi Jessica Li. *ClusterDE: a post-clustering differential expression (DE) method robust to false-positive inflation caused by double dipping*. Version 0.9.0. July 2023. DOI: 10.5281/zenodo.8161964. URL: https://doi.org/10.5281/zenodo.8161964.

Supplementary Material for "Synthetic control removes spurious discoveries from double dipping in single-cell and spatial transcriptomics data analyses"

Dongyuan Song, Siqi Chen, Christy Lee, Kexin Li, Xinzhou Ge, Jingyi Jessica Li

S1 Supplementary Methods

S1.1 Practical guidelines for using ClusterDE

ClusterDE is designed to identify potential cell-type or spatial-domain marker genes through pairwise comparisons of ambiguous cell or spatial clusters obtained from cell or spatial clustering. Generally, the goal of cell clustering based on gene expression is to infer distinct cell types, where marker genes exhibit bimodal distributions between two cell types. Similarly, the goal of spatial clustering for spatially resolved transcriptomics (SRT) data is to infer distinct spatial domains, where marker genes show discontinuous expression changes at domain boundaries.

To unify and simplify terminology, we do not strictly differentiate between "cells" and "spots" (for spot-resolution SRT) or between "cell clusters" (potential cell types) and "spatial clusters" (potential spatial domains) in the following text. In practice, we recommend using ClusterDE by following these steps.

1. Given a set of clusters, identify two clusters that may represent potentially distinct cell types/subtypes or spatial domains. For users employing Seurat for scRNA-seq data analysis, the `BuildClusterTree` function can be used to construct a hierarchical tree of the cell clusters, allowing examination of two leaf clusters with ambiguous distinctions. For users performing spatial clustering, select two clusters that are spatially adjacent.
2. Subset the data to include only the cells or spots within the two selected clusters.
3. Use the subsetted data as the "target data" input for ClusterDE.
4. Review the DE genes identified by ClusterDE and assess whether the two clusters represent biologically meaningful and distinct cell types/subtypes or spatial domains.

Notably, ClusterDE does not automatically determine whether two clusters should be merged, unlike methods designed to directly assess clustering quality for scRNA-seq data [1, 2, 3, 4, 5]. Instead, ClusterDE emphasizes identifying reliable post-clustering DE genes that may serve as potential cell-type or spatial-domain markers. This approach allows researchers to gain biological insights by analyzing the specific genes that distinguish potentially ambiguous clusters, aiding in the determination of whether these clusters represent distinct cell types or spatial domains. Thus, unlike clustering quality assessment methods, ClusterDE enables researchers to explore the functional and molecular characteristics of clusters.

In step 3 of the above procedure, users can either provide pre-existing cell cluster labels for the target data (the default option in ClusterDE) or allow ClusterDE to re-cluster the target data. If the default option is selected, ClusterDE clusters only the synthetic null data, and the target DE scores are computed based on the input cell clusters. Alternatively, if re-clustering is chosen, ClusterDE performs clustering on both the target data and the synthetic null data in parallel. However, the downside of this alternative approach is that the resulting target cell clusters may differ from the original input clusters, making it unsuitable for determining whether the original clusters in the target data represent distinct cell types or spatial domains.

S1.2 ClusterDE method details

Notations for the double-dipping problem in post-clustering DE analysis. The target data is represented as $\mathbf{Y} = [Y_{ij}] \in \mathbb{N}_{\geq 0}^{n \times m}$, where n observations are rows, m genes are columns, and Y_{ij} denotes the UMI count of gene $j = 1, \ldots, m$ in observation $i = 1, \ldots, n$. In scRNA-seq data, each observation corresponds to a cell, while in SRT data, observations can vary in spatial resolution—for instance, each spot may contain a single cell or multiple cells. For simplicity, we may refer to spots as cells. Each observation i is represented as an m-dimensional gene expression vector, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im})^\mathsf{T}$.

In our formulation of the post-clustering DE problem for scRNA-seq data, the n cells belong to two latent cell types, which may or may not be identical, and are partitioned into two clusters by a clustering algorithm. We represent cell i's latent cell type using $Z_i \in \{0, 1\}$. We define the "ideal DE test" as the test that decides whether a gene has equal mean expression in two cell types. For gene j, we assume that its expression levels in cells of type 0, $\{(Y_{ij}|Z_i = 0)\}_{i=1}^n$, share a common mean, denoted as $\mu_{0j} = \mathbb{E}[Y_{ij}|Z_i = 0]$, and its expression levels in cells of type 1, $\{(Y_{ij}|Z_i = 1)\}_{i=1}^n$, share a common mean, denoted as $\mu_{1j} = \mathbb{E}[Y_{ij}|Z_i = 1]$. The ideal DE test is then formulated with the following null hypothesis H_{0j} and alternative hypothesis H_{1j}:

$$H_{0j} : \mu_{0j} = \mu_{1j} \quad \text{vs.} \quad H_{1j} : \mu_{0j} \neq \mu_{1j} \,.$$

Hence, gene j is a true DE gene if and only if H_{0j} does not hold. If all n cells belong to a single cell type (i.e., the two latent cell types are identical), then all m null hypotheses, $H_{01}, \ldots, H_{0m}$, hold simultaneously.

Note that the null hypothesis depends on the DE test used. In most cases, the DE test evaluates the equality of means. However, the default DE test in Seurat, the Wilcoxon rank-sum test, assesses whether a gene's expression distributions are identical between the two latent cell types. Nonetheless, the above 'equal mean' formulation can be easily generalized.

Since Z_i's are unobserved, standard scRNA-seq data analysis partitions cells into two clusters using a clustering algorithm g (e.g., the Louvain algorithm in Seurat) applied to $\mathbf{Y}$. We denote cell i's cluster membership as $\widehat{Z}_i = g_{\mathbf{Y}}(\mathbf{Y}_i) \in \{0, 1\}$, where $g_{\mathbf{Y}} : \{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\} \to \{0, 1\}$ is the clustering function constructed from the clustering algorithm g and the data $\mathbf{Y}$, mapping a cell's gene expression vector to a cluster membership.

For spot i in SRT data, in addition to the gene expression vector $\mathbf{Y}_i$, we also observe the two-dimensional spatial location $X_i = (x_{i1}, x_{i2})^\mathsf{T}$. Typically, spatial clustering algorithms incorporate both gene expression data $\mathbf{Y}$ and spatial locations $X = [X_1, \cdots, X_n]^\mathsf{T}$ when clustering spots. Consequently, the cluster membership for spot i is given by $\widehat{Z}_i = g_{\mathbf{Y},X}(\mathbf{Y}_i, x_{i1}, x_{i2}) \in \{0, 1\}$, where $g_{\mathbf{Y},X}$ represents the clustering function utilizing both data types.

After clustering, standard scRNA-seq and SRT analyses conduct a DE test for each gene based on $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ and $\widehat{Z}_1, \ldots, \widehat{Z}_n$. In this process, the data $\mathbf{Y}$ is used twice—once in clustering and again in DE analysis—a phenomenon referred to as the "double-dipping (DD) issue." The standard post-clustering DE analysis in Seurat has this DD issue and employs a statistical test that differs from the ideal DE test. Specifically, for gene j, we define $\mu_{0j}^{\mathrm{DD}} = \mathbb{E}[Y_{ij}|\widehat{Z}_i = 0]$ and $\mu_{1j}^{\mathrm{DD}} = \mathbb{E}[Y_{ij}|\widehat{Z}_i = 1]$. The post-clustering DE analysis in Seurat corresponds to the following null hypothesis H_{0j}^{DD} and alternative hypothesis H_{1j}^{DD}:

$$H_{0j}^{\mathrm{DD}} : \mu_{0j}^{\mathrm{DD}} = \mu_{1j}^{\mathrm{DD}} \quad \text{vs.} \quad H_{1j}^{\mathrm{DD}} : \mu_{0j}^{\mathrm{DD}} \neq \mu_{1j}^{\mathrm{DD}} \,.$$

Hence, gene j would be falsely identified as a cell-type marker gene if H_{0j}^{DD} is rejected but H_{0j} holds, leading to an inflated FDR in identifying cell-type marker genes. Fig. S26 provides a toy example illustrating this issue; the same problem also arises in post-clustering DE analysis for SRT data.

ClusterDE Step 1: synthetic null data generation. Previous studies on scRNA-seq data have shown that, within a single cell type, each gene's UMI counts can be well-modeled by a negative binomial (NB) distribution [6, 7, 8]. Furthermore, the joint UMI counts for all genes can be effectively approximated by a multivariate negative binomial (MVNB) distribution defined using the Gaussian copula [9]. Based on these findings, ClusterDE employs an MVNB distribution specified by the Gaussian copula as the scRNA-seq null model, representing a single "hypothetical" cell type. In step 1 of ClusterDE, the null model is fitted to the target data $\mathbf{Y}$, and synthetic null data are subsequently sampled from the fitted null model. The assumption behind this null model is that a single cell type is a homogeneous population where the marginal count distribution for each gene follows an NB distribution, and the gene-gene correlation structure is specified by the Gaussian copula. Moreover, if users have prior knowledge of a more appropriate marginal distribution, ClusterDE can generate synthetic null data from other models, including multivariate Gaussian, multivariate Poisson, multivariate zero-inflated Poisson, and multivariate zero-inflated NB distributions.

For SRT data, we define a null model representing a single "hypothetical" spatial domain. In this model, each gene exhibits smooth spatially variable expression across the spots, without abrupt changes that would suggest the presence of more than one spatial domain. The gene-gene correlation structure is still modeled using the Gaussian copula.

The concept of fitting a null model to the target data, regardless of whether the target data aligns with the null model, underpins the widely used likelihood-ratio test in statistics [10]. In this test, the maximum likelihood under the null hypothesis is calculated from the target data and compared with the maximum likelihood under a more flexible alternative hypothesis, also calculated from the target data. The null hypothesis is rejected only if the maximum likelihood under the null is significantly smaller than that under the alternative. ClusterDE extends this principle by sampling synthetic null data from the null model fitted to the target data using maximum likelihood estimation. This approach allows any post-clustering DE pipeline, regardless of complexity, to be applied simultaneously to both the synthetic null data and the target data. A contrastive strategy is then used to identify trustworthy DE genes—those with DE scores significantly higher in the target data than in the synthetic null data.

1. **Specifying a null model**

   Under the null model for scRNA-seq data, we assume that Y_{ij}, gene j's UMI count in cell i, independently (across cells, not genes) follows the $\mathrm{NB}(\mu_j, \sigma_j)$ distribution with the probability mass function:

$$\mathbb{P}(Y_{ij} = y; \mu_j, \sigma_j) = \frac{\Gamma\left(y + \frac{1}{\sigma_j}\right)}{\Gamma\left(\frac{1}{\sigma_j}\right)\Gamma(y+1)} \left(\frac{1}{1+\sigma_j\mu_j}\right)^{\frac{1}{\sigma_j}} \left(\frac{\sigma_j\mu_j}{1+\sigma_j\mu_j}\right)^{y};$$

$$y \in \{0, 1, 2, \cdots\},$$

   where μ_j and σ_j are the mean and dispersion parameters of the NB distribution for gene j. That is,

$$Y_{1j}, \ \cdots, \ Y_{nj} \overset{\text{i.i.d.}}{\sim} \mathrm{NB}(\mu_j, \sigma_j),$$

   with "i.i.d." short for "independent and identically distributed," meaning that the n cells' counts for gene j represent a random sample from $\mathrm{NB}(\mu_j, \sigma_j)$.

   For SRT data, we have additional information: the two-dimensional spatial location of each cell/spot, $X_i = (x_{i1}, x_{i2})^{\mathsf{T}}$. In the null model for SRT data, each gene's mean expression is modeled as a smooth surface over spatial locations. Specifically, we assume that gene j's mean expression at spot i is given by $\log(\mu_{ij}) = \alpha_j + f_j^{GP}(x_{i1}, x_{i2}, K)$, where α_j denotes gene j's specific intercept, and the $f_j^{GP}(\cdot)$ represents a penalized Gaussian process regressor with K bases specific to gene j. The parameter K defines the upper bound of the "wiggliness" of the smooth surface. In our default ClusterDE setting, we use $K = 4$, the minimum possible value for K, ensuring the smoothest spatial pattern that a two-dimensional Gaussian process regressor can represent. This choice reflects the smoothest spatial variation that aligns with the null hypothesis, avoiding overfitting to gene expression measurement noise. That is,

$$Y_{ij} \overset{\text{ind.}}{\sim} \mathrm{NB}(\mu_{ij}, \sigma_j),$$

   where $\mu_{ij} = \alpha_j + f_j^{GP}(x_{i1}, x_{i2}, 4)$. In other words, we assume that under the null model, each gene within a single domain exhibits a smooth spatial expression surface.

   Let F_j denote the cumulative distribution function (CDF) of $\mathrm{NB}(\mu_j, \sigma_j)$ for scRNA-seq data or $\mathrm{NB}(\mu_{ij}, \sigma_j)$ for SRT data, where $j = 1, \ldots, m$ and $i = 1, \ldots, n$, the MVNB distribution specified by the Gaussian copula is

$$\begin{matrix} \left(\Phi^{-1}(F_1(Y_{11})), \cdots, \Phi^{-1}(F_m(Y_{1m}))\right)^{\mathsf{T}}, \\ \vdots \\ \left(\Phi^{-1}(F_1(Y_{n1})), \cdots, \Phi^{-1}(F_m(Y_{nm}))\right)^{\mathsf{T}} \end{matrix} \overset{\text{i.i.d.}}{\sim} N_m(\mathbf{0}, \mathbf{R}),$$

where Φ is the CDF of the standard Gaussian distribution $N(0,1)$, and $N_m(\mathbf{0}, \mathbf{R})$ is an m-dimensional Gaussian distribution with an m-dimensional $\mathbf{0}$ mean vector and an m-by-m correlation matrix $\mathbf{R}$ (which is also the covariance matrix because all m Gaussian variables have unit variances). This null model assumes that, after each gene is transformed to a standard Gaussian random variable, the n cells represent a random sample from an m-dimensional Gaussian distribution with zero means, unit variances, and a correlation matrix $\mathbf{R}$. Technically, to address the issue that the Gaussian copula is unidentifiable for discrete F_j's, each F_j is converted into a continuous CDF using a random interpolation procedure called the "distribution transform" (detailed below) [11].

In summary, the scRNA-seq null model parameters include $\{\mu_j, \sigma_j\}_{j=1}^m$ and $\mathbf{R}$, while the SRT null model is specified by $\{\alpha_j, f_j^{GP}, \sigma_j\}_{j=1}^m$ and $\mathbf{R}$.

2. **Fitting the null model to target data (parameter estimation)**

First, F_j is fitted for each gene $j = 1, \ldots, m$. For the scRNA-seq null model, the parameters $\{\mu_j, \sigma_j\}_{j=1}^m$ are estimated using maximum likelihood estimation for the m NB distributions. Specifically, $Y_{1j}, \ldots, Y_{nj}$ are used to estimate μ_j and σ_j as $\widehat{\mu}_j$ and $\widehat{\sigma}_j$, respectively, for each $j = 1, \ldots, m$. For the SRT null model, $Y_{1j}, \ldots, Y_{nj}$ and spatial locations $X_1, \ldots, X_n$ are used to estimate μ_{ij} as $\widehat{\mu}_{ij}$ and σ_j as $\widehat{\sigma}_j$ through a penalized Gaussian process. The resulting fitted CDFs are denoted as $\widehat{F}_1, \ldots, \widehat{F}_m$.

Second, to estimate the correlation matrix $\mathbf{R}$, each Y_{ij} is transformed into $U_{ij} = V_{ij} \cdot \widehat{F}_j(Y_{ij}) + (1 - V_{ij}) \cdot \widehat{F}_j(Y_{ij} + 1)$, where $V_{ij} \overset{\text{i.i.d.}}{\sim} \text{Uniform}[0,1]$, so that $U_{ij} \sim \text{Uniform}[0,1]$. This procedure is referred to as the "distribution transform" to convert the discrete random variable Y_{ij} into U_{ij}, a continuous Uniform$[0,1]$ random variable [11]. Then, $\mathbf{R}$ is estimated as the sample correlation matrix of the transformed data

$$\left(\Phi^{-1}(U_{11}), \cdots, \Phi^{-1}(U_{1m})\right)^\mathsf{T}, \\ \cdots, \\ \left(\Phi^{-1}(U_{n1}), \cdots, \Phi^{-1}(U_{nm})\right)^\mathsf{T}$$

and denoted as $\widehat{\mathbf{R}}$.

In summary, the fitted null model is specified by $\widehat{F}_1, \ldots, \widehat{F}_m$ and $\widehat{\mathbf{R}}$.

3. **Sampling from the fitted null model (synthetic null data generation)**

First, n Gaussian vectors of m dimensions are independently sampled from $N_m(\mathbf{0}, \widehat{\mathbf{R}})$ as

$$(\widetilde{Z}_{11}, \cdots, \widetilde{Z}_{1m})^\mathsf{T}, \ \cdots, \ (\widetilde{Z}_{n1}, \cdots, \widetilde{Z}_{nm})^\mathsf{T}.$$

Second, The n Gaussian vectors are converted to NB count vectors as

$$\widetilde{\mathbf{Y}}_1 := \left(\widehat{F}_1^{-1}(\Phi(\widetilde{Z}_{11})), \cdots, \widehat{F}_m^{-1}(\Phi(\widetilde{Z}_{1m}))\right)^\mathsf{T}, \\ \cdots, \\ \widetilde{\mathbf{Y}}_n := \left(\widehat{F}_1^{-1}(\Phi(\widetilde{Z}_{n1})), \cdots, \widehat{F}_m^{-1}(\Phi(\widetilde{Z}_{nm}))\right)^\mathsf{T},$$

which represent the n synthetic null cells, each of which contains m genes' synthetic null counts sampled from the null model.

In summary, the target data is an n-by-m count matrix $\mathbf{Y}$, with the n real cells $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ as the rows. Similarly, the synthetic null data is an n-by-m count matrix $\widetilde{\mathbf{Y}}$, with the n synthetic null cells $\widetilde{\mathbf{Y}}_1, \ldots, \widetilde{\mathbf{Y}}_n$ as the rows. Note that there is no one-to-one correspondence between the real cells and the synthetic null cells, as the synthetic null cells are independently sampled from the null model.

ClusterDE Step 2: cell clustering. While ClusterDE supports any clustering algorithm, we followed common practice by using the R package Seurat (version 4.2.0) for scRNA-seq data and the R package BayesSpace (version 1.14.0) for SRT data. Specifically, we applied the same clustering algorithm to both the target data and the synthetic null data in parallel, resulting in two clusters for each dataset.

For scRNA-seq data, the default Seurat clustering procedure includes the following steps, applied separately to both the target data and the synthetic null data, each of which is stored as a Seurat object, denoted as `Seurat.obj`.

1. Normalize each cell to have a total count of 10,000; then perform log(normalized count + 1) transformation.
   `NormalizeData(Seurat.obj, normalization.method = "LogNormalize", scale.factor = 10000)`
2. Select 2,000 highly variable genes.
   `FindVariableFeatures(Seurat.obj, selection.method = "vst", nfeatures = 2000)`
3. Scale the data.
   `ScaleData(Seurat.obj)`
4. Run PCA on the data.
   `RunPCA(Seurat.obj, features = VariableFeatures())`
5. Compute cells' k-nearest neighbors.
   `FindNeighbors(Seurat.obj, dims = 1:30, nn.method = "rann", k.param = 20)`
6. Perform Louvain clustering on the cells.
   `FindClusters(Seurat.obj, resolution)`
   Since the Louvain clustering cannot pre-specify the cluster number, we tried resolutions starting from the default resolution of 0.5 and adjusted the resolution until two clusters were found.

For SRT data, we employed the spatial clustering algorithm, BayesSpace, for spatial domain detection. Both the target and synthetic null datasets were stored as SingleCellExperiment objects, each denoted as `sce.obj`.

1. Perform log-normalization and then apply PCA.
   `spatialPreprocess(sce.obj, platform="ST", n.PCs=15, log.normalize=TRUE)`
2. Cluster the spots and add the predicted domain labels to the `sce.obj`.
   `spatialCluster(sce.obj, q=2, platform="ST", d=7, init.method="mclust", model="t", gamma=2, nrep=1000, burn.in=100, save.chain=TRUE)`

After applying the above clustering procedure, we obtained the cluster labels $\widehat{Z}_1, \ldots, \widehat{Z}_n$ from the target data $\mathbf{Y}$, and $\widetilde{Z}_1, \ldots, \widetilde{Z}_n$ from the synthetic null data $\widetilde{\mathbf{Y}}$, respectively, where $\widehat{Z}_i, \widetilde{Z}_i \in \{0, 1\}$, $i = 1, \ldots, n$. Again, there exists no one-to-one correspondence between $\widehat{Z}_1, \ldots, \widehat{Z}_n$ and $\widetilde{Z}_1, \ldots, \widetilde{Z}_n$.

ClusterDE Step 3: DE analysis. ClusterDE supports any DE test in principle. For scRNA-seq data, we employ five DE tests included in the Seurat `FindMarkers` function: the Wilcoxon rank-sum test (Wilcoxon, the default), t-test, negative binomial generalized linear model (NB-GLM), logistic regression model predicting cluster membership with a likelihood-ratio test (LR), and likelihood-ratio test for single-cell gene expression (bimod; see [12]). For SRT data, we use only the Wilcoxon rank-sum test and t-test, which are commonly applied to identify spatial-domain marker genes.

For a given DE test applied to the target dataset, ClusterDE computes a P value P_j for each gene j to test the double-dipping null hypothesis $H_{0j}^{\text{DD}} : \mu_{0j}^{\text{DD}} = \mu_{1j}^{\text{DD}}$, where $\mu_{0j}^{\text{DD}} = \mathbb{E}[Y_{ij} | \widehat{Z}_i = 0]$ and $\mu_{1j}^{\text{DD}} = \mathbb{E}[Y_{ij} | \widehat{Z}_i = 1]$. The target DE score for gene j is then defined as $S_j := -\log_{10} P_j$.

In parallel, on the synthetic null data, ClusterDE calculates a P value $\widetilde{P}_j$ for each gene j to testing the null hypothesis $\widetilde{H}_{0j}^{\text{DD}} : \widetilde{\mu}_{0j}^{\text{DD}} = \widetilde{\mu}_{1j}^{\text{DD}}$, where $\widetilde{\mu}_{0j}^{\text{DD}} = \mathbb{E}[\widetilde{Y}_{ij} | \widetilde{Z}_i = 0]$ and $\widetilde{\mu}_{1j}^{\text{DD}} = \mathbb{E}[\widetilde{Y}_{ij} | \widetilde{Z}_i = 1]$. The null DE score of gene j is defined as $\widetilde{S}_j := -\log_{10} \widetilde{P}_j$.

In summary, ClusterDE generates target DE scores $S_1, \ldots, S_m$ and null DE scores $\widetilde{S}_1, \ldots, \widetilde{S}_m$ for m genes.

ClusterDE Step 4: FDR control. Given the target DE scores $S_1, \ldots, S_m$ and the null DE scores $\widetilde{S}_1, \ldots, \widetilde{S}_m$, we use the FDR-control method Clipper [13] to identify DE genes at a target FDR threshold $q \in (0, 1)$. Given a set of identified DE genes (i.e., discoveries), the FDR is defined as

$$\text{FDR} := \mathbb{E}\left[\frac{\#\text{ false discoveries}}{\#\text{ discoveries } \vee 1}\right],$$

where $a \vee b$ denotes the maximum of two numbers a and b.

To ensure valid FDR control, Clipper requires each gene to have a contrast score such that the true non-DE genes (i.e., non-marker genes defined based on the ideal DE test) have contrast scores symmetric about zero. In ClusterDE, the contrast score for gene j is defined as

$$C_j := S_j - \widetilde{S}_j \,.$$

Then ClusterDE uses Clipper to find a contrast score cutoff T within $\mathcal{C}$ (the set of non-zero contrast score values) given the target FDR threshold q:

$$T := \min\left\{t \in \mathcal{C} : \frac{\text{card}\left(\{j : C_j \leq -t\}\right) + 1}{\text{card}\left(\{j : C_j \geq t\}\right) \vee 1} \leq q\right\}$$

and outputs $\{j \in \{1, \cdots, m\} : C_j \geq T\}$ as discoveries. Here, card(A) defines the cardinality (i.e., size) of a set A. This contrast score thresholding procedure for FDR control is from the knockoffs method [14].

Under the assumption that the majority of genes are true non-DE genes, the distribution of all genes' contrast scores is expected to have a mode at zero, thereby satisfying Clipper's symmetry requirement for contrast scores under the null hypothesis. Ideally, slightly less than 50% of all genes' contrast scores would be negative. However, in practice, this symmetry requirement may not hold in real datasets. For example:

1. The contrast score distribution might have a positive mode, resulting in too few negative contrast scores and inflated false discoveries.
2. Conversely, the distribution could have a negative mode, leading to too many negative contrast scores and reduced statistical power.

To address this, ClusterDE checks the symmetry of contrast scores around zero using Yuen's trimmed mean test (via the function `yuen.t.test()` from the R package PariedData (version 1.1.1)). This test examines the symmetry of the contrast score distribution after excluding the smallest 10% and largest 10% of the contrast scores.

If symmetry is rejected by Yuen's trimmed mean test, ClusterDE adjusts the contrast score distribution to better satisfy the symmetry requirement. In detail, ClusterDE employs "robust fitting of linear models" using the function `rlm()` from the R package MASS (version 7.3-60). A linear model is fitted between the target DE scores (response variable y) and the null DE scores (explanatory variable x), and the predicted values ($\hat{y}$) are used as the adjusted null DE scores. Then the adjusted contrast scores, defined as the differences between the target DE scores and the adjusted null DE scores, are expected to meet the symmetry requirement more closely.

To maintain conservatism in the adjustment process, ClusterDE applies a one-sided ("greater than") Yuen's trimmed mean test at a significance level of 0.001. Hence, adjustment is performed only when there are too few negative contrast scores, a scenario that would likely inflate false discoveries by Clipper.

S1.3 Implementation of the TN test and Countsplit

We compared ClusterDE with two existing methods—the TN test [15] and Countsplit [16]—both designed to address the double-dipping issue in post-clustering DE analysis.

For the TN test, we used the Python module truncated-normal (version 0.4). We followed the GitHub tutorial for the implementation (`https://github.com/jessemzhang/tn_test/blob/master/experiments/experiments_pbmc3k.ipynb`). In the clustering step, we used the same procedure as in ClusterDE Step 2. In the DE analysis step, unlike ClusterDE and Countsplit, the TN test has its own DE test, so we did not use any DE tests from the R package Seurat (version 4.2.0).

For Countsplit, we used the R package countsplit (version 1.0) to split the original count matrix (i.e., the target data used by ClusterDE) into a training matrix (for clustering) and a test matrix (for DE analysis). In the clustering step, we used the same procedure as in ClusterDE Step 2. In the DE analysis step, we used the five DE tests provided in the R package Seurat (version 4.2.0).

S1.4 Alternative strategies for synthetic null data generation

Although the model-X knockoffs method was originally developed for feature selection in multivariate predictive models (e.g., the Lasso) [17], rather than for marginal DE tests (where each feature is analyzed independently), we compared it to ClusterDE because both methods share the concept of generating synthetic null data (or negative-control data) based on real data.

To implement the model-X knockoffs method for post-clustering DE analysis, we used the R package knockoff (version 0.3.6) to construct knockoff data (i.e., the negative control). For feature selection, we applied the default `glmnet` method for binary logistic regression, treating the cluster labels as the response variable (y) and the genes as features. Testing this approach on 50 simulated datasets with logFC $= 2.6$ (see "Simulation setting with two cell types and 200 true DE genes"), we observed that it consistently selected 0 DE genes, resulting in a power of 0.

Moreover, we tested three alternative strategies for synthetic null data generation in ClusterDE Step 1: the knockoff data constructed above, permuted data (where each gene was independently permuted across all cells), and scDesign3 without copula modeling (assuming gene independence) (Fig. S9). Our results on simulated datasets showed that scDesign3 provided more robust FDR control and better statistical power compared to these three alternative strategies for synthetic null data generation (Fig. S10).

S1.5 Simulation designs

To benchmark post-clustering DE methods in terms of FDR and statistical power, we required ground truths for DE genes and non-DE genes. We used the R package scDesign3 (version 0.99.0) [9] to simulate realistic data with true DE genes and non-DE genes, based on model parameters estimated from real data.

1. **Simulation designs for scRNA-seq data**
   Under each simulation setting, we simulated 200 replicates. Each replicate contained $n = 998$ cells and $m = 9239$ genes, matching the dimensions of the naïve cytotoxic T cells in the Zhengmix4eq dataset [18] after Seurat's default preprocessing step, which removed genes expressed at extremely low levels. In the following, we let i and j denote the indices of cells and genes, respectively, $i = 1, \ldots, n$; $j = 1, \ldots, m$. The simulation involved estimating the following **one-cell-type model parameters** from the naïve cytotoxic T cells using `scDesign3`. For details of the model formulation, refer to the section "ClusterDE Step 1: synthetic null data generation."
   - Per-gene NB mean parameter $\mu_j \in \mathbb{R}^+$, $j = 1, \ldots, m$;
   - Per-gene NB dispersion parameter $\sigma_j \in \mathbb{R}^+$, $j = 1, \ldots, m$;
   - Gene-gene correlation matrix in the Gaussian copula $\mathbf{R} \in [-1, 1]^{m \times m}$.

   Using the estimated model parameters $\{\widehat{\mu}_j, \widehat{\sigma}_j\}_{j=1}^m$ and $\widehat{\mathbf{R}}$, we designed two settings: (1) one cell type and (2) two cell types with 200 true DE genes.

   Simulation setting with one cell type. All $n = 998$ cells were simulated as belonging to one cell type, modeled by an MVNB distribution specified by the Gaussian copula with correlation matrix $\widehat{\mathbf{R}}$. Gene j's counts followed an NB distribution with mean $\widehat{\mu}_j$ and dispersion $\widehat{\sigma}_j$, $j = 1, \ldots, m$. The R package `scDesign3` was used to simulate a cell-by-gene count matrix $\mathbf{Y} \in \mathbb{N}_{\geq 0}^{n \times m}$, representing the one-cell-type target data (Fig. 1c) in simulation studies.

   Simulation setting with two cell types and 200 true DE genes. All $n = 998$ cells were assigned to two cell types, each modeled by its own MVNB distribution specified by the Gaussian copula. To simulate each replicate, we randomly selected 200 true DE genes with distinct mean parameters μ_j^0 and μ_j^1 for the two cell types, based on $\widehat{\mu}_j$.
   For the two cell types, we simulated three cell-type size ratios, $r \in \{1, 4, 9\}$, such that cells $i = 1, \ldots, \lfloor n/(r+1) \rfloor$ belonged to cell type 0, and cells $i = \lfloor n/(r+1) \rfloor + 1, \ldots, n$ belonged to cell type 1. For each replicate, we specified 200 true DE genes with the index set $J_{\mathrm{DE}} \subset \{1, \cdots, m\}$. For each true DE gene $j \in J_{\mathrm{DE}}$, we set its mean parameter in cell type 0 as the estimate, i.e., $\mu_j^0 = \widehat{\mu}_j$. Then we

modified its mean parameter in cell type 1, μ_j^1, using a pre-specified log fold change, logFC, with a 50% probability of up-regulation and a 50% probability of down-regulation:

$$\mu_j^1 = \begin{cases} \widehat{\mu}_j \times 2^{\text{logFC}}, & \text{if } Z_j = 1 \\ \widehat{\mu}_j \times 2^{-\text{logFC}}, & \text{if } Z_j = 0 \end{cases}, \quad \text{with } Z_j \sim \text{Ber}(0.5), \quad j \in J_{\text{DE}} \,.$$

For each true non-DE gene j, we set

$$\mu_j^0 = \mu_j^1 = \widehat{\mu}_j, \quad j \in \{1, \cdots, m\} \backslash J_{\text{DE}} \,.$$

The pre-specified logFC determines the differences between the two cell types and is expected to have an inverse relationship with the severity of double dipping (i.e., greater differences between the two cell types lead to better clustering, which in turn reduces the severity of double dipping). Accordingly, we simulated two cell types using a sequence of logFC values

$$\text{logFC} = 1.05, 1.1, \ldots, 1.95, 2, 2.1, \ldots, 2.9, 3.$$

For each logFC value, we simulated cells from cell types 0 and 1, each modeled by an MVNB distribution specified by the Gaussian copula with correlation matrix $\widehat{\mathbf{R}}$. Specifically, gene j's counts in cell types 0 and 1 followed NB distributions with distinct mean parameters μ_j^0 and μ_j^1, respectively, and a shared dispersion parameter $\widehat{\sigma}_j$, $j = 1, \ldots, m$. For detailed simulation procedures, refer to the section "ClusterDE Step 1: synthetic null generation." The R package scDesign3 was used to simulate a cell-by-gene count matrix $\mathbf{Y} \in \mathbb{N}_{\geq 0}^{n \times m}$, which served as the two-cell-type target data (Fig. S2) in simulation studies.

2. **Simulation designs for SRT data**
We assume that the expression of non-domain marker genes changes gradually across spatial locations, while the expression of domain marker genes shows significant discontinuities (sudden changes) at domain boundaries. In our modeling of SRT data, gene j's counts are assumed to follow a negative binomial (NB) distribution with spot-specific mean μ_{ij} and a common dispersion parameter σ_j $(i = 1, \ldots, n; j = 1, \ldots, m)$. Additionally, all genes are modeled to follow an MVNB distribution specified by the Gaussian copula with correlation matrix $\mathbf{R}$. We simulated three settings: (1) one spatial domain; (2) two domains with 200 true domain marker genes; (3) three domains with 200 and 168 true domain marker genes between adjacent domains.

Simulation setting with one spatial domain. A single domain ($n = 989$ spots) was simulated using Layer 3 of slice 151673 from the LIBD human dorsolateral prefrontal cortex (DLPFC) dataset as the reference. We modeled gene j's mean expression as $\log(\mu_{ij}) = \alpha_j + f_j^{GP}(x_{i1}, x_{i2}, K)$, where α_j denotes gene j's specific intercept, and $f_j^{GP}(\cdot)$ represents a smooth function modeled by a penalized Gaussian Process regressor with K bases. We set $K = 10$ to constrain the mean expression surface to be smooth. The R package scDesign3 was used to simulate a spot-by-gene count matrix $\mathbf{Y} \in \mathbb{N}_{\geq 0}^{n \times m}$, based on $\log(\widehat{\mu}_{ij}) = \widehat{\alpha}_j + \widehat{f}_j^{GP}(x_{i1}, x_{i2}, K)$, $\widehat{\sigma}_j$ $(j = 1, \ldots, m)$, and $\widehat{\mathbf{R}}$ estimated from the reference data. The simulated matrix, along with the spots' 2D locations (identical to those in the reference data), served as the one-domain target data for the simulation studies.

Simulation setting with two spatial domains and 200 true spatial-domain marker genes. All $n = 1207$ spots from Layers 3 and 4 of slice 151673 in the DLPFC dataset were used as reference data to simulate SRT data with two domains. We selected 200 genes as the true domain marker genes for Layer 4, meaning their expression values exhibited abrupt changes between the two domains.
To select these 200 genes, for each gene j, we modeled the mean parameter μ_{ij} as:

$$\log(\mu_{ij}) = \alpha_{0j} + \alpha_{1j}\mathbb{I}(Z_i = 1) + f_j^{GP}(x_{i1}, x_{i2}, K),$$

where α_{0j} is the intercept, α_{1j} represents the domain effect, Z_i is the domain label for spot i ($Z_i = 0$ corresponds to Layer 3, and $Z_i = 1$ corresponds to Layer 4), and $K = 10$. The model was fitted to the reference data to obtain parameter estimates $\widehat{\alpha}_{0j}$, $\widehat{\alpha}_{1j}$, and $\widehat{f}_j^{GP}$, as well as the dispersion estimate $\widehat{\sigma}_j$. The 200 true domain marker genes were then selected as those with the largest differences $\widehat{\alpha}_{1j}/\widehat{\alpha}_{0j} - 1$. For the remaining genes, considered true non-marker genes, we modeled the mean parameter μ_{ij} as:

$$\log(\mu_{ij}) = \alpha_j + f_j^{GP}(x_{i1}, x_{i2}, K).$$

This model was fitted to data from both domains of the reference set to obtain estimates $\widehat{\alpha}_j$, $\widehat{f}_j^{GP}$, and the dispersion $\widehat{\sigma}_j$.

To amplify the differences between the two domains, a constant value of 2 was manually added to enhance the domain effect. Consequently, the mean parameter used to simulate a true domain marker gene j for Layer 4 was defined as:

$$\log(\widehat{\mu}_{ij}) = \widehat{\alpha}_{0j} + (\widehat{\alpha}_{1j} + 2)\mathbb{I}(Z_i = 1) + \widehat{f}_j^{\,GP}(x_{i1}, x_{i2}, K).$$

The R package scDesign3 was used to simulate a spot-by-gene count matrix $\mathbf{Y} \in \mathbb{N}_{\geq 0}^{n \times m}$, based on $\log(\widehat{\mu}_{ij})$, $\widehat{\sigma}_j$ ($j = 1, \ldots, m$), and $\widehat{\mathbf{R}}$ estimated from the reference data. The simulated matrix, along with the spots' 2D locations (identical to those in the reference data), served as the two-domain target dataset for the simulation studies.

Simulation setting with three spatial domains and 200 and 168 true spatial-domain marker genes. All $n = 1878$ spots from Layers 5, 6, and WM of slice 151673 in the DLPFC dataset were used as reference data to simulate SRT data with three domains. We selected 200 and 168 genes as true domain marker genes for Layers 5 and 6, respectively, indicating that their expression values exhibited abrupt changes between the adjacent domains.

To select these true domain marker genes, for each gene j, we modeled the mean parameter μ_{ij} as:

$$\log(\mu_{ij}) = \alpha_{0j} + \alpha_{1j}\mathbb{I}(Z_i = 1) + \alpha_{2j}\mathbb{I}(Z_i = 2) + f_j^{GP}(x_{i1}, x_{i2}, K),$$

where α_{0j} is the intercept, α_{1j} and α_{2j} represent domain effects, Z_i is the domain label for spot i ($Z_i = 0$ corresponds to Layer 5, $Z_i = 1$ to Layer 6, and $Z_i = 2$ to Layer WM), and $K = 10$. The model was fitted to the reference data to obtain parameter estimates $\widehat{\alpha}_{0j}$, $\widehat{\alpha}_{1j}$, $\widehat{\alpha}_{2j}$, and $\widehat{f}_j^{GP}$, as well as the dispersion estimate $\widehat{\sigma}_j$. The 200 true domain marker genes for Layer 5 were first selected as those with the largest $|\widehat{\alpha}_{0j}/\widehat{\alpha}_{1j} - 1|$. Then, an additional set of 200 genes for Layer 6 were selected based on the largest $|\widehat{\alpha}_{1j}/\widehat{\alpha}_{2j} - 1|$, with overlap from the first set removed, resulting in 168 unique genes.

For the remaining genes, considered true non-marker genes, we modeled the mean parameter μ_{ij} as:

$$\log(\mu_{ij}) = \alpha_j + f_j^{GP}(x_{i1}, x_{i2}, K).$$

This model was fitted to data from all three domains of the reference set to obtain estimates $\widehat{\alpha}_j$, $\widehat{f}_j^{GP}$, and the dispersion $\widehat{\sigma}_j$.

To amplify the differences between the three domains, constant values of 1 and 2.5 were manually added to the domain-specific intercepts for Layers 5 and 6, respectively. Consequently, the mean parameter used to simulate a true domain marker gene j for Layer 5 was defined as:

$$\log(\widehat{\mu}_{ij}) = (\widetilde{\alpha}_j + 1) + \widetilde{\alpha}_j\mathbb{I}(Z_i = 1) + \widetilde{\alpha}_j\mathbb{I}(Z_i = 2) + \widehat{f}_j^{GP}(x_{i1}, x_{i2}, K),$$

and the mean parameter used to simulate a true domain marker gene j for Layer 6 was defined as:

$$\log(\widehat{\mu}_{ij}) = \widetilde{\alpha}_j + (\widetilde{\alpha}_j + 2.5)\mathbb{I}(Z_i = 1) + \widetilde{\alpha}_j\mathbb{I}(Z_i = 2) + \widehat{f}_j^{GP}(x_{i1}, x_{i2}, K),$$

where $\widetilde{\alpha}_j$ was set to the minimum of $\widehat{\alpha}_{0j}$, $\widehat{\alpha}_{1j}$, and $\widehat{\alpha}_{2j}$.

The R package scDesign3 was used to simulate a spot-by-gene count matrix $\mathbf{Y} \in \mathbb{N}_{\geq 0}^{n \times m}$, based on $\log(\widehat{\mu}_{ij})$, $\widehat{\sigma}_j$ ($j = 1, \ldots, m$), and $\widehat{\mathbf{R}}$ estimated from the reference data. The simulated matrix, along with the spots' 2D locations (identical to those in the reference data), served as the three-domain target dataset for the simulation studies.

S1.6 Real data analysis

Collection of real data. We collected five scRNA-seq datasets of cell lines. Three datasets—A549, H2228, and HCC827—were obtained from the study [19] and downloaded from `https://github.com/LuyiTian/sc_mixology/tree/master/data`. The other two datasets, HEK293T and JURKAT, were retrieved from the study [20] and downloaded from `https://cf.10xgenomics.com/samples/cell-exp/1.1.0/jurkat/jurkat_filtered_gene_bc_matrices.tar.gz` and `https://cf.10xgenomics.com/samples/cell-exp/1.1.0/293t/293t_filtered_gene_bc_matrices.tar.gz`, respectively. Additionally, we collected eight peripheral blood mononuclear cell (PBMC) datasets from the study [21], which were downloaded from `https://github.com/satijalab/seurat-data`. These datasets originated from the same biological sample with two technical replicates (Rep1/Rep2) measured by four protocols: 10X Genomics Versions 2 and 3, Drop-seq, and inDrop. For each dataset, we selected cells labeled as "CD14$^+$ monocytes" and "CD16$^+$ monocytes." We also used a neuronal cell atlas for the adult *Drosophila* from the study [22]. The processed data, including cell annotations and original cell clustering, are available under accession code GSE142789 at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE142789`. For SRT data, we collected the LIBD human dorsolateral prefrontal cortex (DLPFC) dataset from the study [23] and selected the slice 151673, which can be downloaded from `http://research.libd.org/spatialLIBD/`. Additionally, we downloaded human primary pancreatic cancer (PDAC) SRT data from the study [24], available under accession code GSE111672 `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111672`. From this dataset, we selected the PDAC-B tumor section for real data analysis.

Data preprocessing. We filtered out lowly expressed genes from the datasets as follows. For the cell line datasets A549, H2228, and HCC827, genes expressed in fewer than 20% of cells were removed, while for the cell line datasets HEK293T and JURKAT, genes expressed in fewer than 10% of cells were excluded. For PBMC datasets, we removed genes expressed in fewer than 10% of the selected monocyte cells (CD14$^+$ and CD16$^+$ monocytes). When performing the default Seurat clustering, Seurat automatically filtered out cells with fewer than three expressed genes and genes expressed in fewer than 200 cells. For the *Drosophila* dataset, the original authors in [22] applied stringent filtering criteria: cells were required to have a minimum of 1,000 UMIs, with mitochondrial genes comprising less than 10% of UMIs, and genes expressed in fewer than three cells across all experiments were removed. For Louvain clustering, the authors performed a grid search across resolution parameters and the number of principal components (PCs) to identify optimal clustering parameters, ultimately selecting a resolution of 10 and 120 PCs. Overclustering was evaluated using the `AssessNodes` function in the R package Seurat v.2, which builds a random forest model to separate two potential clusters and reports the out-of-bag error (OOBE). Clusters were required to have an OOBE below 5% and clear differences in between-cluster DE genes. For the slice 151673 in the DLPFC dataset, genes expressed in fewer than 20% of spots were removed. For the PDAC-B tumor section in the PDAC dataset, genes expressed in fewer than 5% of spots were excluded.

Dimensionality reduction and visualization. To visualize the high-dimensional single-cell data, we first applied the PF-logPF transformation to the cell-by-gene count matrix, as described in [25]. We then calculated the top 50 principal components (PCs) of the transformed matrix using the R package irlba (version 2.3.5.1). These PCs were subsequently used as input to the R package umap (version 0.2.10.0) to project the cells from the 50-dimensional PC space to a 2-dimensional UMAP space. For the *Drosophila* dataset, the original authors used 120 PCs as input for t-SNE.

When comparing the target data with the synthetic null data, we calculated the PCs and UMAPs jointly by concatenating the two datasets. This ensured that the target cells and synthetic null cells were projected into the same 2-dimensional UMAP space.

All plots were generated using the R package ggplot2 (version 3.4.2).

For the UMAP visualizations in Fig. 2f, we truncated each gene's normalized expression levels to the 99th percentile to better illustrate gene expression patterns.

Gene set enrichment analysis. We performed gene set enrichment analysis (GSEA) using the R package clusterProfiler (version 4.4.4). The test method used was `fgsea`, with the number of permutations set to

100,000. The gene set "CD14$^+$/CD16$^+$ Monocyte Markers" was obtained from the original study [21] and downloaded from `https://bitbucket.org/jerry00/scumi-dev/raw/61f7f001d20b2fc8fa7c2f4f4147bff1b0d620d8/R/marker_gene/human_pbmc_marker.rda`. The gene set "Housekeeping Genes" (`HSIAO_HOUSEKEEPING_GENES`) was sourced from the Molecular Signature Database (MSigDB) and originated from the study [26].

Gating plots. We used gating plots to compare two clusters at a time. DE genes between the two clusters were identified using the Wilcoxon rank-sum test. DE genes were included as marker genes only if their adjusted P values were less than 0.05 and their absolute log-fold changes between the two clusters exceeded 0.5. Identified markers were categorized into positive and negative markers for each cluster based on the log-fold change. To calculate the marker rate, we determined the proportion of counts attributed to each set of marker genes out of the total gene count for each cell.

Heatmaps. We generated gene expression heatmaps using scaled data, with genes ordered by hierarchical clustering. The heatmaps were created using the `heatmap.2` function from the R package gplots (version 3.1.3).

Machine Learning. For each machine learning model, the *Drosophila* visual atlas scRNA-seq data was subset to include only cell types Tm1 and Tm2, along with one of the three specified gene sets: (1) all 105 DE genes identified by ClusterDE, (2) the top 105 genes ranked by increasing P value from the Wilcoxon rank-sum test in Seurat, and (3) a random selection of 105 genes. The data was split into an 80-20 train-test split, and this process was repeated 100 times. A new model was trained for each split using the genes from the respective set. Random forest models were trained using the `randomForest` function from the R package randomForest (version 4.7.1.1), and SVM models were trained using the `svm` function from the R package e1071 (version 1.7.13).

Validity checks of the contrast scores of ClusterDE and the P values of Seurat, Countsplit, and the TN test For ClusterDE, the primary assumption is that the contrast scores of true non-DE genes are symmetric around zero. In Fig. S5 (left), we evaluated the symmetry of the contrast scores of ClusterDE using five DE tests (Wilcoxon, t-test, NB-GLM, LR, and bimod; corresponding to Fig. S5a–e, left) on a simulated one-cell-type dataset where all genes are true non-DE genes (see "Simulation setting with one cell type" in the Supplementary Methods). The dataset used is one of the 200 simulated replicates.

For Seurat, Countsplit, and the TN test, the validity of their FDR control relies on the assumption that the P values of true non-DE genes follow the Uniform$[0, 1]$ distribution. To examine this, we divided the genes in the same simulated dataset into two groups using hierarchical clustering (via the default R function `hclust()`) applied to the estimated correlation matrix $\widehat{\mathbf{R}}$ from the Gaussian copula. Due to the block pattern in $\widehat{\mathbf{R}}$ (Fig. 1c), the two groups consisted of genes that are highly correlated and those that are less correlated, respectively. We analyzed the P values of the genes in these two groups separately, noting that all genes in this simulated dataset are true non-DE genes. In Fig. S5 (middle and right), we plotted histograms of the P values and the quantile-quantile plots (Q-Q plots) of the negative log-transformed P values for Seurat and Countsplit (both using the five DE tests; corresponding to Fig. S5a–e) and for the TN test (using its own test; the same panels repeated five times in Fig. S5a–e). To quantify the deviation of the P value distribution from the theoretical Uniform$[0, 1]$ distribution, we used the R function `KL.empirical` (from the R package entropy, version 1.3.1)) to calculate the empirical Kullback–Leibler divergence (KL divergence). A larger KL divergence indicates a more severe violation of the P value uniformity assumption. The results demonstrated that Countsplit and the TN test produced nearly uniform P values for the uncorrelated gene group. However, their P values deviated substantially from the uniform distribution for the correlated gene group.

S2 Theoretical justification of ClusterDE

In this section, we explain why ClusterDE can asymptotically control the false discovery rate (FDR) and avoid FDR inflation caused by double dipping. The theoretical justification is based on the results of the mirror statistics framework, which relies on data splitting [27]. Recall that ClusterDE leverages a multivariate count model [9] to generate synthetic null data and employs the P-value-free FDR control framework, Clipper [13], to identify DE genes.

To achieve FDR control in ClusterDE, the contrast scores $\{C_j : j = 1, \ldots, m\}$ must satisfy the following assumptions. Recall that the contrast scores $\{C_j : j = 1, \ldots, m\}$ are derived from the target DE scores $S_1, \ldots, S_m$ and the null DE scores $\widetilde{S}_1, \ldots, \widetilde{S}_m$, and are defined as:

$$C_j := S_j - \widetilde{S}_j \,.$$

Let $\mathcal{N} \subset \{1, \ldots, m\}$ denote the set of true non-DE genes, and define $m_0 = \text{card}(\mathcal{N})$.

Assumption 1. For each true non-DE gene $j \in \mathcal{N}$, the distribution of C_j is symmetric about 0.

Assumption 2. We assume that the contrast scores $\{C_j : j = 1, \ldots, m\}$ are continuous random variables. Moreover, there exist constants $c > 0$ and $\alpha \in (0, 2)$ such that for any threshold $t \in \mathbb{R}$,

$$\text{Var}\left(\sum_{j\in\mathcal{N}} \mathbb{I}(C_j > t)\right) \leq c m_0^{\alpha} \,.$$

Assumption 2 only restricts the correlations among the non-DE genes. In one extreme scenario, if the non-DE genes are all independent, we have:

$$\text{Var}\left(\sum_{j\in\mathcal{N}} \mathbb{I}(C_j > t)\right) = \sum_{j\in\mathcal{N}} \text{Var}\left(\mathbb{I}(C_j > t)\right) \leq c m_0 \,,$$

where c can be any constant no smaller than $1/4$, the maximum variance of a binary random variable. In this case, $\alpha = 1$, and Assumption 2 holds. In the other extreme scenario, if all non-DE genes have identical contrast scores, then:

$$\text{Var}\left(\sum_{j\in\mathcal{N}} \mathbb{I}(C_j > t)\right) = \text{Var}\left(\mathbb{I}(C_j > t)\right) \cdot m_0^2 \,, \quad \forall j \in \mathcal{N} \,,$$

and Assumption 2 does not hold. More generally, if all contrast scores have constant pairwise correlation or can be clustered into a fixed number of groups with constant within-group correlation, then α must be 2, and Assumption 2 does not hold. Hence, Assumption 2 imposes restrictions on the extent and strength of correlations among the contrast scores of non-DE genes but is less stringent than assuming independence among these contrast scores. For instance, if only a small proportion of non-DE genes' contrast scores are correlated—a realistic scenario in biology—Assumption 2 may still hold.

Theorem 1. *We define the false discovery proportion (FDP) at any contrast score cutoff t as*

$$\text{FDP}(t) = \frac{\text{card}\left(\{j \in \mathcal{N} : C_j \geq t\}\right)}{\text{card}\left(\{j : C_j \geq t\}\right) \vee 1} \,,$$

its estimate, used to determine the contrast score cutoff for FDR control, as

$$\widehat{\text{FDP}}(t) = \frac{\text{card}\left(\{j : C_j \leq -t\}\right) + 1}{\text{card}\left(\{j : C_j \geq t\}\right) \vee 1} \,,$$

and the contrast score cutoff given the target FDR $q \in (0, 1)$ as

$$\tau_q = \min\{t > 0 : \widehat{\text{FDP}}(t) \leq q\}.$$

Assume that there exists a constant $t_q > 0$ *such that* $\mathbb{P}(\mathrm{FDP}(t_q) \leq q) \to 1$ *as* $m \to \infty$*. Then, under Assumptions 1 and 2, we have:*

$$\mathrm{FDP}(\tau_q) \leq q + o_m(1),$$

and

$$\limsup_{n,m\to\infty} \mathrm{FDR}(\tau_q) \leq q.$$

The existence of $t_q > 0$ such that $\mathbb{P}(\mathrm{FDP}(t_q) \leq q) \to 1$ as $m \to \infty$ ensures that the data-dependent contrast score cutoff τ_q is bounded from above with probability approaching 1, thus does not diverge to infinity. Theorem 1 can be proved based on the proof in [27].

S3 Supplementary Figures

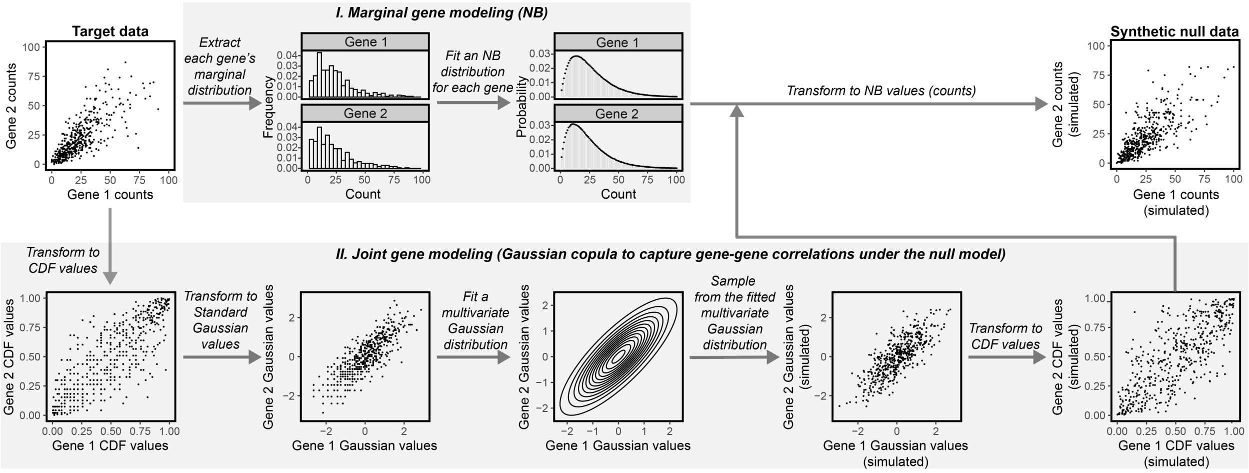

I. Marginal gene modeling (NB)
Target data
Gene 2 counts
Gene 1 counts
Extract each gene's marginal distribution
Gene 1
Gene 2
Frequency
Count
Fit an NB distribution for each gene
Probability
Transform to NB values (counts)
Synthetic null data
Gene 2 counts (simulated)
Gene 1 counts (simulated)
Transform to CDF values
II. Joint gene modeling (Gaussian copula to capture gene-gene correlations under the null model)
Gene 2 CDF values
Gene 1 CDF values
Transform to Standard Gaussian values
Gene 2 Gaussian values
Gene 1 Gaussian values
Fit a multivariate Gaussian distribution
Sample from the fitted multivariate Gaussian distribution
Gene 2 Gaussian values (simulated)
Gene 1 Gaussian values (simulated)
Transform to CDF values
Gene 2 CDF values (simulated)
Gene 1 CDF values (simulated)

Fig. S1: The process for generating synthetic null data from target scRNA-seq data (top left). For illustration, we show the bivariate case (two genes), whereas the actual case involves high-dimensional data with thousands of genes. The null model consists of two components: **marginal gene modeling** and **joint gene modeling**.

For marginal gene modeling (top), each gene's counts are modeled using a negative binomial (NB) distribution. The two NB parameters (mean and dispersion) are estimated from the target data for each gene. For joint gene modeling part (bottom), there are three steps. (1) *Transformation to cumulative distribution function (CDF) values*: Each gene's counts in the target data are transformed into CDF values using either the fitted NB distribution or the empirical distribution (if the target data contains a large number of cells), so the gene's CDF values are uniform between 0 and 1. (2) *Transformation to Gaussian values*: Each gene's CDF values are transformed into quantiles of the standard Gaussian distribution, $N(0, 1)$. (3) *Modeling gene dependence*: A multivariate Gaussian distribution (illustrated here as a bivariate Gaussian) is fitted to the transformed Gaussian values of the genes whose correlations are to be modeled. The correlation matrix of this fitted multivariate Gaussian distribution specifies the Gaussian copula, capturing the gene dependence structure.

After the marginal and joint distributions are modeled, synthetic counts are generated in three steps. (1) *Sampling Gaussian values*: Standard Gaussian values for genes are jointly sampled from the fitted multivariate Gaussian distribution. (2) *Transformation to CDF values*: The sampled Gaussian values are transformed into CDF values of the standard Gaussian distribution. (3) *Transformation to counts*: The CDF values are transformed into quantiles of the fitted NB distribution for each gene, becoming synthetic counts that constitute the synthetic null data (top right).

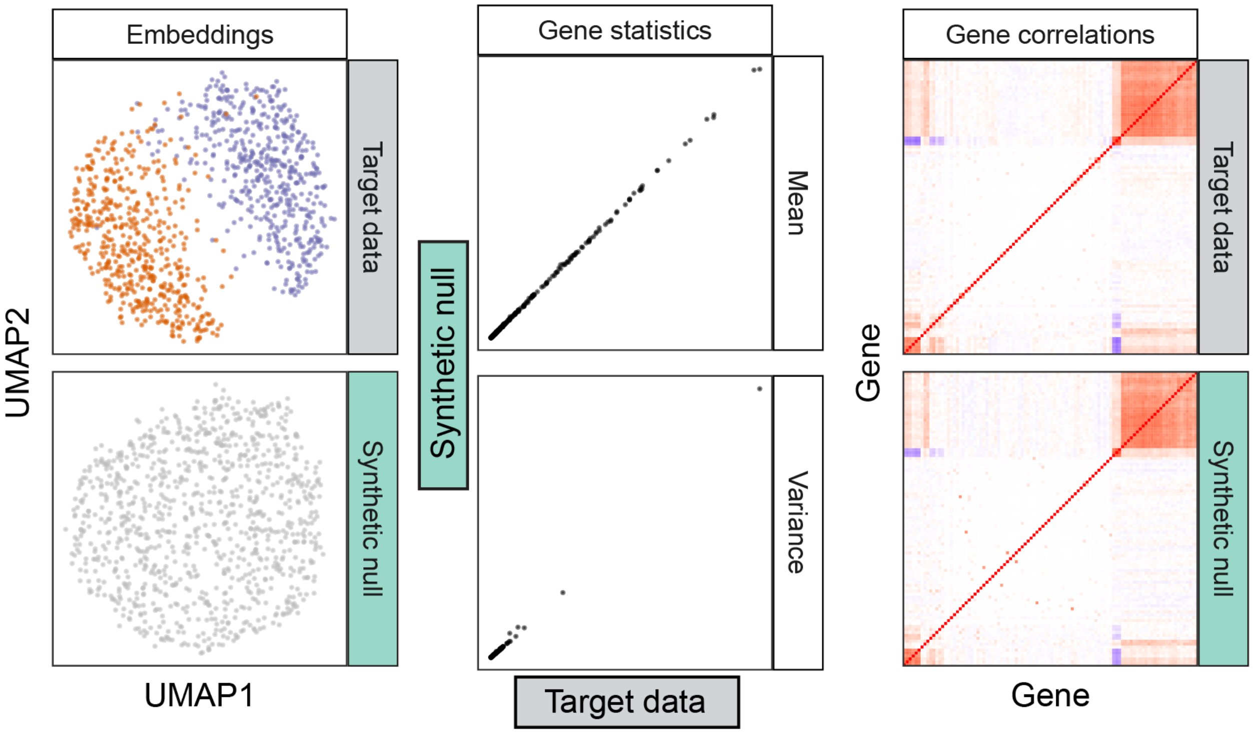


Fig. S2: When the target scRNA-seq data contains cells from two cell types (simulation; see "Simulation setting with two cell types and 200 true DE genes" in the Supplementary Methods), the synthetic null data generated by ClusterDE fills the gap between the two cell types but resembles the target data in other visual aspects of UMAP cell embeddings (left), per-gene expression mean and variance statistics (middle), and gene-gene correlations (right).

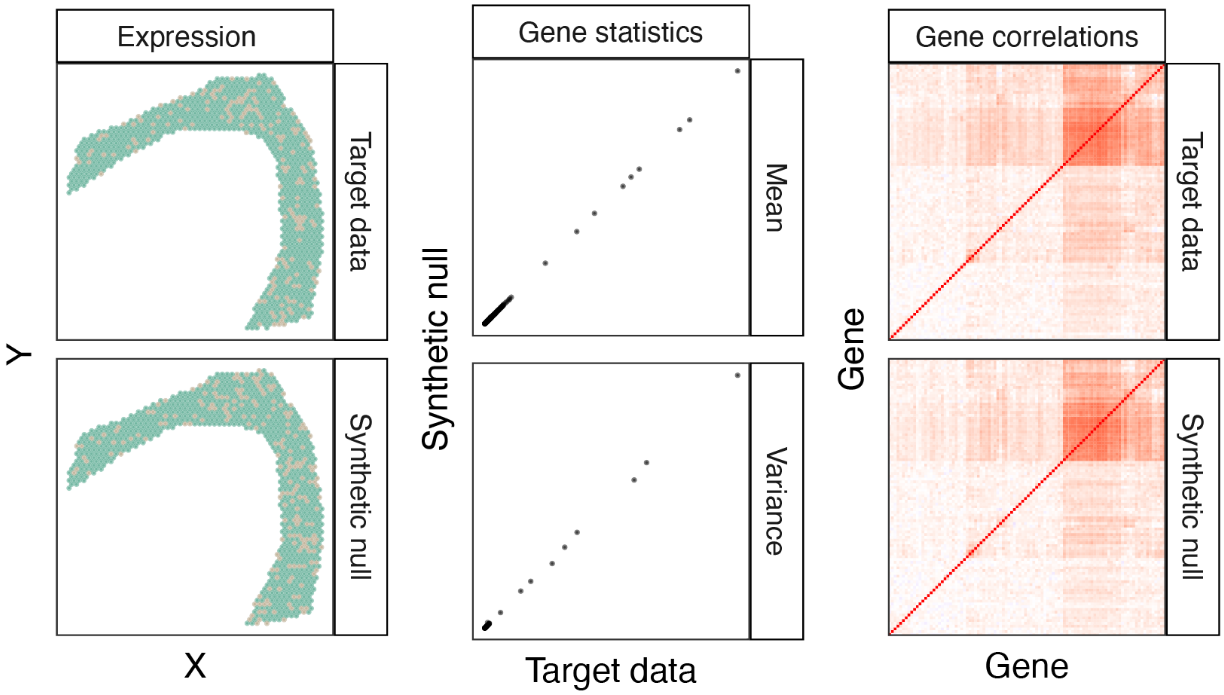


Fig. S3: When the target SRT data contains one spatial domain (simulation; see "Simulation setting with one spatial domain" in the Supplementary Methods), the synthetic null data generated by ClusterDE resembles the target data (left) while preserving per-gene expression mean and variance statistics (middle) as well as gene-gene correlations (right).

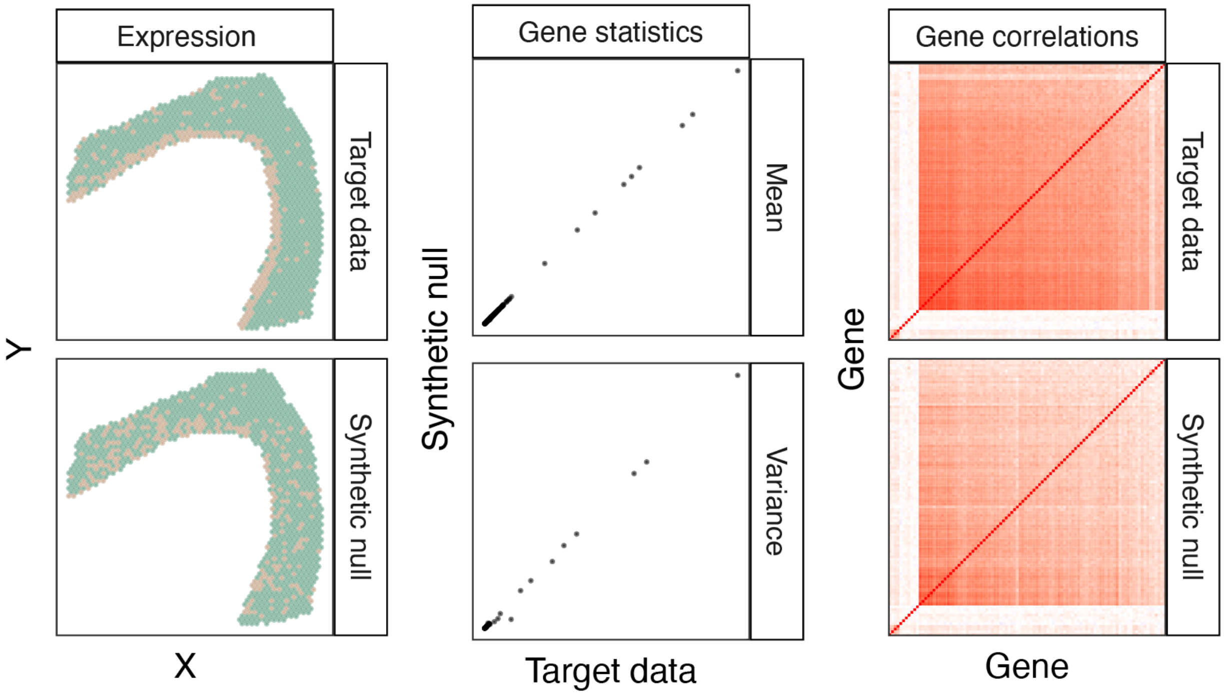


Fig. S4: When the target SRT data contains two spatial domains (simulation; see "Simulation setting with two spatial domains and 200 true spatial-domain marker genes" in the Supplementary Methods), the synthetic null data generated by ClusterDE blurs the domain boundary (left) while preserving per-gene expression mean and variance statistics (middle) as well as gene-gene correlations (right).

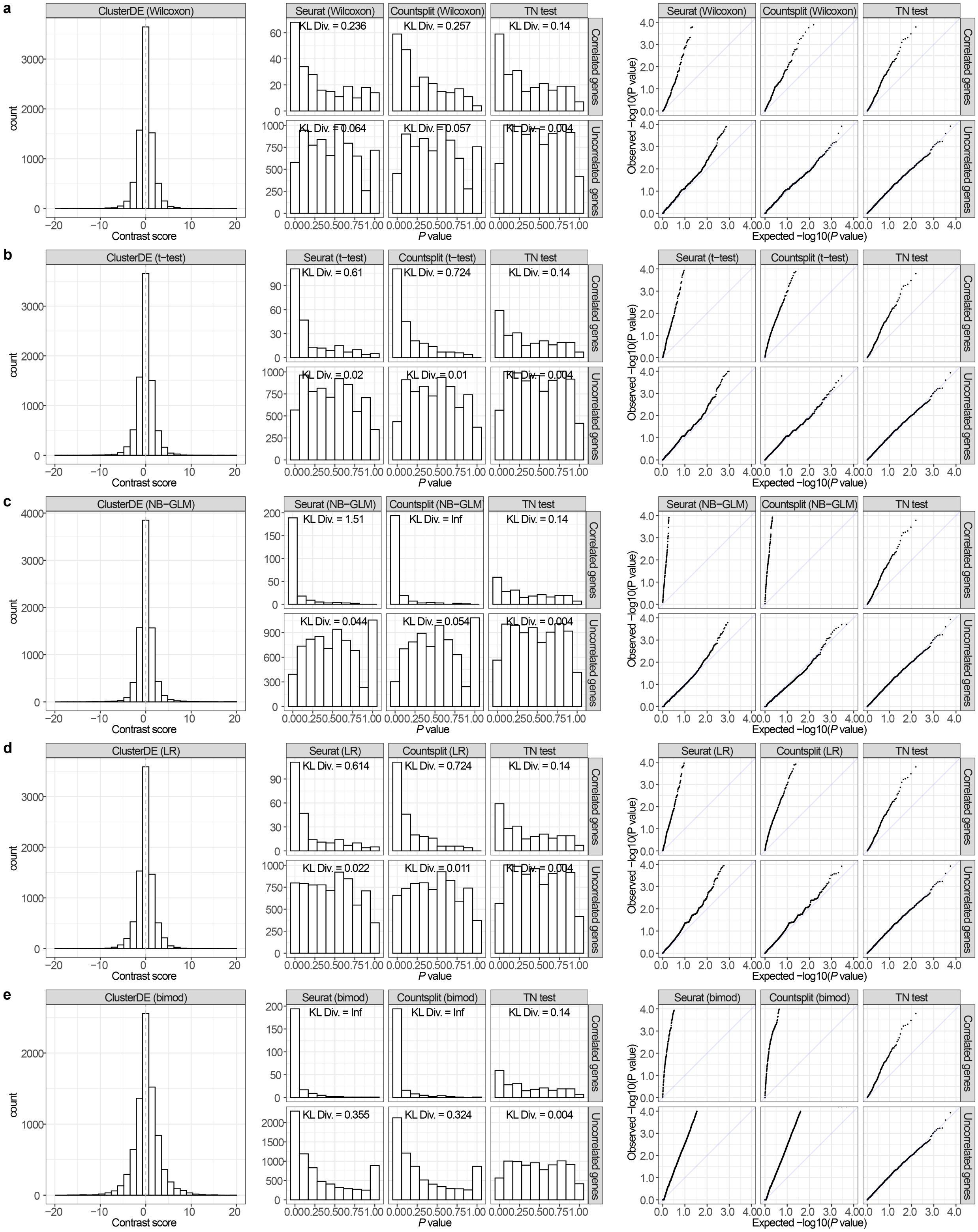

a
ClusterDE (Wilcoxon)
Seurat (Wilcoxon)
Countsplit (Wilcoxon)
TN test
KL Div. = 0.236
KL Div. = 0.257
KL Div. = 0.14
KL Div. = 0.064
KL Div. = 0.057
KL Div. = 0.004
b
ClusterDE (t−test)
Seurat (t−test)
Countsplit (t−test)
TN test
KL Div. = 0.61
KL Div. = 0.724
KL Div. = 0.14
KL Div. = 0.02
KL Div. = 0.01
KL Div. = 0.004
c
ClusterDE (NB−GLM)
Seurat (NB−GLM)
Countsplit (NB−GLM)
TN test
KL Div. = 1.51
KL Div. = Inf
KL Div. = 0.14
KL Div. = 0.044
KL Div. = 0.054
KL Div. = 0.004
d
ClusterDE (LR)
Seurat (LR)
Countsplit (LR)
TN test
KL Div. = 0.614
KL Div. = 0.724
KL Div. = 0.14
KL Div. = 0.022
KL Div. = 0.011
KL Div. = 0.004
e
ClusterDE (bimod)
Seurat (bimod)
Countsplit (bimod)
TN test
KL Div. = Inf
KL Div. = Inf
KL Div. = 0.14
KL Div. = 0.355
KL Div. = 0.324
KL Div. = 0.004
count
Contrast score
P value
Correlated genes
Uncorrelated genes
Observed −log10(P value)
Expected −log10(P value)

Fig. S5: Validity checks of the contrast scores of ClusterDE and P values of Seurat, Countsplit, and the TN test on an exemplary one-cell-type dataset, where no post-clustering DE genes are true cell-type markers by simulation design (see "Simulation setting with one cell type" in the Supplementary Methods). The panels (rows) **a**–**e** represent the five DE tests in Seurat used in ClusterDE, Seurat, and Countsplit (see "ClusterDE step 3: DE analysis" in the Supplementary Material); since the TN test has its own DE test, its results are the same in the panels **a**–**e**. The first column shows that the ClusterDE contrast scores for all genes (true non-marker genes) are approximately symmetric around 0, which meets the assumption of ClusterDE for the FDR control. The second column shows the histograms of the P values of the correlated genes (top) and the uncorrelated genes (bottom) from Seurat, Countsplit, and the TN test. A larger empirical Kullback-Leibler divergence (KL div.) between the P value distribution and the theoretical Uniform$[0, 1]$ distribution represents a more severe violation of the P value uniformity assumption. The results show that Countsplit (for four out of the five DE tests) and the TN test have close-to-uniform P values for the uncorrelated genes, but their P values exhibit a severe departure from the uniform distribution for the correlated genes. The third column contains the quantile-quantile plots of the negative log-transformed P values corresponding to the second column.

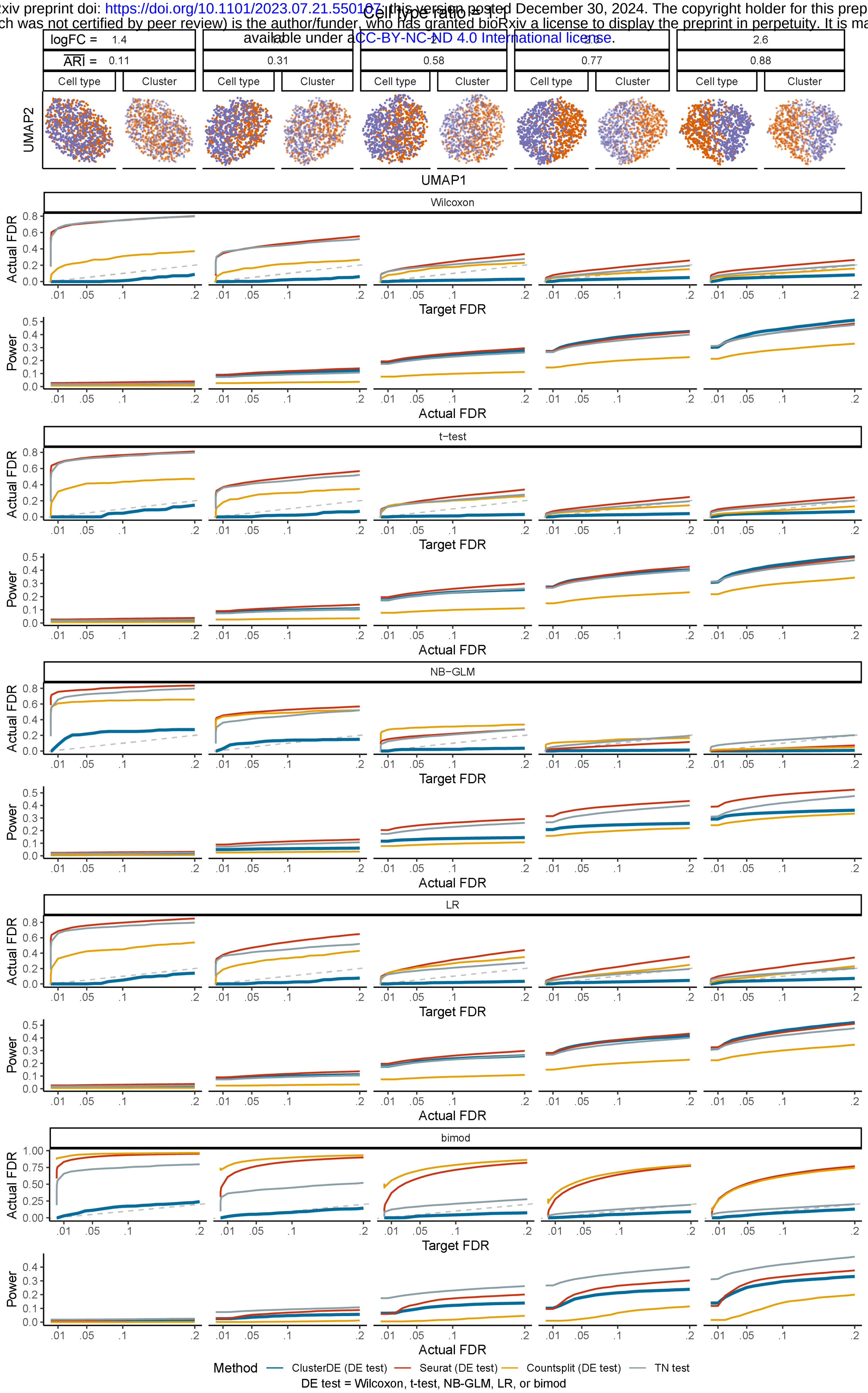

Cell type ratio = 1:1
logFC = 1.4
1.7
2
2.3
2.6
ARI = 0.11
0.31
0.58
0.77
0.88
Cell type
Cluster
UMAP2
UMAP1
Wilcoxon
t−test
NB−GLM
LR
bimod
Actual FDR
Target FDR
Power
.01 .05 .1 .2
0.0 0.2 0.4 0.6 0.8
0.0 0.1 0.2 0.3 0.4 0.5
0.00 0.25 0.50 0.75 1.00
Method
ClusterDE (DE test)
Seurat (DE test)
Countsplit (DE test)
TN test
DE test = Wilcoxon, t-test, NB-GLM, LR, or bimod

Fig. S6: Comparison of the FDRs and power of ClusterDE with three existing methods (Seurat, Countsplit, and the TN test) under varying levels of double-dipping severity when the two cell types have a size ratio of 1 : 1. The log fold change (logFC) represents the average gene expression difference between the two cell types in the simulation (see "Simulation setting with two cell types and 200 true DE genes" in the Supplementary Methods). A small logFC leads to a small adjusted Rand index (ARI), indicating poor agreement between cell clusters and cell types and representing a more severe double-dipping issue. Across different levels of double-dipping severity and the five DE tests, ClusterDE consistently controls the FDRs below the target thresholds (gray diagonal dashed line) and, under the default Wilcoxon rank-sum test, achieves comparable or higher power relative to the three existing methods at equivalent actual FDR levels.

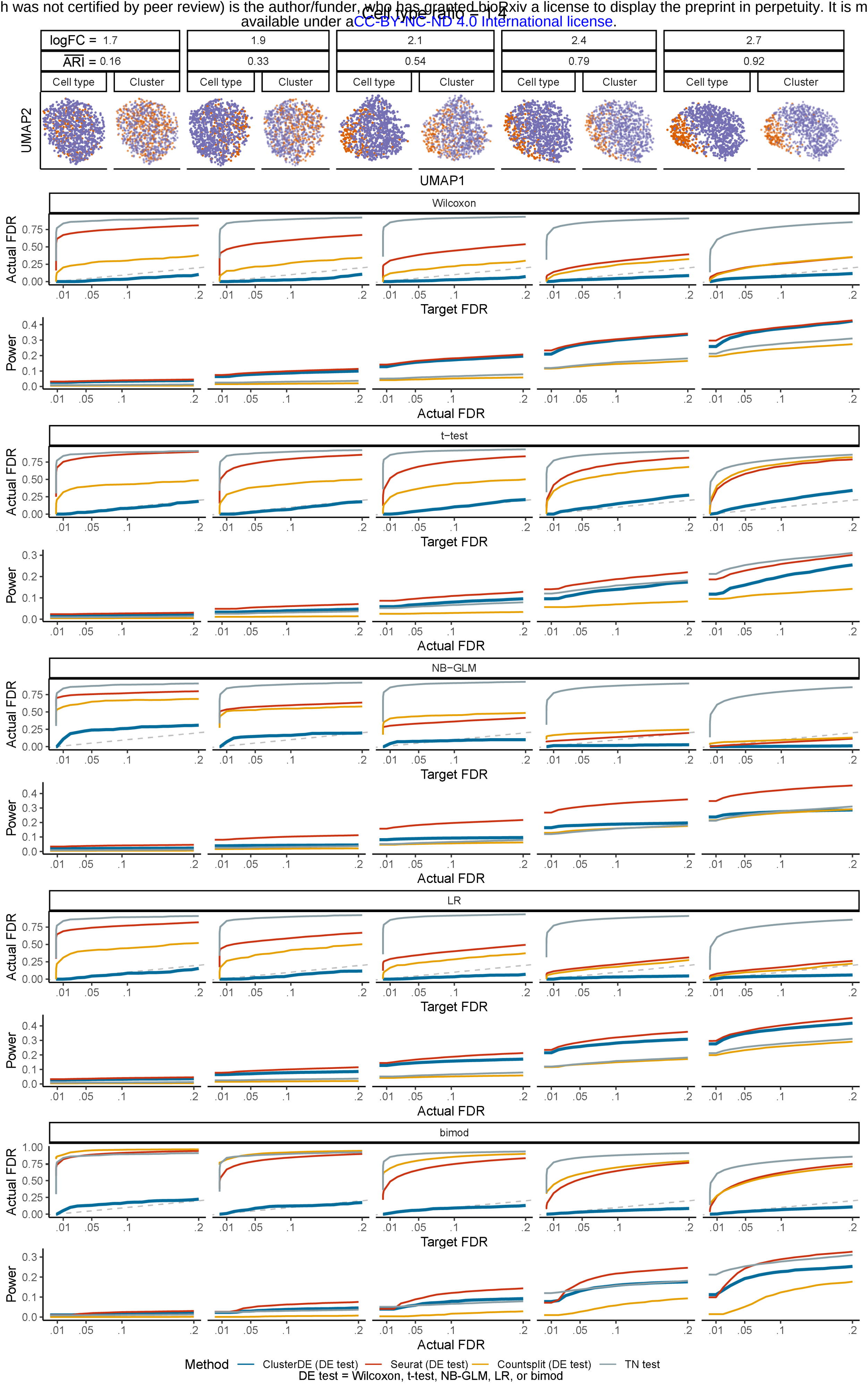

Cell type ratio = 1:4
logFC = 1.7
1.9
2.1
2.4
2.7
$\overline{\text{ARI}}$ = 0.16
0.33
0.54
0.79
0.92
Cell type
Cluster
UMAP2
UMAP1
Wilcoxon
t−test
NB−GLM
LR
bimod
Actual FDR
0.00
0.25
0.50
0.75
1.00
Target FDR
Power
0.0
0.1
0.2
0.3
0.4
.01
.05
.1
.2
Actual FDR
Method
ClusterDE (DE test)
Seurat (DE test)
Countsplit (DE test)
TN test
DE test = Wilcoxon, t-test, NB-GLM, LR, or bimod

Fig. S7: Comparison of the FDRs and power of ClusterDE with three existing methods (Seurat, Countsplit, and the TN test) under varying levels of double-dipping severity when the two cell types have a size ratio of 1 : 4. The log fold change (logFC) represents the average gene expression difference between the two cell types in the simulation (see "Simulation setting with two cell types and 200 true DE genes" in the Supplementary Methods). A small logFC leads to a small adjusted Rand index (ARI), indicating poor agreement between cell clusters and cell types and representing a more severe double-dipping issue. Across different levels of double-dipping severity and the five DE tests, ClusterDE consistently controls the FDRs below the target thresholds (gray diagonal dashed line) and, under the default Wilcoxon rank-sum test, achieves comparable or higher power relative to the three existing methods at equivalent actual FDR levels.

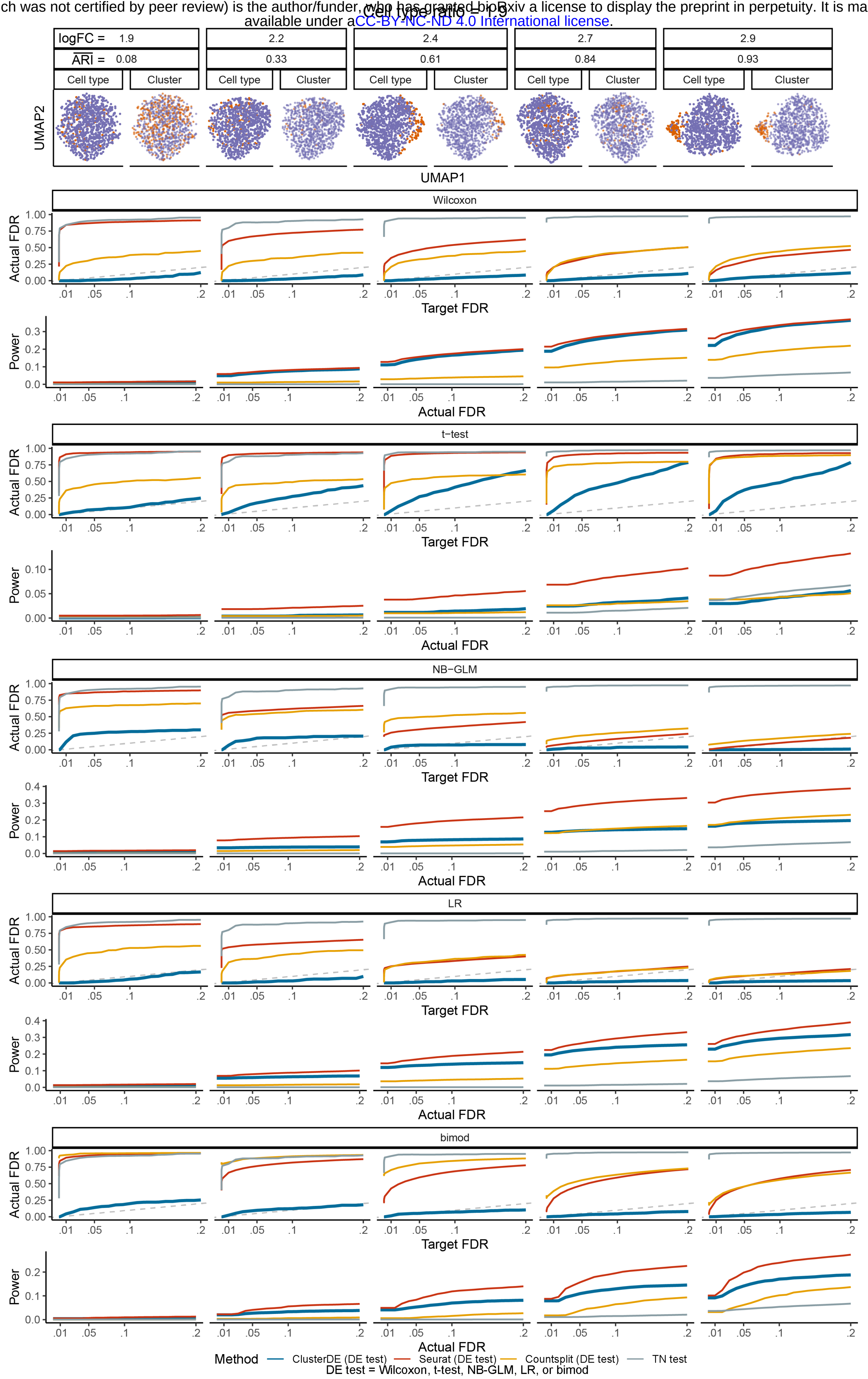

Cell type ratio = 1:9
logFC = 1.9
2.2
2.4
2.7
2.9
$\overline{\text{ARI}}$ = 0.08
0.33
0.61
0.84
0.93
Cell type
Cluster
UMAP2
UMAP1
Wilcoxon
t−test
NB−GLM
LR
bimod
Actual FDR
1.00
0.75
0.50
0.25
0.00
Target FDR
Power
0.4
0.3
0.2
0.1
0.0
0.10
0.05
0.00
.01
.05
.1
.2
Actual FDR
Method
ClusterDE (DE test)
Seurat (DE test)
Countsplit (DE test)
TN test
DE test = Wilcoxon, t-test, NB-GLM, LR, or bimod

Fig. S8: Comparison of the FDRs and power of ClusterDE with three existing methods (Seurat, Countsplit, and the TN test) under varying levels of double-dipping severity when the two cell types have a size ratio of 1 : 9. The log fold change (logFC) represents the average gene expression difference between the two cell types in the simulation (see "Simulation setting with two cell types and 200 true DE genes" in the Supplementary Methods). A small logFC leads to a small adjusted Rand index (ARI), indicating poor agreement between cell clusters and cell types and representing a more severe double-dipping issue. Across different levels of double-dipping severity and the five DE tests, ClusterDE consistently controls the FDRs below the target thresholds (gray diagonal dashed line) and, under the default Wilcoxon rank-sum test, achieves comparable or higher power relative to the three existing methods at equivalent actual FDR levels.

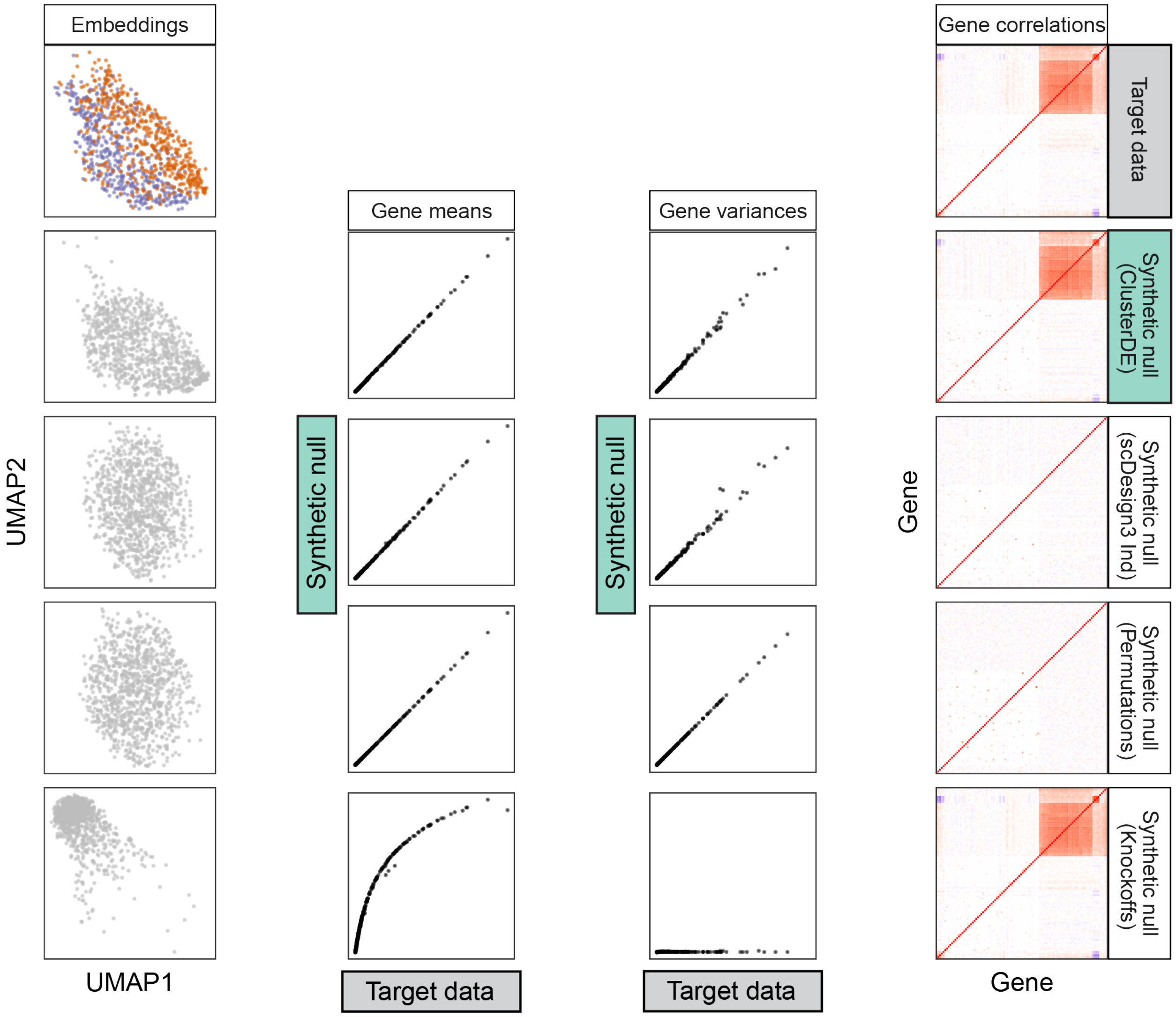


Fig. S9: When the target data contains cells from two cell types (simulation; see "Simulation setting with two cell types and 200 true DE genes" in the Supplementary Methods), the synthetic null data generated by ClusterDE (second row) effectively fills the gap between the two cell types while closely resembling the target data in other visual aspects of UMAP cell embeddings (left), per-gene expression mean and variance statistics (middle), and gene-gene correlations (right). In contrast, the synthetic null data generated by scDesign3 without copula modeling ("scDesign3 Ind," third row), permutations (fourth row) and the model-X knockoffs (fifth row) do not resemble the target data. Specifically, the synthetic null data generated by the model-X knockoffs preserves the gene-gene correlations of the target data but does not preserve per-gene expression mean and variance statistics. Conversely, the synthetic null data generated by scDesign3 Ind and permutations preserves per-gene expression mean and variance statistics but does not preserve gene-gene correlations.

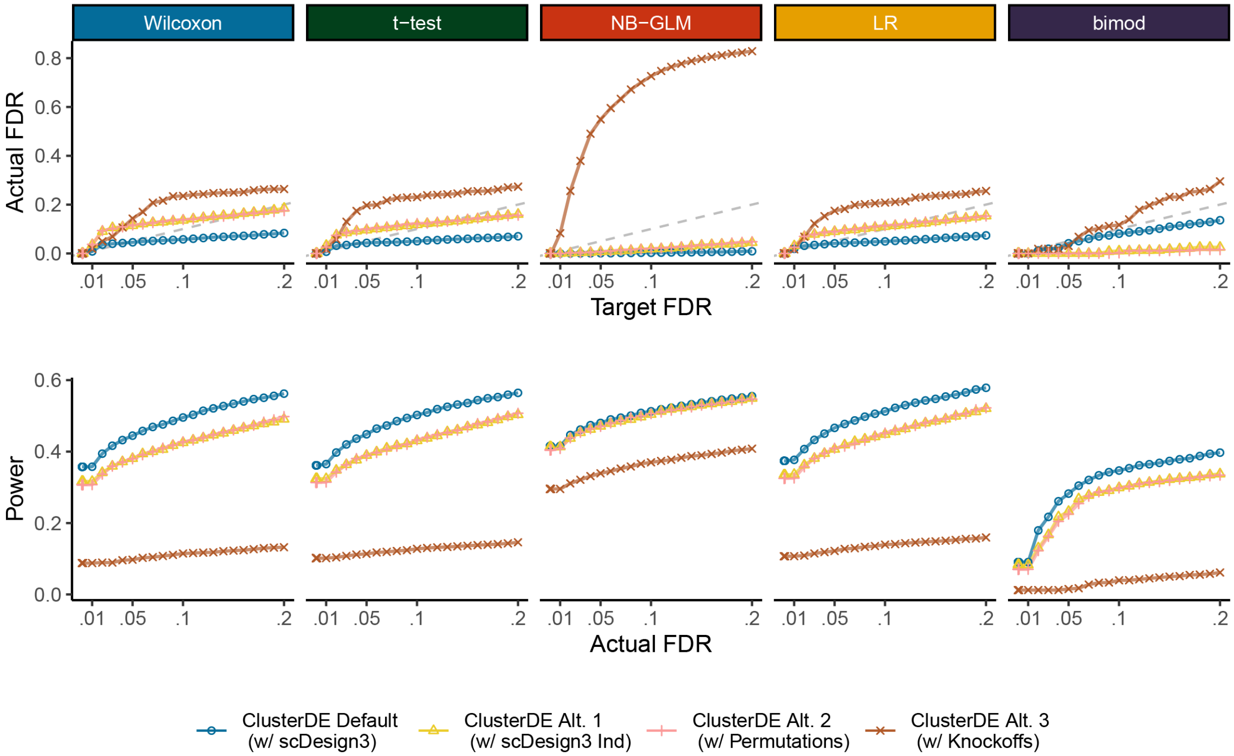


Fig. S10: Comparison of the FDRs and power of ClusterDE with four strategies for synthetic null data generation in Step 1: scDesign3 (the default in ClusterDE), scDesign3 Ind (without copula modeling, assuming gene independence), independent permutations of all genes across cells, and the model-X knockoffs. The default ClusterDE, employing scDesign3, demonstrates superior performance by effectively controlling the FDR and achieving higher power compared to the three alternative strategies.

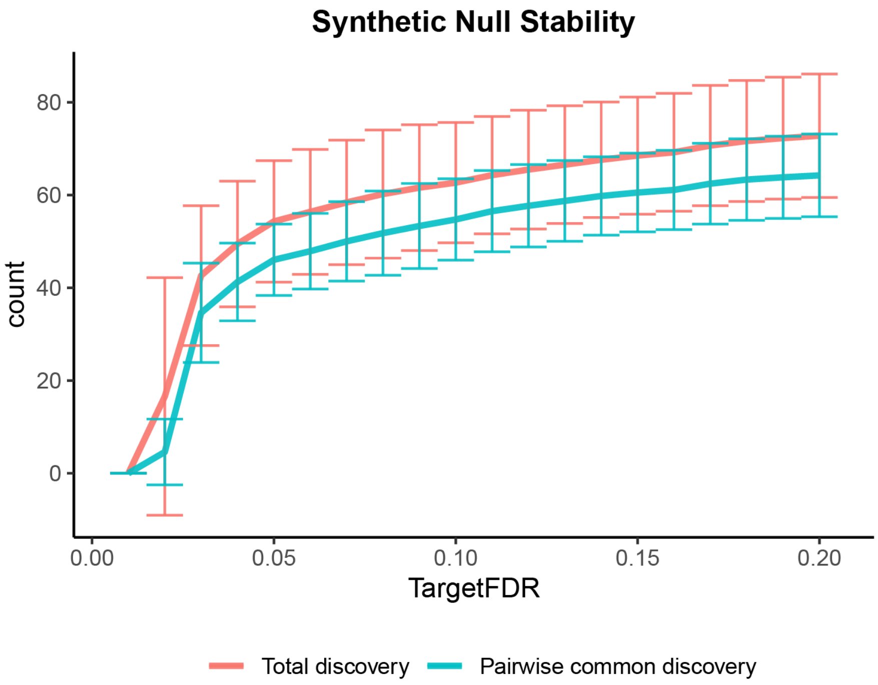


Fig. S11: Stability of the DE genes identified by ClusterDE with respect to the randomness of synthetic null data generation. Using a target dataset simulated with two cell types (see "Simulation setting with two cell types and 200 true DE genes" in the Supplementary Methods), 50 synthetic null datasets are generated with 50 random seeds, and DE genes are identified by ClusterDE at varying target FDRs for each synthetic null dataset. The red curve represents the mean number of DE genes identified at each target FDR, with the standard deviation shown as half the height of the vertical error bars, calculated across the 50 random seeds. The cyan curve represents the mean number of DE genes shared between results from two random seeds, with the standard deviation (half the height of the vertical error bars) calculated across all $\binom{50}{2}$ pairs of random seeds at each target FDR. The results indicate that DE genes identified by ClusterDE are relatively stable and robust to the randomness of synthetic null data generation.

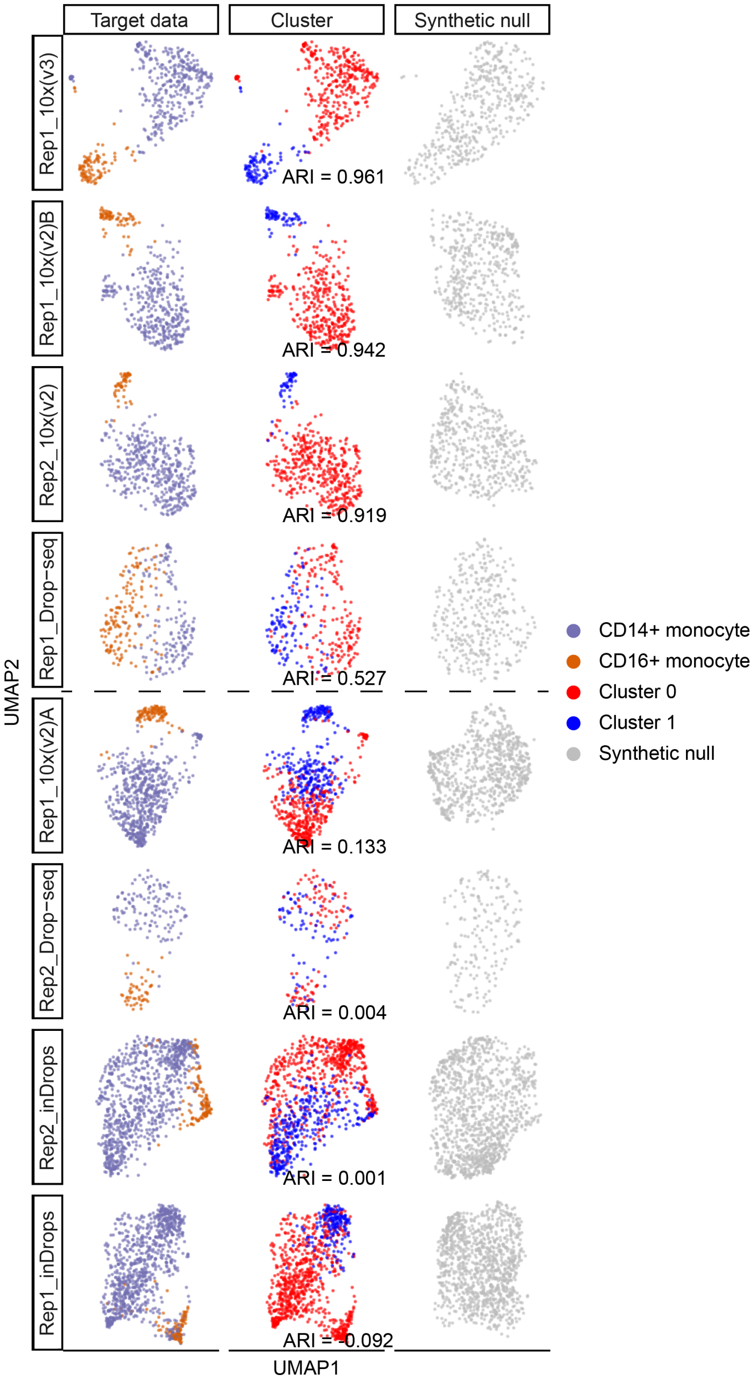


Fig. S12: UMAP visualizations and Seurat clustering accuracies (ARIs) for the eight PBMC monocyte datasets (ordered by ARIs from high to low). The first and second columns show the UMAP visualizations of the eight datasets (the target data), with cells labeled by monocyte subtypes (the first column) or Seurat clusters (the second column). The third column shows the UMAP visualizations of the synthetic null data corresponding to the eight target datasets. The horizontal dashed line separates the eight datasets into two groups (rows 1–4 and rows 5–8) based on clustering accuracy. Monocyte-subtype markers are expected to be more likely identified as post-clustering DE genes in the top four datasets (higher ARIs) compared to the bottom four datasets (lower ARIs).

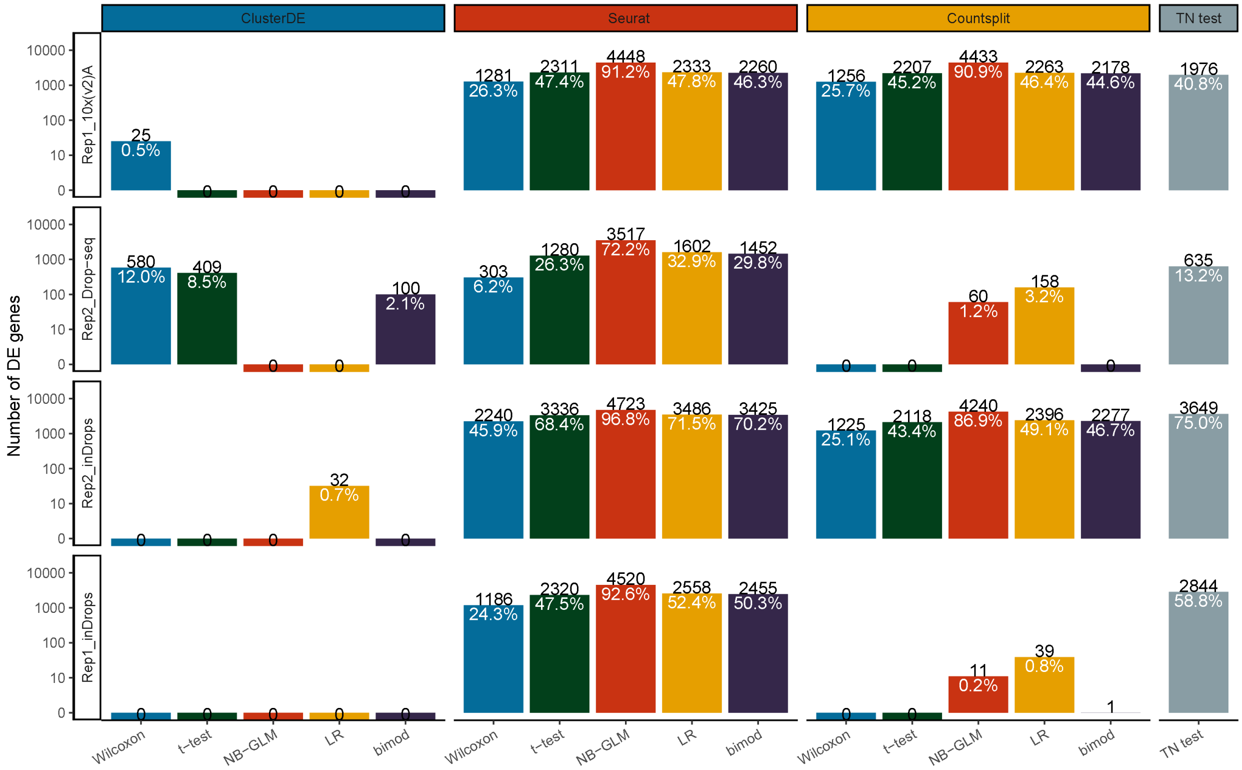


Fig. S13: ClusterDE avoids false discoveries under double dipping. Although the datasets contain two monocyte subtypes, the clustering results poorly align with the subtype labels (the bottom four datasets in Fig. S12), and therefore, few or no DE genes should be identified. The numbers in black indicate the numbers of DE genes identified, while the numbers in white represent the proportions of DE genes among all genes. As expected, in most cases, ClusterDE does not idetify any DE genes.

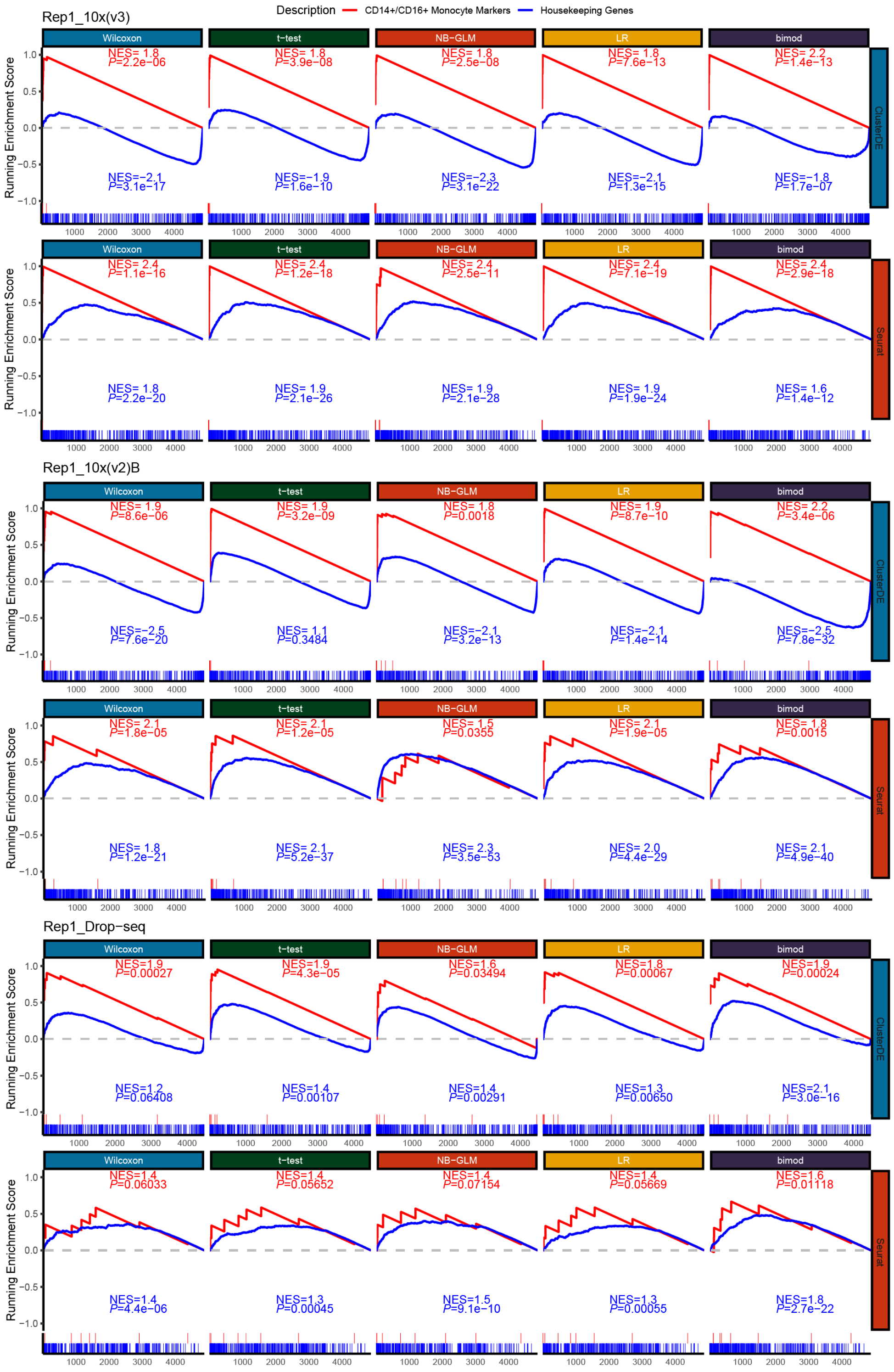


Fig. S14: Gene set enrichment analysis (GSEA) of the ranked DE gene lists identified by ClusterDE and Seurat using five DE tests across three PBMC monocyte datasets. The red lines represent the enrichment of the "CD14$^+$/CD16$^+$ Monocyte Marker Genes" set, while the blue lines represent the enrichment of the "Housekeeping Genes" set. The short vertical lines at the bottom indicate the rank distributions of genes from the two gene sets within each ranked DE gene list. The normalized enrichment score (NES) quantifies the direction and magnitude of enrichment, and the P value reflects the statistical significance of the enrichment.

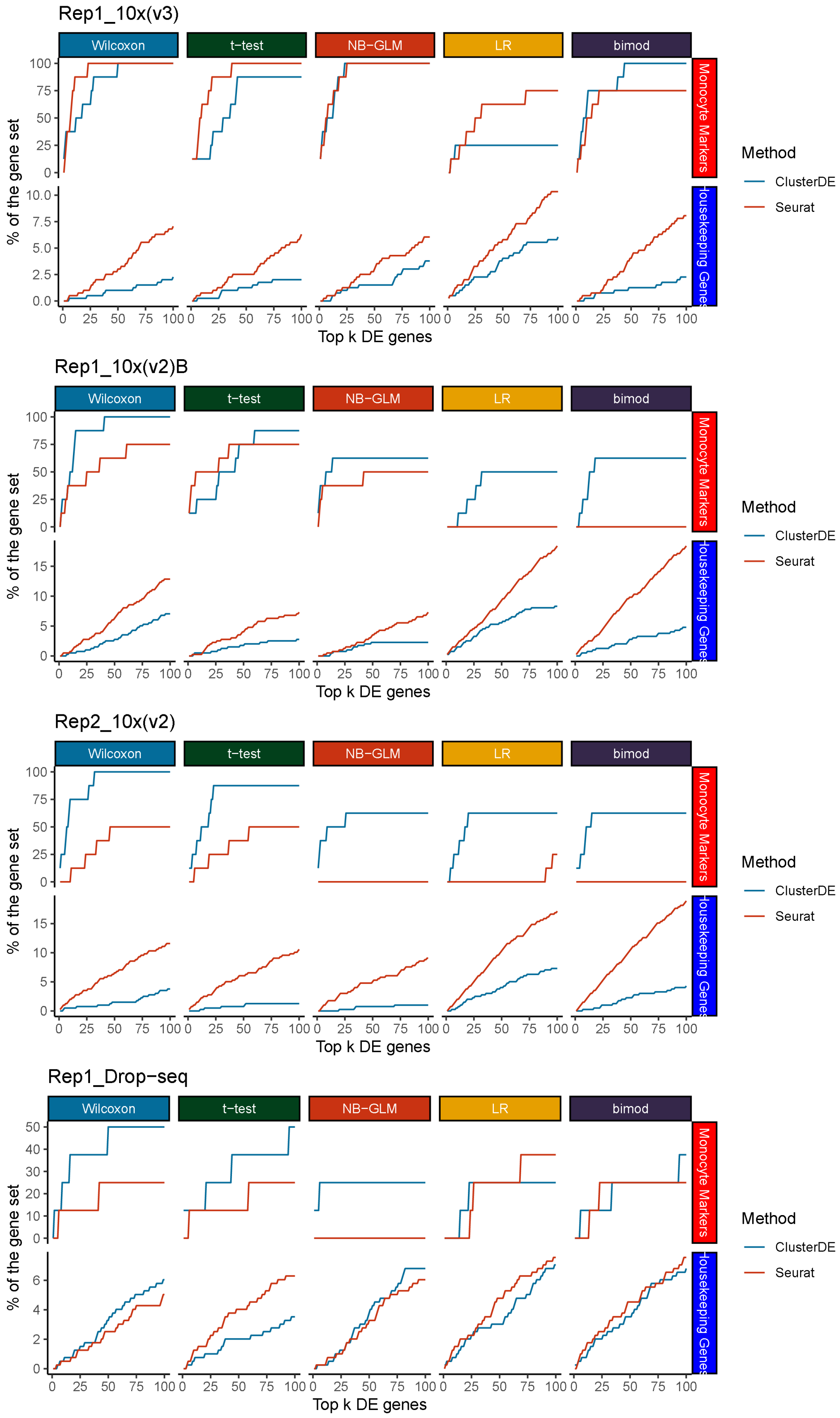


Fig. S15: Overlaps between monocyte markers (or housekeeping genes) and the top k DE genes, with k ranging from 1 to 100. The four panels correspond to four PBMC monocyte datasets. The horizontal axis represents k (e.g., $k = 100$ corresponds to the top 100 DE genes from the DE gene lists), while the vertical axis indicates the proportion of detected monocyte subtype markers (top row in each panel) or housekeeping genes (bottom row in each panel) among the top k DE genes identified by each of the five DE tests (columns). In most cases, ClusterDE (blue line) identifies a higher proportion of monocyte subtype markers and a lower proportion of housekeeping genes compared to Seurat (red line).

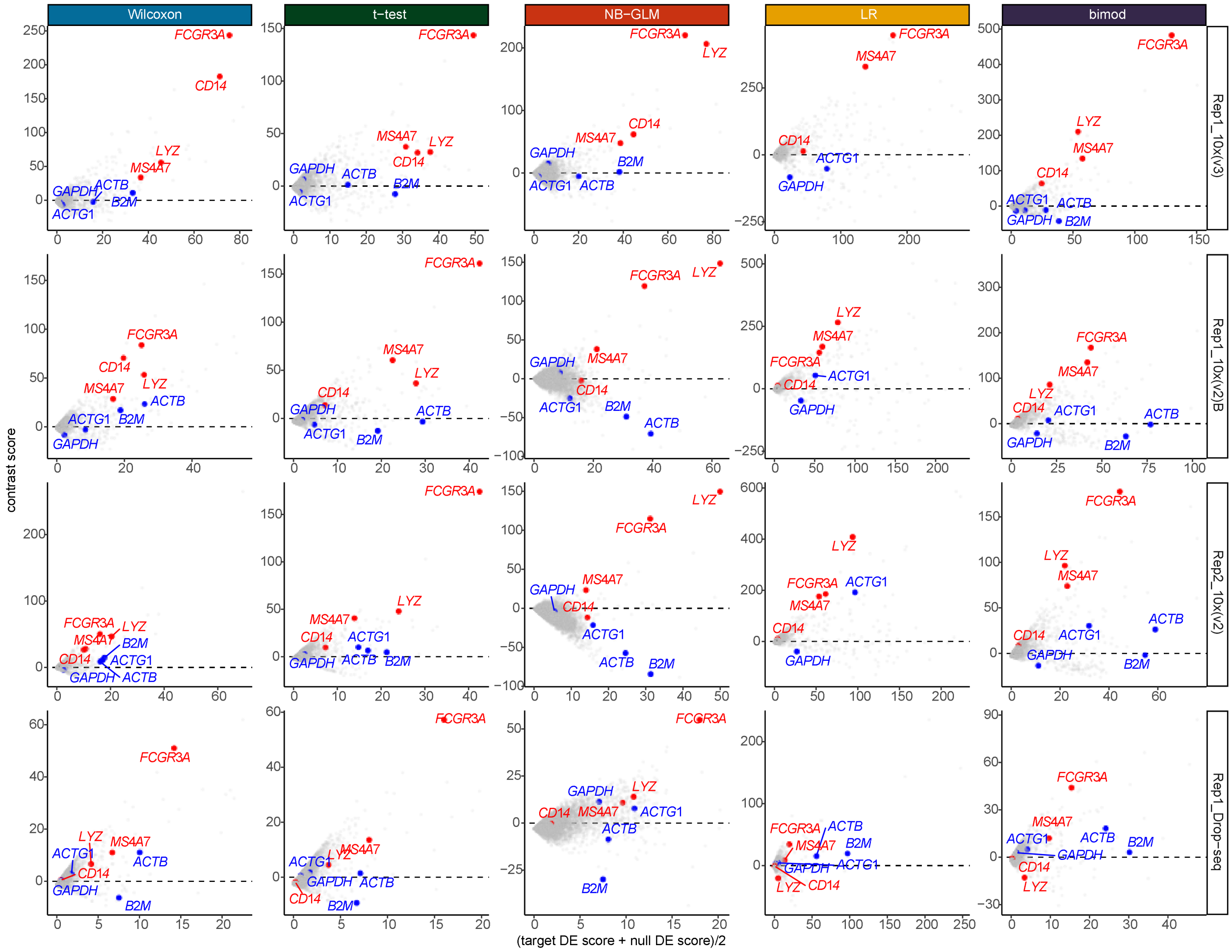


Fig. S16: Minus-average (MA) plots of ClusterDE contrast scores (target DE score minus null DE score) vs. the averages of target DE scores and null DE scores. Red points represent four well-known CD14$^+$/CD16$^+$ subtype markers, and blue points represent four well-known housekeeping genes. The dashed black line indicates contrast scores of 0. For housekeeping genes, their DE scores are large in both the target data and the synthetic null data, resulting in contrast scores centered around 0. As a result, these housekeeping genes would be ranked highly by Seurat (which only considers target DE scores) but not by ClusterDE. On the other hand, monocyte subtype markers have much larger DE scores in the target data than in the synthetic null data, leading to high contrast scores and top rankings by ClusterDE.

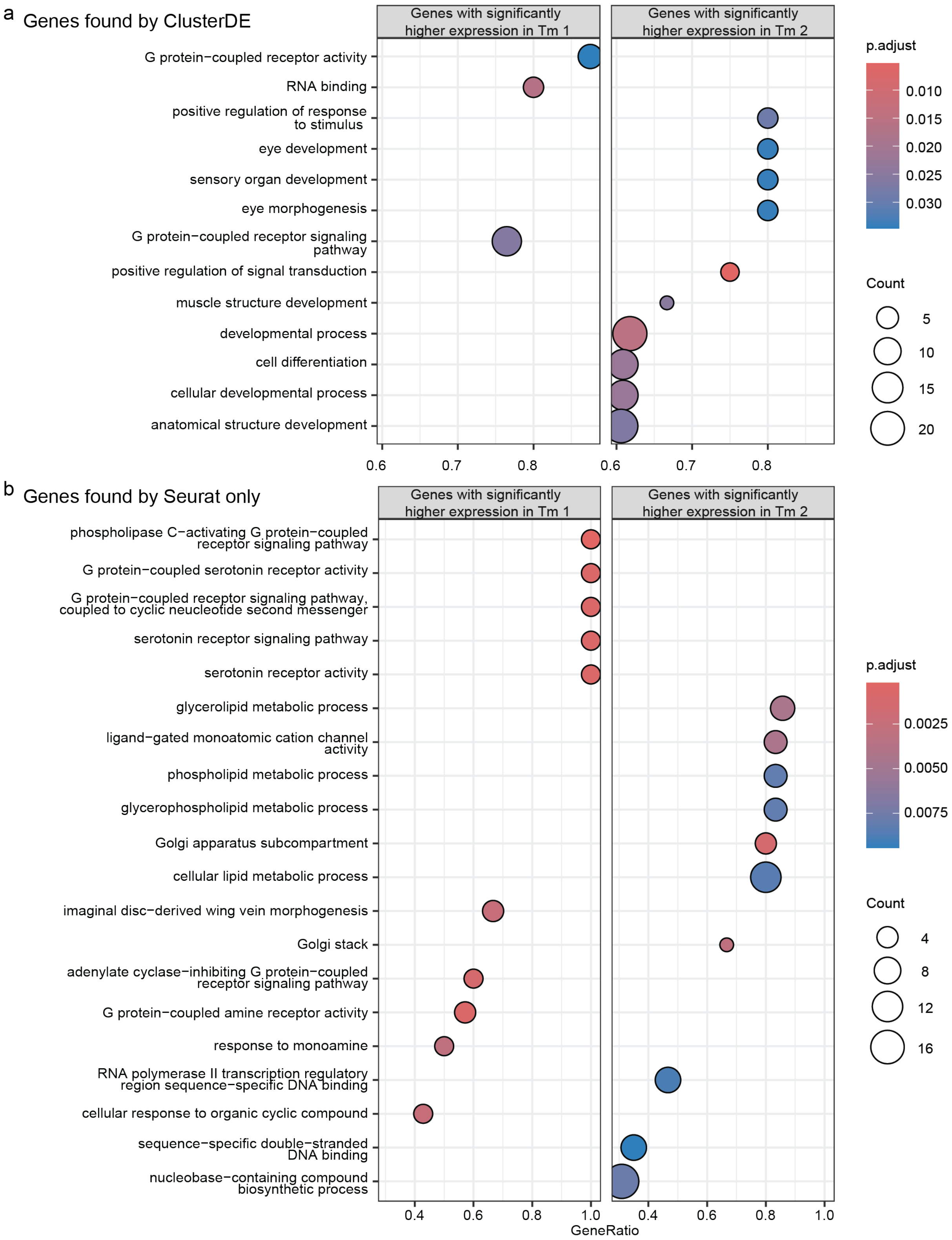


Fig. S17: GSEA analysis for post-clustering DE genes between Tm1 and Tm2 cell types, identified by cell clustering, in the *Drosophila* neuronal visual system scRNA-seq dataset. **a**, Results for DE genes found by ClusterDE. **b**, Results for DE genes found by Seurat using the Wilcoxon rank-sum test (not identified by ClusterDE).

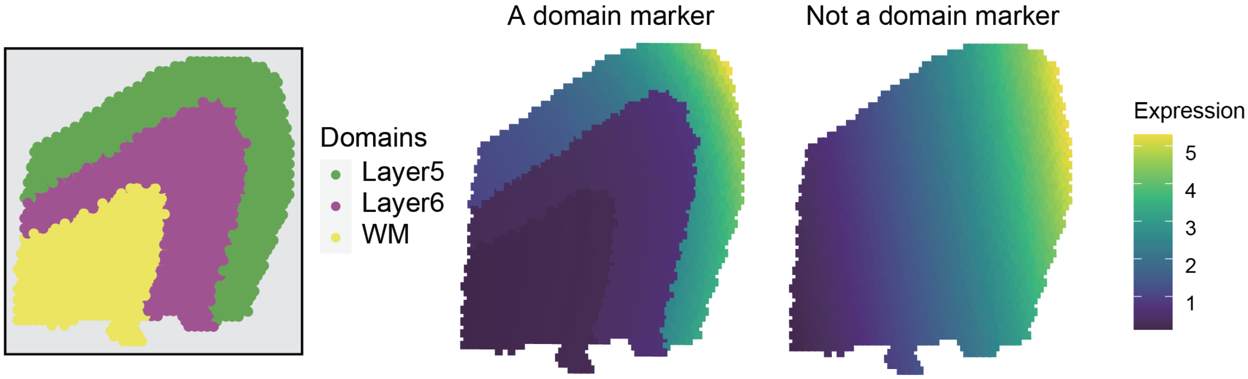


Fig. S18: Illustration of the definition for spatial-domain marker genes under the ClusterDE framework. The left panel depicts three domains. The right panel presents examples: a domain marker gene showing clear discontinuity at the domain boundary and a non-marker gene exhibiting a continuous transition across two domains.

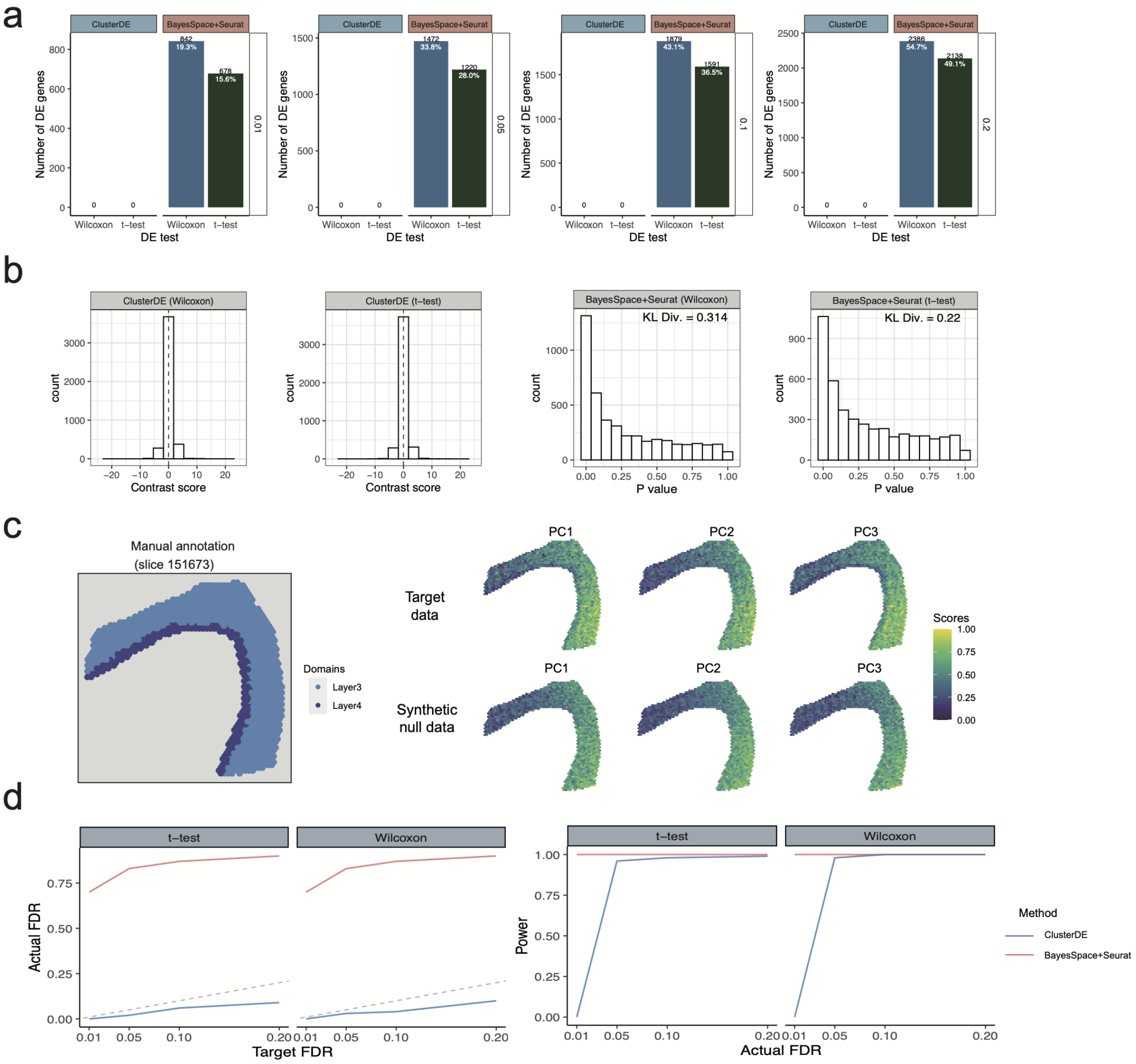


Fig. S19: ClusterDE achieves reliable FDR control and good statistical power in identifying spatial-domain marker genes from simulated SRT datasets containing one or two spatial domains. **a**, Numbers of DE genes identified by ClusterDE and BayesSpace+Seurat (at the target FDRs of 0.01, 0.05, 0.1, and 0.2) in the simulated data containing only one domain. The numbers in black indicate the numbers of DE genes, while the numbers in white represent their proportions among all genes. The statistical tests used in the DE step are the Wilcoxon rank-sum test (Wilcoxon) and t-test. **b**, Validity checks for the contrast scores of ClusterDE and the P values of BayesSpace+Seurat on the simulated one-domain dataset. The first and second columns show that the ClusterDE contrast scores of all genes are approximately symmetric around 0, satifying ClusterDE's assumption for FDR control. The third and fourth columns show histograms of P values obtained from BayesSpace+Seurat. A larger empirical Kullback-Leibler divergence (KL div.) between the P value distribution and the theoretical Uniform[0,1] distribution indicates a more severe violation of the P value uniformity assumption. The results show that P values from BayesSpace+Seurat exhibit severe deviations from the uniform distribution for all genes. **c**, Visualization of manual annotation of real SRT data and the spatial expression patterns of three representative principal components (PC1, PC2, and PC3) among the top five principal components in the target data and synthetic null data, respectively. **d**, The FDRs and power of ClusterDE and BayesSpace+Seurat using t-test and Wilcoxon at various target FDRs.

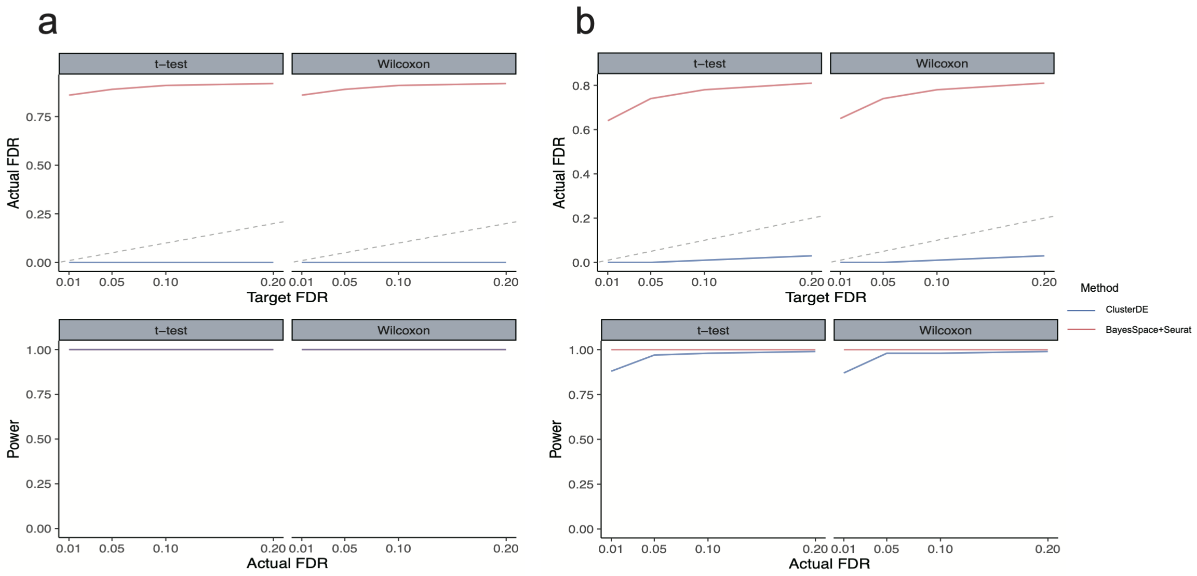


Fig. S20: ClusterDE achieves reliable FDR control and good statistical power in identifying spatial-domain marker genes from a simulated SRT dataset containing three domains (Fig. S18; see “Simulation setting with three spatial domains, including 200 and 168 true spatial-domain marker genes” in the Supplementary Methods). **a**, The FDRs and power of ClusterDE and BayesSpace+Seurat for identifying marker genes between Layer5 and Layer6 at various target FDRs. **b**, The FDRs and power of ClusterDE and BayesSpace+Seurat for identifying marker genes between Layer6 and WM at various target FDRs.

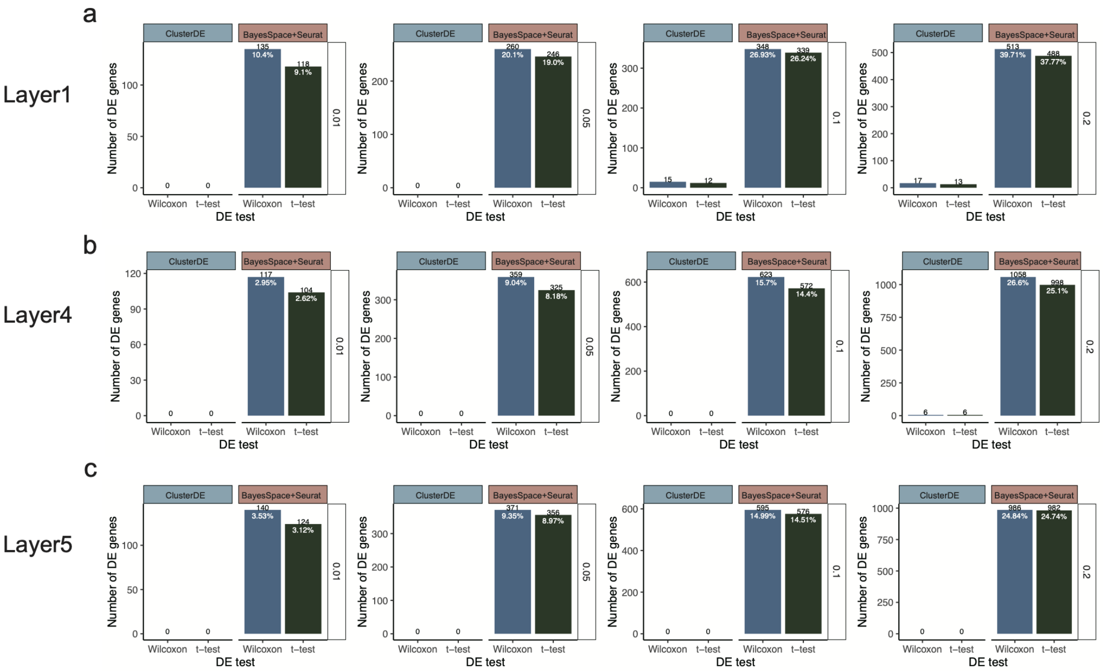
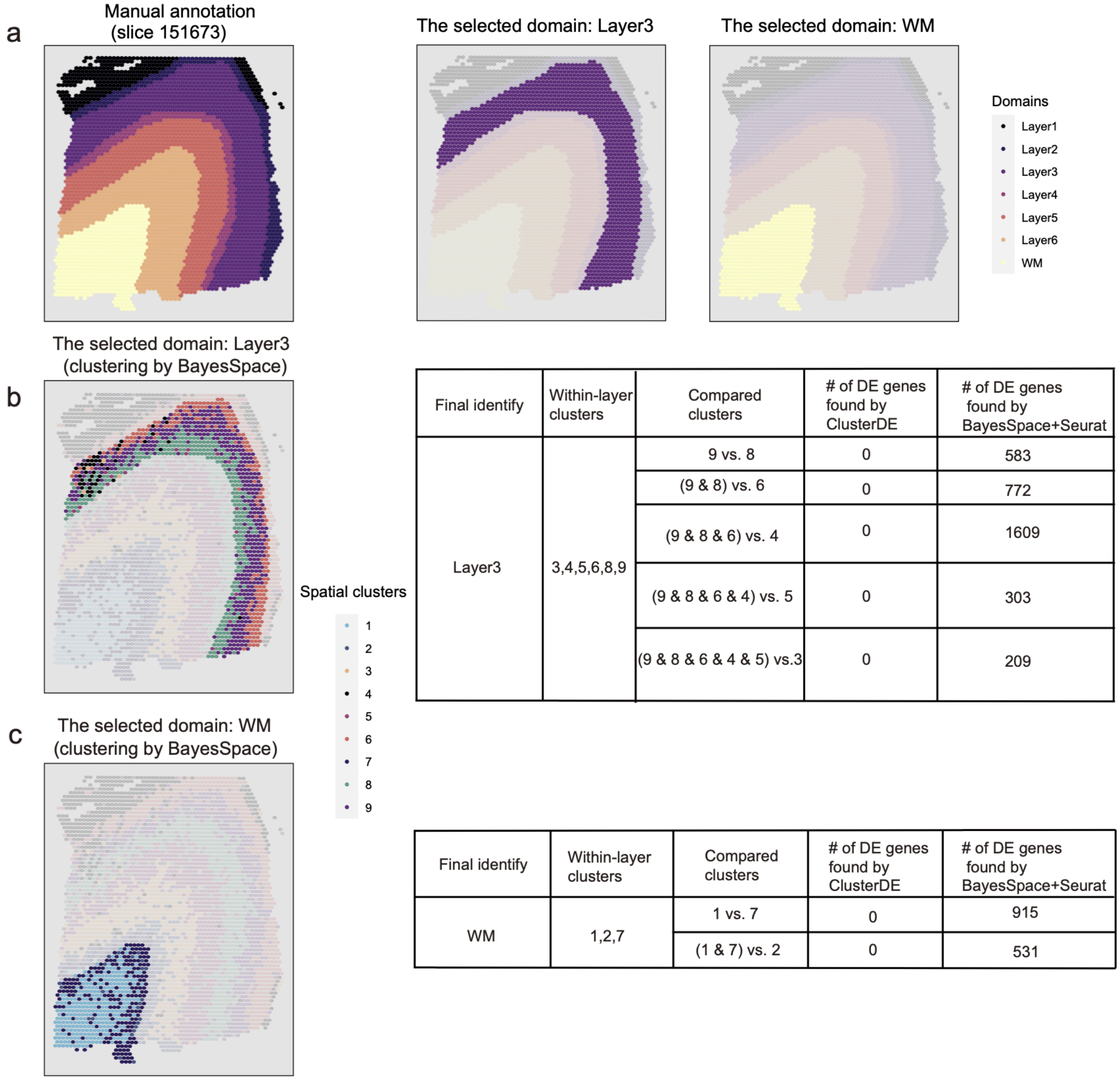


Fig. S21: ClusterDE achieves reliable FDR control in identifying spatial-domain marker genes from three target SRT datasets, each derived from a single domain of the real DLPFC SRT dataset (Fig. S22). The panels (rows) **a**–**c** represent the three target datasets: Layer1, Layer4, and Layer5. The numbers of DE genes identified by ClusterDE and BayesSpace+Seurat at target FDR levels of 0.01, 0.05, 0.1, and 0.2 are shown for each target dataset. Numbers in black indicate the numbers of identified DE genes, while numbers in white represent their proportions among all genes. The statistical tests used for the DE analysis are the Wilcoxon rank-sum test (Wilcoxon) and t-test.



Final identify	Within-layer clusters	Compared clusters	# of DE genes found by ClusterDE	# of DE genes found by BayesSpace+Seurat
Layer3	3,4,5,6,8,9	9 vs. 8	0	583
		(9 & 8) vs. 6	0	772
		(9 & 8 & 6) vs. 4	0	1609
		(9 & 8 & 6 & 4) vs. 5	0	303
		(9 & 8 & 6 & 4 & 5) vs.3	0	209

Final identify	Within-layer clusters	Compared clusters	# of DE genes found by ClusterDE	# of DE genes found by BayesSpace+Seurat
WM	1,2,7	1 vs. 7	0	915
		(1 & 7) vs. 2	0	531

Fig. S22: Application of ClusterDE to the LIBD human dorsolateral prefrontal cortex (DLPFC) SRT dataset. **a**, Visualization of the entire dataset with manual annotation and selected domains: Layer3 (purple) and WM (yellow). **b**, Left panel: Spatial clustering results for Layer3 using BayesSpace. Right panel: Numbers of DE genes identified at the target FDR of 0.05 by ClusterDE and BayesSpace+Seurat (using the Wilcoxon rank-sum test). **c**, Left panel: Spatial clustering results for WM using BayesSpace. Right panel: Numbers of DE genes identified at the target FDR of 0.05 by ClusterDE and BayesSpace+Seurat (using the Wilcoxon rank-sum test).

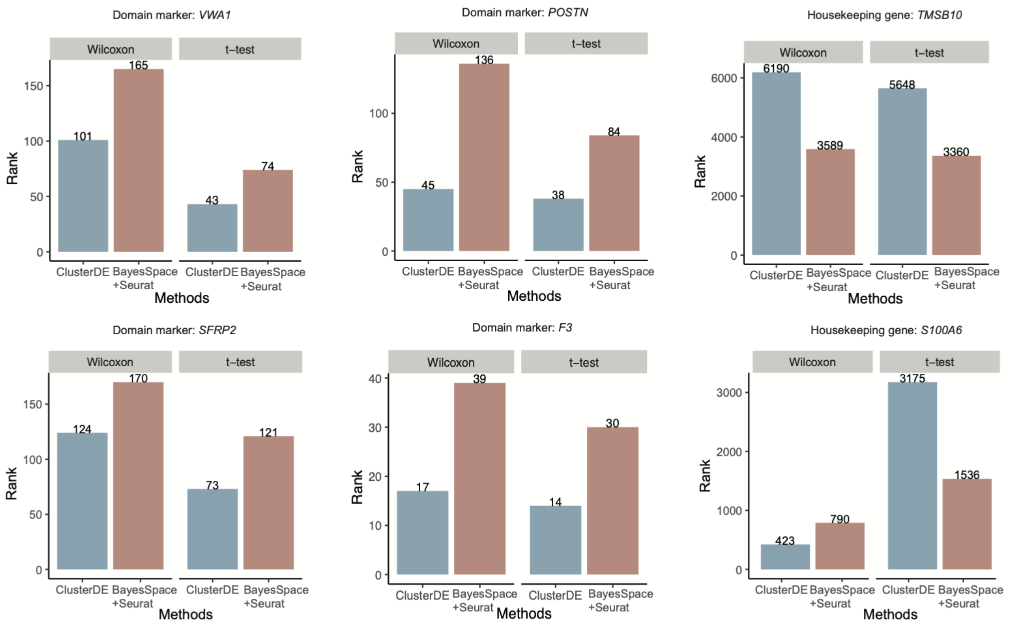


Fig. S23: Application of ClusterDE to the PDAC-B cancer slice from the human primary pancreatic cancer (PDAC) SRT data. Bar plots show the rankings of domain markers and housekeeping genes identified by ClusterDE and BayesSpace+Seurat. The top row shows the rankings of three exemplary genes between the Duct epithelium and Interstitium: the domain markers *VWA1* and *POSTN* and the housekeeping gene *TMSB10*. The bottom row shows the rankings of three exemplary genes between the Interstitium and Cancer region: the domain markers *SFRP2* and *F3* and the housekeeping gene *S100A6*.

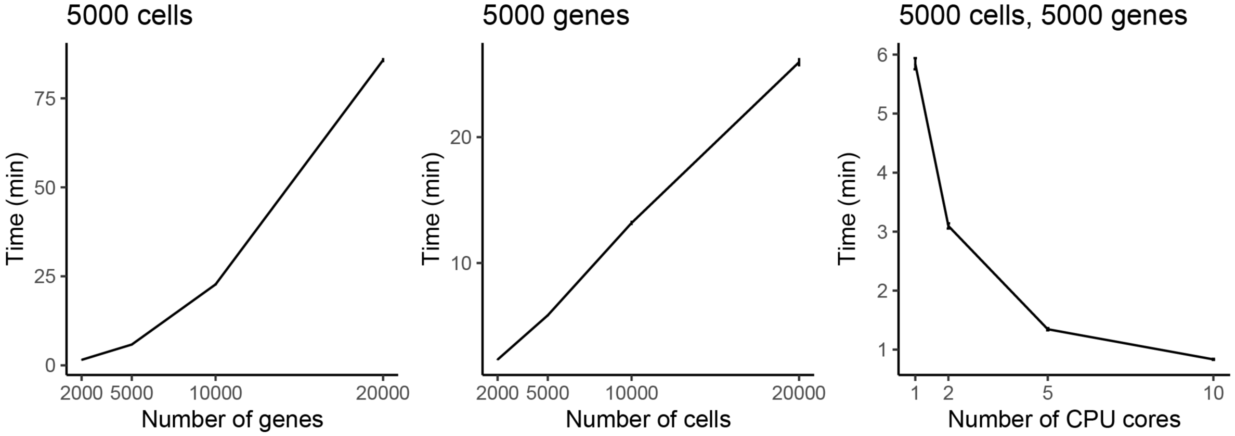


Fig. S24: Time complexity of synthetic null data generation. The computational time scales approximately quadratically with the number of genes, linearly with the number of cells, and is inversely proportional to the number of CPU cores under parallel computing.

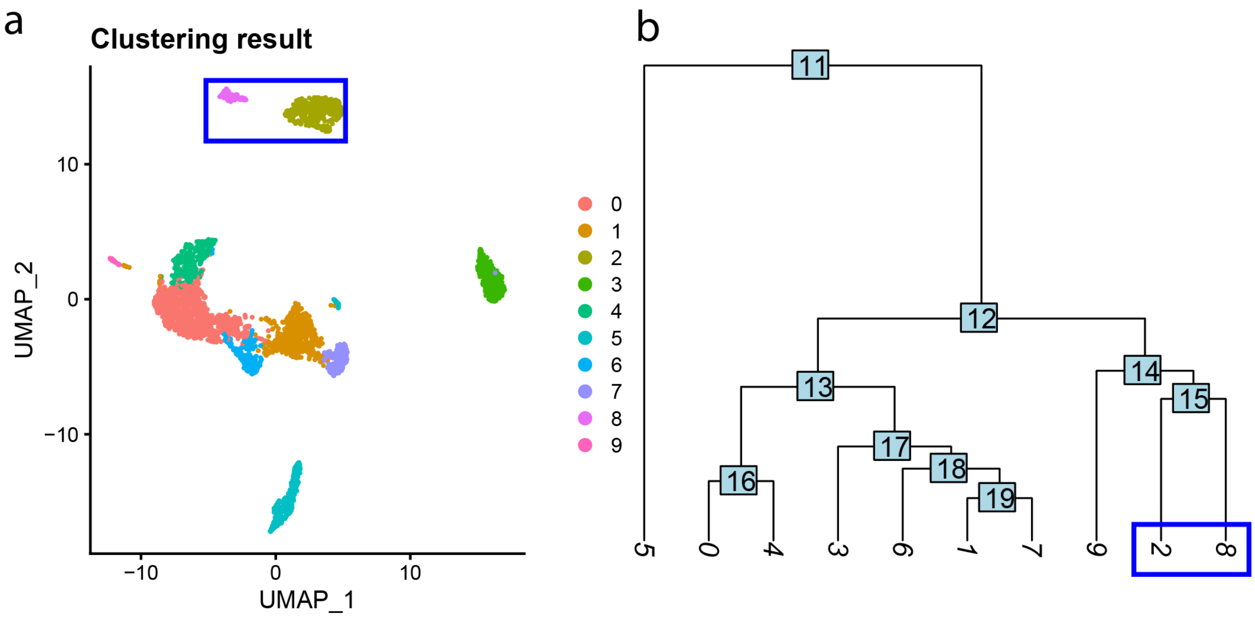


Fig. S25: Demonstration of ClusterDE's application in the presence of multiple cell clusters. **a**, UMAP visualization of Seurat clusters found in the PBMC dataset Rep1 10x(v3). The blue box highlights two neighboring clusters that roughly correspond to CD14^{+}/CD16^{+} monocytes. **b**, Cluster tree constructed by Seurat, with clusters 2 and 8 corresponding to the two clusters highlighted in the blue box in **a**. ClusterDE is recommended for annotating two neighboring cell clusters in UMAP or a cluster tree, as it identifies more reliable marker genes. This is particularly important because neighboring clusters are more likely to be spurious.

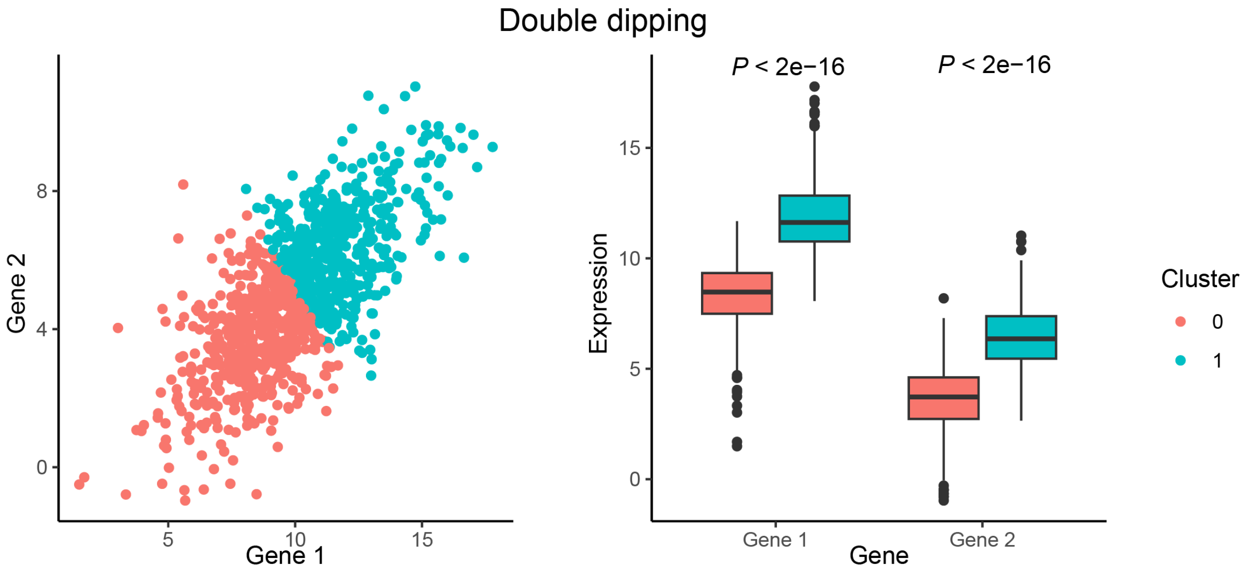


Fig. S26: A toy example illustrating the double-dipping issue. The expression levels of two genes follow a bivariate Gaussian distribution, representing a homogeneous cell type. However, when K-means clustering is applied to divide the cells into two clusters, the two genes are artificially forced to exhibit different distributions between the clusters. As a result, both genes are incorrectly identified as post-clustering DE genes, leading to the false conclusion that the two clusters represent distinct cell types defined by these genes.

References

[1] Baolin Liu et al. “An entropy-based metric for assessing the purity of single cell populations”. In: *Nature communications* 11.1 (2020), p. 3155.

[2] Jiyuan Fang et al. “Clustering Deviation Index (CDI): a robust and accurate internal measure for evaluating scRNA-seq data clustering”. In: *Genome Biology* 23.1 (2022), p. 269.

[3] Wei Vivian Li. “Phitest for analyzing the homogeneity of single-cell populations”. In: *Bioinformatics* 38.9 (2022), pp. 2639–2641.

[4] Maria Mircea et al. “Phiclust: a clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations”. In: *Genome Biology* 23.1 (2022), pp. 1–24.

[5] I.N. Grabski, K. Street, and R.A. Irizarry. “Significance analysis for clustering with single-cell RNA-sequencing data”. In: *Nat Methods* 1.1 (2023), p. 1.

[6] Valentine Svensson. “Droplet scRNA-seq is not zero-inflated”. In: *Nature Biotechnology* 38.2 (2020), pp. 147–150.

[7] Tae Hyun Kim, Xiang Zhou, and Mengjie Chen. “Demystifying “drop-outs” in single-cell UMI data”. In: *Genome biology* 21.1 (2020), p. 196.

[8] Ruochen Jiang et al. “Statistics or biology: the zero-inflation controversy about scRNA-seq data”. In: *Genome biology* 23.1 (2022), pp. 1–24.

[9] Dongyuan Song et al. “scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics”. In: *Nature Biotechnology* (2023), pp. 1–6.

[10] Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*. Vol. 3. Springer, 1986.

[11] Tianyi Sun et al. “scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured”. In: *Genome biology* 22.1 (2021), p. 163.

[12] Andrew McDavid et al. “Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments”. In: *Bioinformatics* 29.4 (2013), pp. 461–467.

[13] Xinzhou Ge et al. “Clipper: p-value-free FDR control on high-throughput data from two conditions”. In: *Genome biology* 22.1 (2021), pp. 1–29.

[14] Rina Foygel Barber and Emmanuel J Candès. “Controlling the false discovery rate via knockoffs”. In: *Annals of Statistics* (2015).

[15] Jesse M Zhang, Govinda M Kamath, and N Tse David. “Valid post-clustering differential analysis for single-cell RNA-Seq”. In: *Cell systems* 9.4 (2019), pp. 383–392.

[16] Anna Neufeld et al. “Inference after latent variable estimation for single-cell RNA sequencing data”. In: *Biostatistics* 25.1 (2024), pp. 270–287.

[17] Emmanuel Candes et al. “Panning for gold:‘model-X’knockoffs for high dimensional controlled variable selection”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.3 (2018), pp. 551–577.

[18] Angelo Duò, Mark D Robinson, and Charlotte Soneson. “A systematic performance evaluation of clustering methods for single-cell RNA-seq data”. In: *F1000Research* 7 (2018).

[19] Luyi Tian et al. “Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments”. In: *Nature methods* 16.6 (2019), pp. 479–487.

[20] Grace XY Zheng et al. “Massively parallel digital transcriptional profiling of single cells”. In: *Nature communications* 8.1 (2017), p. 14049.

[21] Jiarui Ding et al. “Systematic comparison of single-cell and single-nucleus RNA-sequencing methods”. In: *Nature biotechnology* 38.6 (2020), pp. 737–746.

[22] Mehmet Neset Özel et al. “Neuronal diversity and convergence in a visual system developmental atlas”. In: *Nature* (2021), pp. 88–95.

[23] Kristen R Maynard et al. “Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex”. In: *Nature neuroscience* 24.3 (2021), pp. 425–436.

[24] Reuben Moncada et al. “Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas”. In: *Nature biotechnology* 38.3 (2020), pp. 333–342.

[25] A Sina Booeshaghi et al. “Depth normalization for single-cell genomics count data”. In: *bioRxiv* (2022), pp. 2022–05.

[26] Li-Li Hsiao et al. “A compendium of gene expression in normal human tissues”. In: *Physiological genomics* 7.2 (2001), pp. 97–104.

[27] Chenguang Dai et al. “False discovery rate control via data splitting”. In: *Journal of the American Statistical Association* (2022), pp. 1–18.