Detecting regions of differential abundance between scRNA-Seq datasets

Jun Zhao [1,2,3], Ariel Jaffe[4], Xiuyuan Cheng[5], Richard Flavell[3,6], Yuval Kluger[1,2,4,†]

Abstract

Recently, single cell RNA sequencing (scRNA-Seq) has been applied to compare transcriptomic landscapes between biological conditions or states. This comparison allows the discovery of cell subpopulations that are differentially abundant between biological conditions or states. Typically, these subpopulations are determined by identification of clusters or cell types in which the proportion of cells from the two states deviates significantly from the corresponding proportion in the entire cell populations. This approach might be suboptimal as the most differentially abundant subpopulations may not perfectly overlap with these clusters or cell types. Here, we develop DA-seq, a multiscale algorithm that detects cell subpopulations with the highest disproportionate cell abundances between two samples. DA-seq does not require prior partitioning of the data based on clustering or cell subtypes. To evaluate DA-seq, we applied it to one simulated and two real scRNA-Seq datasets. The first scRNA-Seq dataset was obtained from a study on melanoma responders and non-responders to checkpoint immunotherapy. The second datasets is from a study that compares two time points of developing embryonic mouse skin.

1 Introduction

scRNA-Seq technology became an invaluable tool for profiling biological systems through its ability to measure the expression level of all genes over thousands to millions of cells [1, 2]. Recently, the task of comparing single cell datasets across different conditions became prevalent in biological and medical research [3, 4]. Here, we provide an approach for addressing analytic challenges involved in this task.

A standard approach for characterizing transcriptional differences between scRNA-Seq datasets generated from nonidentical biological conditions or states is through clustering. Namely, the transcriptional profiles of all cells included in the union of these datasets are partitioned into clusters of cells based on unsupervised clustering or known predefined markers. Next, the proportions of cells of different biological conditions or

†To whom correspondence should be addressed, yuval.kluger@yale.edu
[1]Interdepartmental Program in Computational Biology and Bioinformatics, [2]Department of Pathology, [3]Department of Immunobiology, [4]Program in Applied Mathematics, Yale University, New Haven, CT 06511. [5]Department of Mathematics, Duke University, Durham, NC 27708. [6] Howard Hughes Medical Institute, Chevy Chase, MD, USA

states are compared for each cluster. A cluster in which these proportions are significantly different is considered differentially abundant (DA). Subsequently, a subset of differentially expressed genes characterizing each DA cluster is determined. Such approaches were applied, for example, to compare responders and non-responders to cancer treatment [5], to quantify changes in PBMC cells under rest and interferon β stimulation [3], and to compare colon single cell data before and after treatment [6]. ClusterMap is a variant of this clustering-based approach [7]. In ClusterMap the clustering step is applied separately to samples from two conditions or states. Thereafter, the datasets are merged by matching between similar clusters which allows comparisons between the samples at the cluster level.

Clustering based approaches might be suboptimal in cases where the cell subpopulations associated with differential abundance between conditions are distributed among several adjacent clusters, or alternatively, encompass only a part of a cluster. Another scenario where clustering approaches are not suitable is for datasets in which heterogeneous cell populations do not fall into well defined separate clusters. Cydar is an alternative approach that does not require initial clustering and was developed for finding differential abundances between samples analyzed in mass cytometry experiments [8]. The algorithm performs multiple differential abundance tests between the samples within different spheres centered at randomly selected positions in the marker space. Its performance is dependent on selection of the positions of these spheres, their radius size and their number. Thus, new methods that consider differential abundance around each cell (point) using multiscale approach may overcome these limitations.

Here, we develop DA-seq, an approach for detecting differentially abundant cell subpopulations. Our approach is based on the following steps. First, we define a multi-scale *score vector* that measures, for each cell, the proportion of cells from both conditions among its k nearest neighbors, for several values of k. Second, we apply a top-down clustering scheme to detect regions that strongly differentiate between two conditions. In contrast to the standard clustering-based procedure discussed above, our clustering step groups cells that are similar in both their expression levels and DA profile. Finally, we apply a standard gene detection method implemented in *Seurat* [3] to capture genes that are dominant in the DA regions, and thus relevant to the differences between conditions. We demonstrate the advantage of our approach both on simulated and real -world datasets published in [5] and [9]. For these publicly available datasets, we compare the results obtained with our approach to the findings in the original papers.

1.1 Problem setup

In our work, we analyze and compare the gene expression profiles of scRNA-Seq samples in two biological conditions or states. Let $X = \{\boldsymbol{x}_i\}_{i=1}^{n_x}$and $Y = \{\boldsymbol{y}_i\}_{i=1}^{n_y}$ be two sets of m dimensional vectors, providing the expression levels of m genes for n_x and n_y cells from two biological conditions or states, respectively. Typically, the number of genes is ~ 30000, while the number of cells can vary. Our goal is to detect local regions in the gene space that differentiate between X and Y. We assume that the vectors x_i, y_i are independent realizations of two m dimensional density functions, denoted $f_x(x), f_y(y)$,

$$f_x(x), f_y(y) : \mathbb{R}^m \to \mathbb{R}.$$

In typical studies that compare two states or conditions the corresponding density functions are expected to be similar in most positions in the gene space except for certain locations where they deviate from each other. Let Ω denote these locations in the m-dimensional space. Formally, for any vector $v \in \mathbb{R}^m$ and some

threshold Δ,

$$\begin{cases} |f_x(\boldsymbol{v}) - f_y(\boldsymbol{v})| < \Delta & \boldsymbol{v} \notin \Omega \\ |f_x(\boldsymbol{v}) - f_y(\boldsymbol{v})| \geq \Delta & \boldsymbol{v} \in \Omega. \end{cases} \tag{1}$$

Our first task is to detect contiguous regions where the differences between $f_x(\boldsymbol{v})$ and $f_y(\boldsymbol{v})$ are significant. From this point on, we refer to the regions in Ω as differentially abundant regions or in brief DA regions. Once a DA region is identified, our secondary task is to detect genes that characterize its cells.

The first task can be thought of as a *local* version of the two sample problem. In theory, given a sufficiently large dataset, a straightforward approach would be to estimate the distance between the density functions $f_x(X)$ and $f_y(Y)$ for multiple locations in the gene space. In [8], for example, this approach was suggested for mass cytometry data. There are two main problems, however, using this approach for scRNA datasets: (i) In single cell datasets we usually measure the gene expression of thousands of genes. Thus, the curse of dimensionality makes it almost impossible to accurately estimate the distance at specific points via standard means such as histograms or kernel density estimates, see [10, 11]. (ii) Our goal is to detect *all the regions* that differentiate between X and Y. To do so one must test sufficiently many locations to cover the distributions' support in the gene space. This fact, however, exhibits a tradeoff. On one hand, we would like to test many locations, in order to have better coverage of the distributions' support in the gene space. On the other hand, by increasing the number of tests we reduce the overall testing power, due to the multiple testing problem. In the following sections, we describe how DA-seq addresses these concerns.

2 Methods

One of the common models used in high dimensional settings is *the manifold assumption*, where the data are assumed to reside in one or several unknown low-dimensional manifolds, see [12]. The underlying geometry of the manifold can be captured with a nearest neighbor graph.

Our first step is to compute the k nearest neighbors for each cell in the combined dataset $X \cup Y$. We expect that the k nearest neighbors of a cell positioned in a non DA region include cells from samples X and Y such that their proportion in this neighborhood is similar to the proportion between the total number of cells in X and the total number of cells in Y. In contrast, for cells in a DA region where, say, $f_x > f_y$, we expect to see a large percentage of neighbors that belong to X. Let $x_i(k), y_i(k)$ denote the number of k nearest neighbor cells of cell i, such that $x_i(k) + y_i(k) = k$, where $x_i(k) \in X, y_i(k) \in Y$. Recall that n_x, n_y denote the total number of cells in X and Y, respectively. We define the following score to facilitate the detection of DA regions,

$$s_i(k) = \frac{\frac{x_i(k)}{y_i(k)} - \frac{n_x}{n_y}}{\frac{x_i(k)}{y_i(k)} + \frac{n_x}{n_y}}. \tag{2}$$

The larger the score the larger the differential abundance between the densities f_x and f_y in a neighborhood of size k around cell i. The numerator in Eq. (2) represents the difference between the local abundance ratio $x_i(k)/y_i(k)$ around cell i and the global abundance ratio n_x/n_y. The denominator normalizes the score such

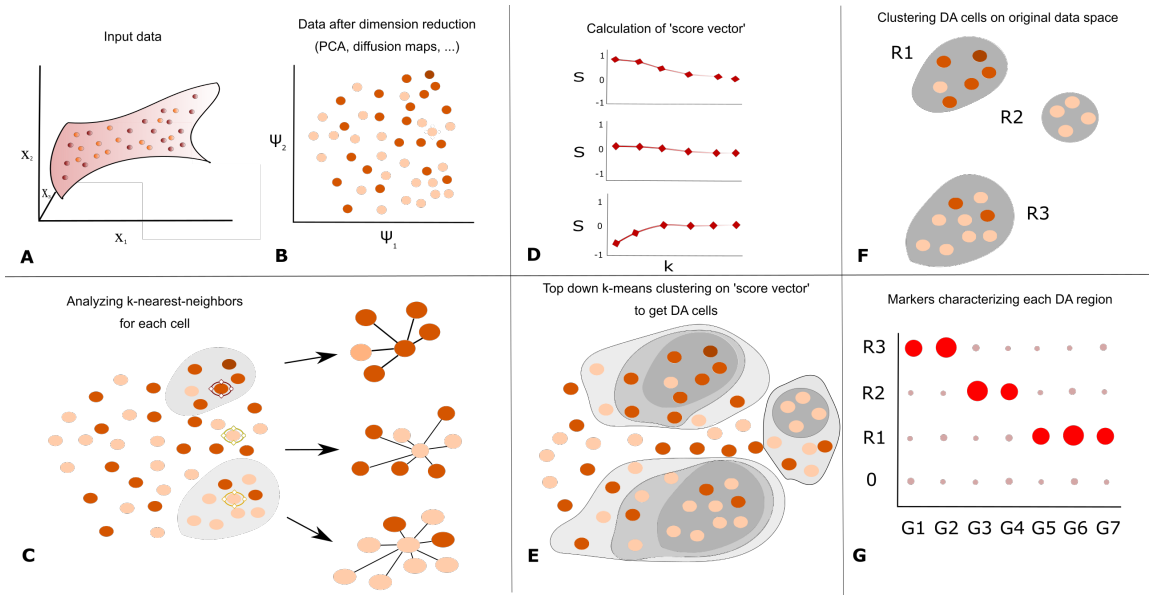

Figure 1: **Schematic demonstration of DA-seq.** (A) The input: Expression level of cells from two biological conditions, type 1 and 2. The high dimensional data is assumed to reside in a low dimensional manifold. (B) (Optional step) The cells representation in diffusion coordinates ψ_1, ψ_2. (C-D) For each cell, we compute a *score vector* according to Eq. (3) for several values of k. (E) A top down clustering of the cells. The first iterations of the clustering scheme are based on the score vector. At each iteration, we keep only the two clusters with high differential abundance of cells of type 1 and 2, while removing the cells in the third cluster where the differential abundance is more moderate. The area around each cell is colored by the iteration number it was removed. Thus, areas colored by dark grey have high differential abundance. (F) The last iteration of the top down clustering is based on the transcriptomic profile. Thus, the cells in the dark grey areas are separated into 3 regions. R2 and R3 have a large proportion of cells of type 1, while R1 has a large proportion of cells of type 2. **G** Genes that characterize the differential regions are detected.

that $-1 < s_i(k) < 1$. We modify the expression of Eq. (2) to

$$s_i(k) = \frac{\frac{x_i(k)}{n_x} - \frac{y_i(k)}{n_y}}{\frac{x_i(k)}{n_x} + \frac{y_i(k)}{n_y}}. \tag{3}$$

This expression is equivalent to the right hand side of Eq. (2) when $y_i(k) > 0$ but it allows us also to define a score when $y_i(k) = 0$. An important parameter in the score measure is the choice of an optimal k. Similarly to the problem of choosing a bandwidth for density estimation, the choice of k exhibits a bias-variance trade off. If k is too large, the score will become non-informative, since a large portion of the neighbors will reside outside of the DA region. However, if k is too small, $s_i(k)$ will have high variance, which will result in detection of a large number of irrelevant points. To cope with this problem, we compute the score for a range of values of k. Thus, instead of a single element, each cell is assigned a score vector $\boldsymbol{s}_i$,

$$\boldsymbol{s}_i = \{s_i(k_1), \ldots, s_i(k_l)\}.$$

Fig. 1D illustrates the qualitative behaviour of the score vector for three different cells. The upper vector corresponds to a cell in a DA region where $f_x > f_y$. Thus, for small values of k, the score is high. As k increases, the score typically decreases as the region at this scale may contain more balanced proportion of cells from samples X and Y and even include neighbors positioned outside of the DA region. The entries of this monotonically decreasing vectors are positive. Similarly, a typical score vector in a DA region where $f_y > f_x$ has monotonically increasing trend and negative entries. In contrast, the entries of a score vector for a cell in a neutral region will have small positive or negative values around 0 for all k.

Once the score vector is computed for all cells, our goal is to detect DA regions by taking into account both the gene expression profile and score vector of each cell. This is different from the standard clustering-based differential abundance approaches that utilize only the pairwise distances between the gene expression profiles of all cells from samples X and Y. To do so, we first perform an iterative top down k means clustering to cluster the data into three clusters, corresponding to $f_x(X) > f_y(Y), f_y(Y) > f_x(X)$ and $f_x(X) \approx f_y(Y)$. In order to increase the resolution of our analysis and obtain smaller regions with stronger differential behaviour, we repeat the clustering step for the cells in regions where $f_x(X) > f_y(Y)$ and $f_y(Y) > f_x(X)$, while removing the cells for the neutral region. Fig. 1E illustrates the result of the repeated clustering steps, where the regions obtained after each iteration are colored with darker shades of grey. After each iteration, we retain a smaller number of cells that reside in nested regions with higher Δ (see Eq. (1)). Thus, the number of iterations is implicitly related to Δ.

Finally, to obtain distinct DA regions from the remaining cells, we perform an additional divisive clustering step based on the transcriptomic data. This is done by applying k-means clustering with outlier removal [13] to the low dimensional embedding of the remaining cells. Algorithm 1 summarizes the different steps included in the computation of the DA regions, and specifies the parameters required as input to apply DA-seq.

Remark 1. *For certain scRNA-Seq samples, data preprocessing by nonlinear dimensionality reduction methods may boost the performance of DA-seq. One such approach is diffusion maps, designed to capture complex geometric structures in the data, see [14]. This approach typically requires a larger embedding size than linear methods such as PCA or CCA. In our implementation of DA-seq we left this step as optional. In appendix B we give a short description of diffusion maps.*

Algorithm 1: DA-seq: detecting differential abundance (DA) regions in scRNA-Seq

Input: Gene expression matrix of samples from two conditions, X and Y
Output: A list of DA regions between X and Y

Step 0. Dimensional reduction of the feature (gene) space
Parameters: number of dimensions
Calculate low dimensional embedding (principal component analysis (PCA) or canonical correlation analysis (CCA), etc.) from merged gene expression data of X and Y.
(Optional) calculate diffusion coordinates from lower dimensional data.

Step 1. Calculate score vector s_i
Parameters: a range of k values
for *each cell i* **do**
  compute its score vector s_i for a given range of k values via Eq. (3)
end

Step 2. Top-down clustering on score vector
Parameters: maximum iteration i_{max}, maximum ratio of remaining cells r_{max}
while $i < i_{max}$ & $r_i < r_{max}$ **do**
  k-means clustering with $k = 3$ on score vectors of remaining cells.
  Remove cells in the neutral cluster.
end

Step 3. Apply expression-based k-means clustering with outlier removal to cells remained in **Step 2**
Parameters: number of clusters k, outlier ratio α
k-means clustering with $tclust$ [13] for remaining cells based on their transcriptomic profiles, after dimensionality reduction (**Step 0**).

The final step involves clustering of the expression profiles of the remaining cells, which allows us to characterize different DA regions in the gene space. Our second task is to use these clusters in order to detect genes that are characteristic to each DA region. For every DA region, we define a gene as characteristic if it is expressed (or not expressed) in a significantly larger fraction of cells within the DA region compared to the fraction of cells it is expressed out of this region. To detect these characteristic genes we used the Wilcoxon rank sum test implemented in the single cell processing package *Seurat* [15].

3 Results

In this section, we apply our method to single cell datasets from two papers. The first is [5], where the aim is to identify factors related to success or failure of immune checkpoint therapy. The second is [9], where cells from dorsolateral skin were sequenced at different time points of mouse embryo development. We compare the results obtained by our method to the results obtained in the original analyses done in [5] and [9]. Our method is designed to zoom into regions of significant differential abundance. These regions

partially overlap with DA clusters that were found by standard clustering-based DA analysis and hence allow us to capture all the results obtained by these methods. Importantly, our approach also allows to uncover markers characteristic to the DA-region that might not be detected by standard approaches. In appendix C we present additional results based on simulated data.

3.1 Distinct T cell subsets from responsive and non-responsive melanoma patients

In [5], the aim is to identify factors related to success or failure of immune checkpoint therapy. To that end, 16,291 immune cells from 48 samples of melanoma patients treated with checkpoint inhibitors were profiled and analyzed. The tumor samples were classified as responders or non-responders based on radiologic assessments. The cells originating from responding tumors and non-responding tumors were labeled using two different colors in the t-SNE plot of Fig. 2A. Comparisons between responders and non-responders yielded important biological insights.

Analysis done in [5]. The first step in [5] was to cluster the 16,291 immune cells into 11 distinct clusters (Fig. 2B). Subsequently, the responders and non responders were compared according to the relative abundance of cells in each of the predefined clusters. Two clusters ($G1, G10$) were enriched in cells from the responder samples, and four clusters ($G3, G4, G6, G11$) were enriched in cells from non-responder samples. Finally, the authors composed a list of genes that are highly expressed in these six differentially abundant clusters.

Comparison between results obtained by DA-seq and standard DA analysis. Applying our methods, we obtained five DA regions, denoted $R1-R5$, see Fig. 2D. Fig. 2C shows the result of four repeated clustering steps done with 10-dimensional score vectors, calculated with k ranging from 50 to 500 in steps of 50. The darker shade in the image corresponds to regions with differential abundance.

In contrast to the method applied in [5], the DA regions obtained by our approach are not constraint to any predefined clusters. Thus, although there are similarities between our findings and those of [5], there are some important differences. Five out of the six DA clusters described in [5] have partial overlaps with our DA regions:

$$G1-R1, \qquad G3-R2, \qquad G6-R3, \qquad G10-R4, \qquad G11-R3.$$

In [5] the clusters $G11, G6$ were found to be two distinct DA regions. In contrast $R3$, which is one of the DA regions detected by DA-seq, partially overlaps with $G6, G11$ and $G9$. We argue that identification of $G6$ and $G11$ as two separate DA regions and the exclusion of $G9$ as potentially relevant for DA is artificial. Unifying the clusters of exhausted lymphocytes allows us to zoom in to detect and transcriptionally characterize cell subpopulations (within this union) that are more specific to differences between responders and non-responders compared to differences detected by inspecting the clusters separately. In addition, though the DA regions $R1, R2$ and $R4$ partially overlap $G1, G3$ and $G10$, they are not identical. Unlike the pre-defined clusters that could bias the differential abundance analysis and result in diffused DA regions, our method simply finds the most discriminative regions irrespective if the underlying data has cluster structure or if it is a contiguous data cloud.

To quantify differential abundance, we define a "DA score" as well as a p-value for each DA region. Details on the computation of both are given in appendix D. Table 1 shows DA scores and p-values for the five

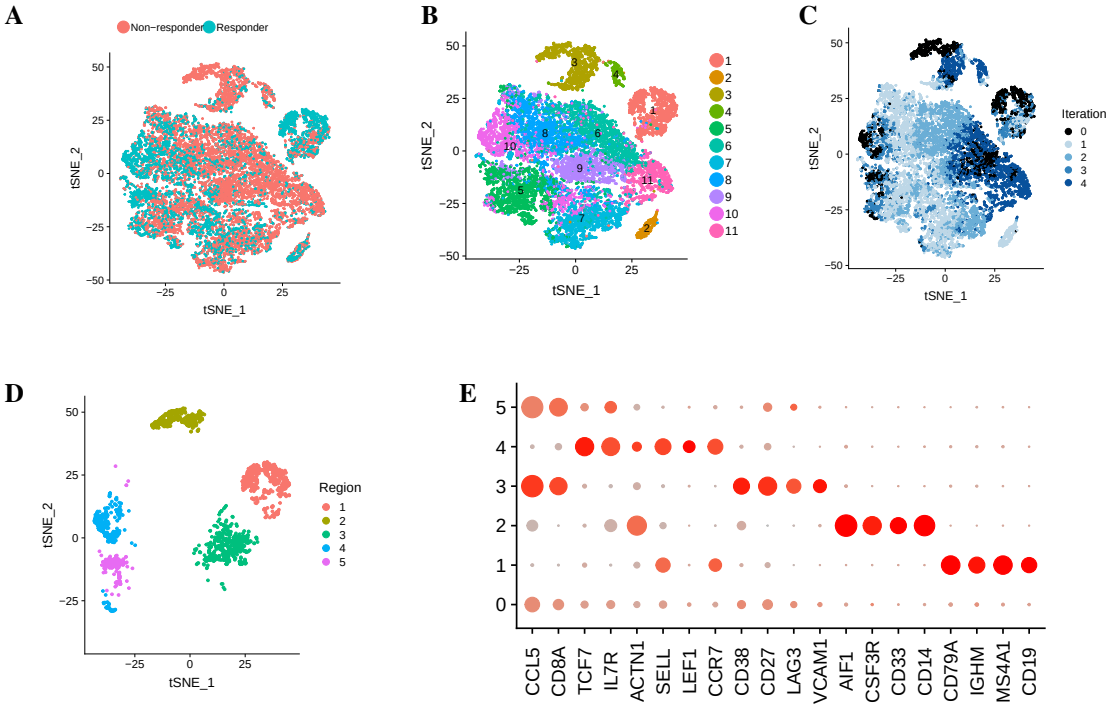

Figure 2: **Differential abundance (DA) results between immune cells from melanoma patients (responders vs non-responders).** t-SNE (t-distributed stochastic neighbor embedding) plot of 16,291 cells, colored by (A) status of response to immune therapy; (B) cluster labels from [5]; (C) which iteration a cell is removed in our iterative clustering process, "0" represents cells that remain after all iterations; (D) distinct DA regions obtained after the final clustering step of the remaining cells from (C); (E) Dotplot for markers characterizing the five selected DA regions; the size of each dot represents the percentage of cells expressing the corresponding gene, color intensity of each dot represents the average gene expression in cells from the corresponding region that express the corresponding gene. The fraction of cells and expected expression values of the genes in non-DA regions is shown in the lowest row of the dot plot, labeled 0.

DA regions detected by DA-seq (a positive DA score represents more responders, and a negative DA score represents more non-responders). The p-values of four ($R1 - R4$) out of the five regions are indicative of significant differential abundance.

Finally, we characterized these DA regions by identifying markers associated with each region, as shown by a dotplot in Fig. 2E. For the DA regions $R1, R2, R5$ and their corresponding clusters in [5] DA-seq detected similar lists of characteristic markers. In the DA region $R4$, DA-seq detected not only the markers of $G10$, but also the additional marker $ACTN1$, which has been reported to be related to T cell maturation. Interestingly, markers of $R3$ ($LAG3, CD27$) define an exhausted lymphocyte population [16, 17], covering three clusters associated with lymphocyte exhaustion. Importantly, the most significant gene relevant to differential abundance in $R3$ is $VCAM1$, whose importance relative to other genes was ranked much lower in [5], and whose role in immune escape has been studied [18, 19, 20, 21, 22].

Region	DA score	p-value
1	-0.801	0.000432
2	0.933	0.000536
3	0.892	0.000145
4	-0.810	0.003482
5	-0.841	0.195056

Table 1: Statistics of DA regions in melanoma data

3.2 Unraveling differentiation patterns of early mouse dermal cells

To further illustrate our approach, we tested it on the scRNA-Seq data obtained from developing embryonic mouse skin [9]. Cells from dorsolateral skin were sequenced at two time points of development, embryonic day E13.5 and E14.5, each with two biological replicates, see Fig. 3A. Dermal cells were selected for analysis, by using the marker $Col1a1$, to study hair follicle dermal condensate (DC) differentiation.

Method used in [9]. The authors embedded the cells via diffusion coordinates to capture the manifold structure of the scRNA-Seq data to study the transcriptional states of the cells. They then used the early DC marker $Sox2$ to identify differentiated DC cells, as well as the diffusion dimension that corresponds to DC cell differentiation, which they called the DC-specific trajectory. By observing this trajectory, they found that although it contained cells from both E13.5 and E14.5, there were notably more E14.5 cells at the terminus, representing differentiated DC cells.

Comparison between methods. In [9] the authors had prior knowledge that differentiated DC cells express $Sox2$. In contrast, with our approach, no prior knowledge was needed, and an unbiased comparison of dermal cells (Fig. 3A) between E13.5 and E14.5 was able to recover the differentiated DC cell population discussed in [9]. Additionally, other differentiation insights were obtained. Fig. 3B shows the result of clustering done with 10-dimensional score vectors, calculated with k ranging from 50 to 500 in steps of 50. As shown in Fig. 3C, we identified and characterized three DA regions, with markers for each DA region shown in 3E. Table 2 shows the 'DA score' and 'p-value' of these regions (a positive DA score represents more cells from E14.5, and a negative DA score represents more cells from E13.5). Out of the three DA regions detected by DA-seq the differential abundances of $R1$ and $R3$ are statistically significant. $R3$ is more abundant in E14.5, which perfectly corresponds to the $Sox2+$ differentiated DC cells (blue region $R3$ in Fig. 3C, $Sox2$ expression shown in Fig. 3D). We also discovered another significant DA region that provided differentiation insights: $R1$ (pink region $R1$ in Fig. 3C), which is more abundant in E13.5, likely representing early mesodermal cells related to somite development ($Scx, Meox1$).

Region	DA score	p-value
1	-0.972	0.0348
2	0.971	0.268
3	0.990	0.0461

Table 2: Statistics of DA regions in mouse dermal data

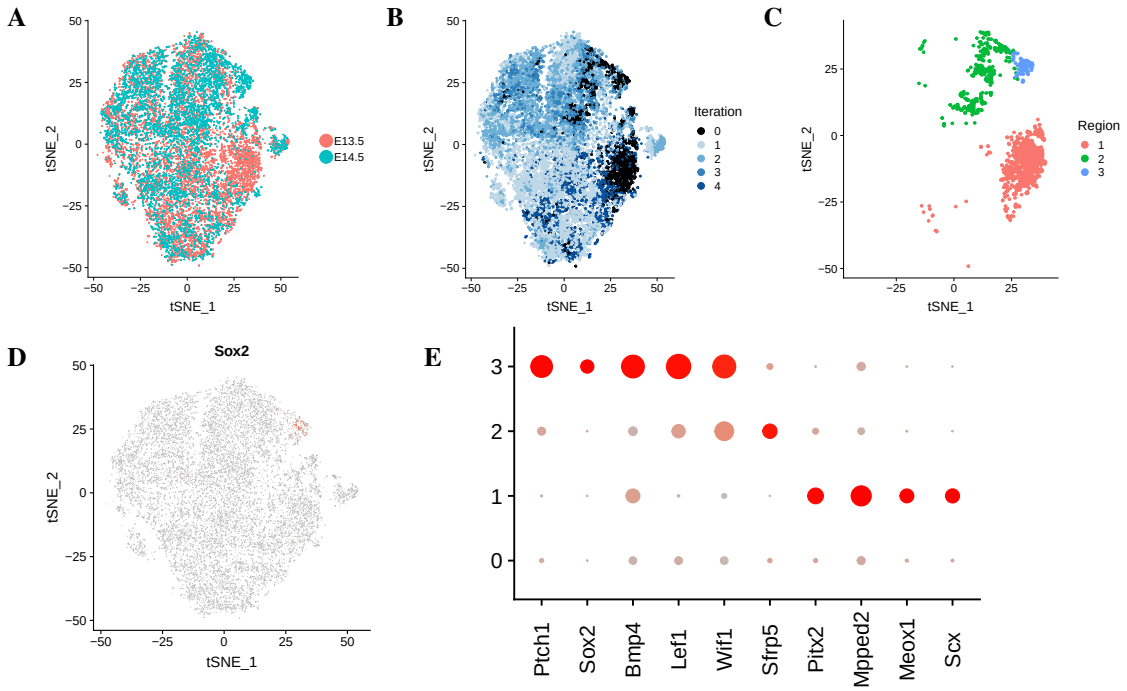

Figure 3: **DA regions between embryonic mouse dermal cells in embryonic day E13.5 and E14.5** t-SNE plot of 15,426 cells, colored by (A) embryonic day; (B) which iteration cells are removed, "0" represents DA cells that remain after 4 iterations; (C) distinct DA regions obtained after clustering the remaining cells from (B); (D) expression level of marker *Sox*2; (E) dot plot of markers for DA regions; the size of each dot represents the percentage of cells expressing the corresponding gene, color intensity of each dot represents the average gene expression in cells from the corresponding region that express the corresponding gene. The fraction of cells and expected expression values of the genes in non-DA regions is shown in the lowest row of the dot plot, labeled 0.

4 Discussion

In this work we presented DA-seq, a novel, multiscale approach for detecting subpopulations of cells that have differential abundance (DA) between scRNA-Seq data from different biological conditions or states. This approach enables to robustly delineate regions of substantially differential abundance between two samples. In contrast to existing methods, the subpopulations of cells we discover are not confined to any predefined clusters or cell subtypes. As illustrated in section 3, this difference impacts our ability to discover the most differentially abundant cell subpopulations, and the genes associated with them.

We note that batch effect removal is a necessary preprocessing step for DA-seq, in cases where there are noticeable batch effects between samples. Without proper calibration, the output DA regions may reflect both biological and technical differences between samples. To address this problem in the context of scRNA-Seq,

multiple batch effect removal methods have been developed [23], [3] and [24].

DA-seq is based on computation of a k-nearest neighbor (k-nn) graph, which may become a computational bottleneck for very large datasets. A standard method to compute k-nn graphs is via application of kd trees [25]. The complexity of constructing a kd tree is $O(n\log(n))$, and the average complexity for finding k-nearest neighbors is bounded by $O(kn\log(n))$. For datasets of the order of millions of cells, fast approximate approaches such as [26, 27] can be applied to increase the scalability of DA-seq.

Code availability An R implementation of DA-seq is freely available at `https://github.com/KlugerLab/DA-seq`.

Acknowledgements

The authors would like to thank Manolis Roulis, Jonathan Levinsohn, and Peggy Myung, for many useful discussions. Y.K. acknowledges support by NIH grants R01HG008383, R01GM131642, and 2P50CA121974.

References

[1] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[2] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8:14049, 2017.

[3] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411, 2018.

[4] Daniel B Burkhardt, Jay S Stanley, Ana Luisa Perdigoto, Scott A Gigante, Kevan C Herold, Guy Wolf, Antonio Giraldez, David van Dijk, and Smita Krishnaswamy. Enhancing experimental signals in single-cell rna-sequencing data using graph signal processing. *bioRxiv*, page 532846, 2019.

[5] Moshe Sade-Feldman, Keren Yizhak, Stacey L Bjorgaard, John P Ray, Carl G de Boer, Russell W Jenkins, David J Lieb, Jonathan H Chen, Dennie T Frederick, Michal Barzily-Rokni, et al. Defining t cell states associated with response to checkpoint immunotherapy in melanoma. *Cell*, 175(4):998–1013, 2018.

[6] James Kinchen, Hannah H Chen, Kaushal Parikh, Agne Antanaviciute, Marta Jagielowicz, David Fawkner-Corbett, Neil Ashley, Laura Cubitt, Esther Mellado-Gomez, Moustafa Attar, et al. Structural

remodeling of the human colonic mesenchyme in inflammatory bowel disease. *Cell*, 175(2):372–386, 2018.

[7] Xin Gao, Deqing Hu, Madelaine Gogol, and Hua Li. Clustermap: compare multiple single cell rna-seq datasets across different experimental conditions. *Bioinformatics*, 2019.

[8] Aaron TL Lun, Arianne C Richard, and John C Marioni. Testing for differential abundance in mass cytometry data. *Nature methods*, 14(7):707, 2017.

[9] Khusali Gupta, Jonathan Levinsohn, George Linderman, Demeng Chen, Thomas Yang Sun, Danni Dong, M Mark Taketo, Marcus Bosenberg, Yuval Kluger, Keith Choate, et al. Single-cell analysis reveals a hair follicle dermal niche molecular differentiation trajectory that begins prior to morphogenesis. *Developmental cell*, 48(1):17–31, 2019.

[10] Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[11] Ery Arias-Castro, Bruno Pelletier, and Venkatesh Saligrama. Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension. *Journal of Nonparametric Statistics*, 30(2):448–471, 2018.

[12] Larry Wasserman and John D Lafferty. Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems*, pages 801–808, 2008.

[13] Heinrich Fritz, Luis A Garcıa-Escudero, and Agustın Mayo-Iscar. tclust: An r package for a trimming approach to cluster analysis. *Journal of Statistical Software*, 47(12):1–26, 2012.

[14] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[15] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single cell data. *BioRxiv*, page 460147, 2018.

[16] Maher K Gandhi, Eleanore Lambley, Jaikumar Duraiswamy, Ujjwal Dua, Corey Smith, Suzanne Elliott, Devinder Gill, Paula Marlton, John Seymour, and Rajiv Khanna. Expression of lag-3 by tumor-infiltrating lymphocytes is coincident with the suppression of latent membrane antigen–specific cd8+ t-cell function in hodgkin lymphoma patients. *Blood*, 108(7):2280–2289, 2006.

[17] Sarah L Buchan, Teresa Manzo, Barry Flutter, Anne Rogel, Noha Edwards, Lei Zhang, Shivajanani Sivakumaran, Sara Ghorashian, Ben Carpenter, Clare L Bennett, et al. Ox40-and cd27-mediated costimulation synergizes with anti–pd-l1 blockade by forcing exhausted cd8+ t cells to exit quiescence. *The Journal of Immunology*, 194(1):125–133, 2015.

[18] Ken-Yu Lin, Dan Lu, Chien-Fu Hung, Shiwen Peng, Lanqing Huang, Chunfa Jie, Francisco Murillo, Jesse Rowley, Ya-Chea Tsai, Liangmei He, et al. Ectopic expression of vascular cell adhesion molecule-1 as a new mechanism for tumor immune evasion. *Cancer research*, 67(4):1832–1841, 2007.

[19] Heidi Harjunpää, Marc Llort Asens, Carla Guenther, and Susanna Carola Fagerholm. Cell adhesion molecules and their roles and regulation in the immune and tumor microenvironment. *Frontiers in immunology*, 10:1078, 2019.

[20] Tzyy Choou Wu. The role of vascular cell adhesion molecule-1 in tumor immune evasion. *Cancer research*, 67(13):6003–6006, 2007.

[21] Martin Schlesinger and Gerd Bendas. Vascular cell adhesion molecule-1 (vcam-1)—an increasing insight into its role in tumorigenicity and metastasis. *International journal of cancer*, 136(11):2504–2514, 2015.

[22] Deok-Hoon Kong, Young Kim, Mi Kim, Ji Jang, and Sukmook Lee. Emerging roles of vascular cell adhesion molecule-1 (vcam-1) in immunological disorders and cancer. *International journal of molecular sciences*, 19(4):1057, 2018.

[23] Uri Shaham, Kelly P Stanton, Jun Zhao, Huamin Li, Khadir Raddassi, Ruth Montgomery, and Yuval Kluger. Removal of batch effects using distribution-matching residual networks. *Bioinformatics*, 33(16):2539–2546, 2017.

[24] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421, 2018.

[25] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[26] Kiana Hajebi, Yasin Abbasi-Yadkori, Hossein Shahbazi, and Hong Zhang. Fast approximate nearest-neighbor search with k-nearest neighbor graph. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[27] George C Linderman, Gal Mishne, Yuval Kluger, and Stefan Steinerberger. Randomized near neighbor graphs, giant components, and applications in data science. *arXiv preprint arXiv:1711.04712*, 2017.

[28] Uri Shaham, Kelly Stanton, Henry Li, Ronen Basri, Boaz Nadler, and Yuval Kluger. Spectralnet: Spectral clustering using deep neural networks. In *International Conference on Learning Representations*, 2018.

A Preprocessing of single cell data

The R package *Seurat* [3] was used for most preprocessing steps for both datasets. Details are described below.

A.1 Melanoma dataset

Transcripts per million (TPM) scRNA-Seq data was obtained from [5]. We then performed data scaling and principal component analysis (PCA) with *Seurat*. To select genes as an input for PCA, we calculated the variance for each gene and kept only genes with variance larger than 6; top ten PCs were retained for the calculation of t-SNE embedding for visualization and diffusion coordinates for DA analysis, following the steps implemented in [5].

A.2 Mouse embryonic dataset

Raw count scRNA-Seq data of two time points E13.5 and E14.5 (two replicates each) was obtained from [9]. For each sample, we used *Seurat* to perform data normalization, scaling, variable gene selection, PCA, clustering, and t-SNE embedding calculation. As in [9], marker $Col1a1$ was used to select dermal clusters, and only cells in clusters with expression of $Col1a1$ were retained for further analysis. After selection of dermal cells, we first merged the biological replicates for each time point, then used canonical correlation analysis (CCA) to merge data from two time points and remove batch effect. Top 15 aligned CCA components were used to calculate the t-SNE embedding and diffusion coordinates of the combined dataset.

B Calculating diffusion map coordinates

The diffusion coordinates are a non linear and low dimensional representation useful for embedding data which is assumed to reside on a manifold. [14]. The coordinates are based on the following construction of a random walk process over the data. First, the affinity between all pairs of points x_i, x_j is computed with a kernel function $k(x_i, x_j)$. The kernel function captures the local geometry of the data. Thus, the choice of $k(x_i, x_j)$ should be guided by the application at hand. In our case, we use the standard Gaussian kernel function,

$$k(x_i, x_j) = \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\|x_i - x_j\|^2/2\sigma^2},$$

where d is the data dimension, and σ a proper scaling parameter. Next, the kernel function is normalized over the rows of $k(x_i, x_j)$ to obtain,

$$P(x_i, x_j) = \frac{1}{\sum_l k(x_i, x_l)} k(x_i, x_j).$$

The rows $P(x_i,)$ have positive elements and sum up to one. Thus, the matrix P has a probabilistic interpretation as *the Markov transition matrix* of the random walk process defined over the data points $\{x_i\}$. In other words, $P(x_i, x_j)$ can be thought of as the probability of transition from point x_i to x_j with a single random step. Similarly, the elements of the t-th power of P, denoted $(P^t)_{ij}$, correspond to the probability of transition between x_i, x_j in t random steps.

The *diffusion distances* $D_t(i, j)$ between points i, j is defined as the l_2 distance between the corresponding distributions $P(i,)$ and $P(j,)$,

$$D_t(i, j)^2 = \sum_{k=1}^{n} (P^t(i, k) - P^t(j, k))^2. \tag{4}$$

Notice, that the dependency on t allows us to represent the data at different scales (t). Increasing t reduces the number of significant eigenvectors and thus puts more weight on the leading ones.

The diffusion distance can be computed, up to a fixed relative error by a spectral analysis of the Markov matrix. Let $\{\lambda_i, \xi_i\}$ be the eigenvalues and corresponding eigenvectors of P in descending order such that

$\lambda_i \geq \lambda_{i+1}$. The diffusion distance can be approximated by the leading K eigenvectors

$$\boldsymbol{D}_t(i,j)^2 = \sum_{k=1}^{K} \lambda_k^{2t} (\xi_k(i) - \xi_k(j))^2 + O(\lambda_{K+1}^{2t}).$$

The diffusion coordinates ϕ_i for a point i are defined as,

$$\phi_i^t = [\lambda_1^t \xi_1(i), \ldots, \lambda_K^t \xi_K(i)] \in \mathbb{R}^K.$$

Notice, that the l_2 distance between ϕ_i and ϕ_j is equal, up to the relative accuracy λ^{2t}, to the diffusion distance $D_t(i,j)^2$, see [14, Proposition 6].

For very large datasets, diffusion map may become a computational bottleneck. To enhance scalability, one can use a fast deep learning based approaches to compute the diffusion maps coordinates (eigenvectors of the graph Laplacian) [28].

C Simulated dataset

In order to test our algorithm with "ground truth", we generated simulated dataset on the basis of the scRNA-Seq data from [5]. Specifically, we used the gene expression profiles from the real data, but assigned labels to cells manually to create artificial DA sites. In our simulation, we randomly selected four DA sites, as shown in Fig S1A. For cells within the DA sites, imbalanced cells labels were assigned (90% label 1, 10% label 2, or vice versa); on the other hand, labels for cells outside the DA sites were randomly assigned with 50% label 1 and 50% label 2. We then applied our algorithm on the simulated dataset. The cells with the most DA neighborhood and final DA regions from our algorithm are shown in Fig. S1B and C. Table S1 shows DA scores and p-values for the four detected DA regions (a positive DA score represents more cells from label 2, and a negative DA score represents more cells from label 2). Four out of four regions are significant. As can be seen, we successfully recovered all four artificial DA sites and did not introduce false positive regions.

Region	**DA score**	***p*-value**
1	-0.536	2.74e-08
2	-0.389	2.54e-07
3	0.581	5.04e-09
4	0.761	3.00e-09

Table S1: Statistics of DA regions in simulated data

C.1 Comparison with Cydar

To compare our algorithm with Cydar [8], we applied Cydar on the simulated dataset. Briefly, Cydar allocates cells into hyperspheres by randomly selecting a proportion of cells as centers and using a fixed

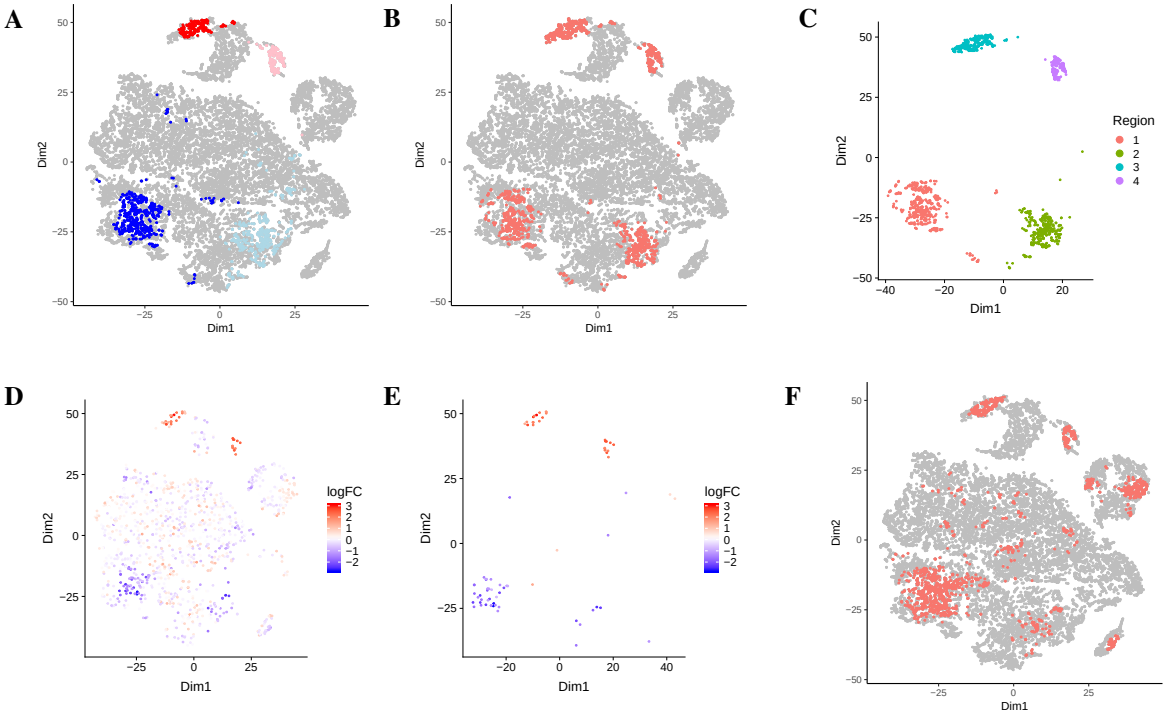

Figure S1: **DA results on simulated dataset.** t-SNE plot of 16,291 cells: (A) colored cells are four artificial DA sites: two sites on the top have more cells from label 2, while two sites on the bottom have more cells from label 1; (B) the most DA cells selected by our algorithm; (C) DA regions after clustering cells from B); (D) hyperspheres with Cydar, positioned at the center cell of each hypersphere, colored with log fold change of the abundance difference; (E) significant hyperspheres with $FDR < 0.05$; (F) cells from significant hyperspheres are highlighted.

radius, then tests differential abundance for each hypersphere with a negative binomial model. In Fig. S1, we plotted the centers of all hyperspheres and colored them based on the DA in terms of log fold change within each sphere (D) and indicated significant hyperspheres with $FDR < 0.05$ (E) obtained from Cydar. Since it is hard to evaluate Cydar's performance with just the hyperspheres, we then merged all significant hyperspheres and highlighted the cells in Fig. S1 F. Clearly, Cydar failed to recover all four DA sites, and identified several false positive sites. Results of Cydar showed limitation of the method on scRNA-Seq data due to: 1) centers of hyperspheres are randomly selected, and sometimes can not cover the whole data; 2) radius of the hypersphere is fixed which typically does not match with the actual size of the DA region. In contrast, our multi-scale approach detects the size of neighborhoods with significant differential abundance.

D DA statistics calculation

For each DA region detected by DA-seq, we compute a DA-score along with a p-value to indicate the significance of the differential abundance in the region. 'DA score' for a region is computed via,

$$DA\ score = \frac{\frac{x(R)}{n_x} - \frac{y(R)}{n_y}}{\frac{x(R)}{n_x} + \frac{y(R)}{n_y}}. \tag{5}$$

where $x(R), y(R)$ denote the number of cells in the DA region R from sample X and Y respectively.

Here we present a method to evaluate the differential abundance of DA regions that is applicable for cases where biological replicates for each state or condition are available. We compute the p-value based on the nonparametric Wilcoxon rank sum test to estimate statistical significance. This calculation is only applicable for datasets with two biological conditions (X,Y) and replicated samples in each condition: $X_1,\ldots,X_{m_X},Y_1,\ldots,Y_{m_Y}$, with m_X,m_Y denoting number of replicates for X,Y. For each sample S, a ratio r of a given DA region R is calculated by,

$$r(S) = \frac{N_{S\cap R}}{N_S}$$

where N_S denotes total number of cells in sample S, and $N_{S\cap R}$ denotes number of cells within region R and from sample S. Then, the test is used to assess whether the difference between $r(X) = \{r(X_1),\ldots,r(X_{m_X})\}$ and $r(Y) = \{r(Y_1),\ldots,r(Y_{m_Y})\}$ is significant. In cases where m_X and m_Y are small, like in [9] ($m_X = 2, m_Y = 2$), we use a standard two sample t-test instead of Wilcoxon. If no biological replicates are available, p-values of DA regions can be computed by other methods such as permutation tests, or partitioning of the data.