Detecting regions of differential abundance between scRNA-seq datasets

Jun Zhao[1,2,3,#], Ariel Jaffe[4,#], Henry Li[4], Ofir Lindenbaum[4], Xiuyuan Cheng[5], Richard Flavell[2], and Yuval Kluger[1,2,4,*]

[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

[#] Co-first authors
[*] To whom correspondence should be addressed: yuval.kluger@yale.edu

Abstract

Comparing the transcriptomic landscapes of two biological conditions has become a key challenge in single cell RNA sequencing analysis. Often, we observe a strong deviation in the number of cells within certain cell subpopulations between the two conditions. Analysis of such differentially abundant (DA) subpopulations may uncover cellular processes that differentiate the biological conditions. Typical methods for identifying DA subpopulations strongly rely on known cell types or unsupervised clustering methods. Here, we develop DA-seq, a multiscale algorithm that detects DA subpopulations not restricted to well separated clusters or known cell types. We applied DA-seq to four scRNA-seq datasets as well as two simulated ones. For the former, we compare our results to previously published findings. We find that for some cases, DA-seq is able to reveal subpopulations undetected in the original studies.

Introduction

Profiling biological systems with scRNA-seq is an invaluable tool, as it enables experimentalists to measure the expression levels of all genes over thousands to millions of cells [28, 41]. A prevalent challenge in scRNA-seq analysis is comparing the transcriptomic profiles of cells from two biological conditions [6, 7, 22]. Often, such comparison reveals subpopulations that are *differntially abundant* (DA). In DA subpopulations, the ratio between the abundances of cells from the two biological conditions differs significantly from the respective ratio in the overall data.

A standard approach to detect DA subpopulations is by first clustering the union of cells from both conditions. This step can be done in a completely unsupervised manner or by using known biomarkers. For each cluster, the proportion of cells from the two biological conditions is measured. A cluster in which these proportions significantly differ from the overall proportion in the data is considered differentially abundant. Finally, a subset of differentially expressed genes characterizing each DA cluster is identified. This approach was applied, for example, to compare responders and non-responders to cancer treatment [31], to quantify changes in PBMC cells under rest and interferon β stimulation [7], and to compare colon single cell data before and after treatment [20]. ClusterMap [13] is a variant of this approach, where the clustering step is applied separately to cells from the two conditions. Subsequently, the datasets are merged by matching similar clusters. This approach, however, might be suboptimal if the heterogeneous cell populations do not fall into well defined separate clusters. For example, DA subpopulations may be distributed among several adjacent clusters or, alternatively, encompass only a part of a cluster. Additionally, the clustering approach may fail in cases where the data represents a continuous process, such as cell cycles or certain developmental programs.

A different approach for identifying DA subpopulations that does not rely on initial clustering was derived by Lun et. al. [27] for mass cytometry data. Their algorithm performs multiple local two sample tests for hyperspheres centered at randomly selected cells. The caveat of this approach is that the selected hyperspheres may only partially overlap with the DA regions or fail to form spatially coherent regions. Accurate delineation of a DA region is essential for identifying the markers that differentiate the DA region from its immediate neighboring subpopulations as well as markers that separate it from the rest of the cells in the dataset.

Here, we develop DA-seq, a multiscale approach for detecting DA subpopulations. In contrast to clustering based approaches, our method can detect DA subpopulations that do not form well separated clusters. For each cell, we compute a multiscale differential abundance score measure. These scores are based on the k nearest neighbors in the transcriptome space for various values of k. The motivation behind multiscale analysis is that by employing a single scale one may miss some of the DA regions if the scale is too large, or detect spurious DA regions if the scale is too small. In the Results section, we show that our method is able to identify DA regions that overlap several clusters or are contained within a cluster. We verify our identified regions by comparing them to previously detected subpopulations.

Problem setup

Our goal is to detect subpopulations of cells that are differentially abundant between two biological states or conditions. Let $X = \{\boldsymbol{x}_i\}_{i=1}^n$ be the gene expression profiles of n cells collected in two experiments. In scRNA-seq, the number of genes is typically ~ 30000, while the number of cells ranges between $10^3 - 10^6$. Every cell is assigned a binary label $y_i \in \{-1, 1\}$, indicating from which of the two experiments it was collected. We assume that $\{\boldsymbol{x}_i\}$ are independent realizations of an m dimensional density function denoted $f_{\boldsymbol{x}}(\boldsymbol{x}) : \mathbb{R}^m \to \mathbb{R}$.

Our primary goal is to detect regions in the gene space where the conditional distribution $f_{\boldsymbol{x}|y}(\boldsymbol{x}|y = 1)$ differs significantly from $f_{\boldsymbol{x}|y}(\boldsymbol{x}|y = -1)$. We say that a point $\boldsymbol{x}$ is in a differentially abundant (DA) region if

$$|f_{\boldsymbol{x}|y}(\boldsymbol{x}|y = 1) - f_{\boldsymbol{x}|y}(\boldsymbol{x}|y = -1)| > \tau, \tag{1}$$

for some threshold τ. Next, we detect marker genes for each of the DA regions.

One approach to find DA regions is based on *local* two sample tests [8, 10, 19]. A global two sample test determines whether two sets of samples were generated by the same distribution. In contrast, local sample tests also detect the locations of any discrepancies between them. This approach was suggested, among others, by Lun et. al [27] and by Freeman et. al. [10]. These methods compute a test statistic in local neighborhoods around randomly selected cells $\{\boldsymbol{x}_i\}$. This statistic quantifies the discrepancy between two estimated conditional distributions, $\hat{f}_{x|y}(\boldsymbol{x}_i|y = 1)$ and $\hat{f}_{x|y}(\boldsymbol{x}_i|y = 0)$, and provides a local p-value for each $\boldsymbol{x}_i$. To correct for multiple testing, the Benjamini-Hoechberg procedure [3] is then applied. Lun et. al. [27] utilized this approach to find DA locations for mass cytometry data.

A different approach for obtaining DA regions was derived by Cazais et. al. [8], where a measure of local discrepancy is computed for all the points in the dataset instead of a random subset. The points with the highest measure of discrepancy are then aggregated into spatially coherent clusters. Thus, in contrast to the local two sample tests, the output of this approach is a small number of DA regions. In this work, we derive a new approach to detect DA regions for scRNA-seq. Our approach is based on a multiscale measure of differential abundance computed for each cell. Similarly to [8], our method obtains spatially coherent DA regions by clustering the points with the highest score. In addition to the DA regions, the output of our approach is a list of genes that characterize each of the regions. In the next section , we describe our approach in detail.

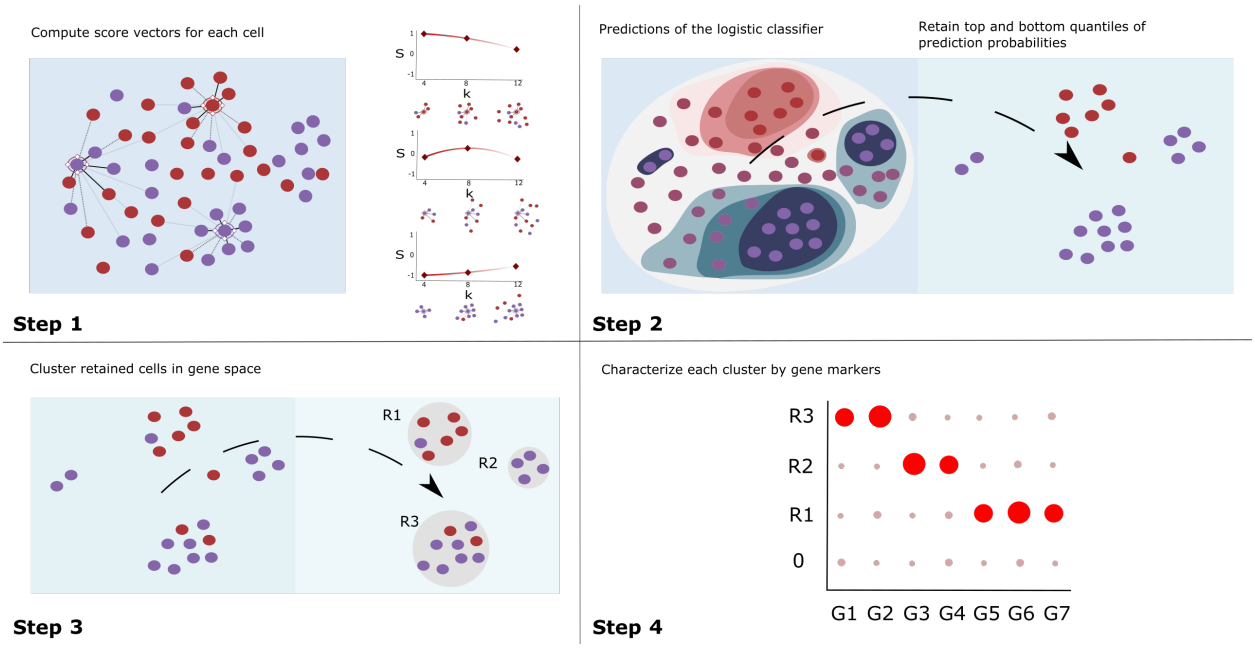

Figure 1. Schematic demonstration of DA-seq. Our method focuses on obtaining DA regions based on cells from two biological conditions, colored by red and purple. **Step 1:** Computing a *multiscale score* measure for each cell according to Eq. (5) for several values of k (e.g. $k = 4, 8, 12$). **Step 2:** Training a logistic classifier to predict the biological condition of each cell based on the multiscale score (see Eq. (6)). The algorithm retains only cells for which the output is above a threshold τ_h or below τ_l and hence may reside in DA regions. **Step 3:** Clustering the cells retained in step 2 with outlier removal to obtain spatially coherent DA regions. These regions are denoted $R1, R2$ and $R3$. **Step 4:** Applying feature selection via stochastic gates to detect subsets of genes that characterize each of the DA regions. For example, $G1$ and $G2$ characterize $R3$ etc.

Materials and Methods

In this section we develop DA-seq, a novel multiscale approach to detect the DA regions defined in Eq. (1). DA-seq consists of four main steps described in the following four subsections. Briefly, these steps involve (i) Computation of a multiscale score vector for each cell. (ii) Applying a logistic regression classifier to predict the biological state of all cells, given the score vectors as input. The classification probability allows us to evaluate whether a transcriptome of a cell is localized in a neighborhood dominated by cells from one of the biological states. (iii) Clustering the cells whose classification probabilities are greater than a prescribed threshold (high measure of differential abundance) into spatially coherent regions. (iv) Applying our novel ℓ_0 based feature selection neural network model to select genes that differentiate a DA region from the rest of the data. The rest of this section describes these steps in detail. Algorithm 1 outlines the different steps of DA-seq.

Step 1: Computing a multiscale score measure. First, we apply Bayes rule to Eq. (1) and obtain

$$f_{\boldsymbol{x}|y}(\boldsymbol{x}|y=1) - f_{\boldsymbol{x}|y}(\boldsymbol{x}|y=0) = f_{\boldsymbol{x}}(\boldsymbol{x})\left(\frac{f_{y|\boldsymbol{x}}(y=1|\boldsymbol{x})}{f_y(y=1)} - \frac{f_{y|\boldsymbol{x}}(y=-1|\boldsymbol{x})}{f_y(y=-1)}\right).$$

For a specific cell $\boldsymbol{x}_i$, we define at a scale k the k-nn estimator of $f_{y|\boldsymbol{x}}(y=1|\boldsymbol{x}_i)$ and $f_{y|\boldsymbol{x}}(y=-1|\boldsymbol{x}_i)$ by

$$\hat{f}_{y|\boldsymbol{x}}(y=1|\boldsymbol{x}_i) = \frac{n_A(\boldsymbol{x}_i;k)}{k} \qquad \hat{f}_{y|\boldsymbol{x}}(y=-1|\boldsymbol{x}_i) = \frac{n_B(\boldsymbol{x}_i;k)}{k}, \tag{2}$$

where $n_A(\boldsymbol{x}_i;k)$ and $n_B(\boldsymbol{x}_i;k)$ denote the number of cells among the k nearest neighbors of $\boldsymbol{x}_i$ with positive and negative labels, respectively. In addition, we define the estimators $\hat{f}_y(y=1)$ and $\hat{f}_y(y=-1)$ by

$$\hat{f}_y(y=1) = n_A/n \qquad \hat{f}_y(y=1) = n_B/n, \tag{3}$$

where n_A and n_B are the total number of cells with positive and negative labels respectively. Inserting (2) and (3) into (2) yields

$$\hat{f}_{\boldsymbol{x}|y}(\boldsymbol{x}|y=1) - \hat{f}_{\boldsymbol{x}|y}(\boldsymbol{x}|y=0) = \frac{f_{\boldsymbol{x}}(\boldsymbol{x})n}{k}\left(\frac{n_A(\boldsymbol{x};k)}{n_A} - \frac{n_B(\boldsymbol{x};k)}{n_B}\right). \tag{4}$$

The estimator in Eq. (4) depends on the choice of the number of neighbors k. However, a single global value for k may be appropriate only for certain regions in the data while being completely suboptimal in other regions. We therefore compute $n_A(\boldsymbol{x};\boldsymbol{k})$ and $n_B(\boldsymbol{x};\boldsymbol{k})$ at l different nearest neighborhood scales $\boldsymbol{k} = [k_1,\ldots,k_l]$. We define the normalized score vector $\boldsymbol{s}(\boldsymbol{x}_i;\boldsymbol{k}) = [s_1(\boldsymbol{x}_i;k_1),\ldots,s_l(\boldsymbol{x}_i;k_l)]$ by

$$s_i(\boldsymbol{x}_i;k_j) = \left(\frac{n_A(\boldsymbol{x}_i;\boldsymbol{k})}{n_A} - \frac{n_B(\boldsymbol{x}_i;\boldsymbol{k})}{n_B}\right) \Bigg/ \left(\frac{n_A(\boldsymbol{x}_i;\boldsymbol{k})}{n_A} + \frac{n_B(\boldsymbol{x}_i;\boldsymbol{k})}{n_B}\right). \tag{5}$$

The numerator in (5) is proportional to the $k-nn$ estimator in (4) while the denominator normalizes the values to be within the range $[-1,1]$.

Fig. 1(Step 1) illustrates the qualitative behaviour of the score vector $\boldsymbol{s}(\boldsymbol{k})$ for three cells located in different regions. The upper vector contains positive entries and corresponds to a cell $\boldsymbol{x}_i$ in a DA region where $f_{y|\boldsymbol{x}}(y=1|\boldsymbol{x}_i) > f_{y|\boldsymbol{x}}(y=-1|\boldsymbol{x}_i)$. Thus, the score is high for small values of k. As k increases, the score typically decreases since at this scale the neighbors may contain a more balanced proportion of cells with positive and negative labels and even include neighbors positioned outside of the DA region.

Step 2: Computing a DA statistic for each cell. The output of step 1 consists of multiscale score vectors. Cells in DA regions whose neighborhoods are enriched with positive labels tend to be closer to each other in the l-dimensional score space than cells whose neighborhoods are enriched with negative labels or not enriched by any of the labels.

Our task in step 2 is to map the l dimensional score vector $\boldsymbol{s}(\boldsymbol{x};\boldsymbol{k})$, defined in (5), into a single DA statistic for each cell. To that end, we use a logistic regression classifier. The classifier is trained to predict the class label y_i of each cell given its l-dimensional score vector $\boldsymbol{s}(\boldsymbol{x}_i;\boldsymbol{k})$. Specifically, we compute a vector $\boldsymbol{w}^*$ that minimizes the following loss,

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{n}(1-y_i)\log(1-\sigma(\boldsymbol{s}(\boldsymbol{x}_i;k)^T\boldsymbol{w}) - \sum_{i=1}^{n} y_i \log\sigma(\boldsymbol{s}(\boldsymbol{x}_i;k)^T\boldsymbol{w}),$$

where σ is the sigmoid function. In other words, the logistic classifier assigns a numerical value between 0 and 1 based on each cell's score vector,

$$d_i = \sigma(\boldsymbol{s}(\boldsymbol{x}_i;k)^T\boldsymbol{w}^*) \in [0,1].$$

The classifier is trained to increase d_i if $y_i = 1$ and decrease its value if $y_i = 0$. Notably, as we explain below, the properties of the logistic classifier imply that a high value of d_i is a strong indication that the cell is located in a (score-vector space) region enriched with positive labels, and vice versa.

Figure 1 (Step 2) illustrates the output of the logistic regression classifier. The figure shows a heatmap, where each cell is colored by the prediction of the logistic classifier. The cells that reside in DA regions are determined by thresholding the predictions of the logistic regression. Thus, we determine that $\boldsymbol{x}_i$ is part of a positive DA region if $d_i > \tau_h$ and of a negative DA region if $d_i < \tau_l$.

Logistic regression, smoothness and early stopping. As explained above, the predictions of the logistic regression classifier are taken as a DA measure for each cell. The output of the logistic regression classifier is a prediction function which is locally *smooth* over the input space. This means that the places where the prediction function returns a value close to 1 (or 0) correspond to regions with significantly more cells of class 1 than class 2 (or vice versa). In contrast, non-significant DA regions do not fall in the upper or lower quantiles of the prediction function, because of the comparable abundance of the two classes. This smoothness property of the prediction function is important, as it makes our DA method less sensitive to finite-sample effects in the data and the influence of a few outlier cell measurements. In other words, the imposed smooth prediction function of the logistic regression prevents the detection from overfitting and can be viewed as a regularization in our method. Theoretically, the smoothness of the prediction function assumes that the transition from one significant DA region (say of class 1) to another (say of class 2) is relatively slow and cannot have sharp interface, which is fulfilled by our DA problem due to a large portion of neutral regions in the space of cell measurements. The level of smoothness is quantitatively controlled by the strength of regularization, which can be found by e.g. cross validation as is standard practice in machine learning. We have also observed in experiments that regularizing the smoothness of the prediction function improves the accuracy of DA-seq.

To apply regularization and guarantee the smoothness of the prediction function, we train our classifiers for differentiating cells from two samples with early stopping. Early stopping is a frequently used technique for training neural network models in practice. Recent research has shown that training overparameterized models with early stopping leads to functions with precisely these properties. [2, 30] prove a correspondence between early stopping and learning low frequency harmonic functions on the feature space. [23] derives guarantees on the robustness of classifying functions to label noise when trained with early stopping. The algorithmic parameter which controls the strength of regularization, and hence the level of smoothness of the trained prediction function, is the number of steps in the early stopping training. In our preliminary experiments with training a neural network classifier for DA detection, we also successfully applied early stopping, see more in the discussion section.

Step 3: Clustering the DA cells into spatial regions. This step involves clustering the subset of cells whose prediction probabilities are above the threshold. The clustering is performed in the original gene space or in its compressed principal component subspace. This allows us to characterize DA regions in the gene space or any desired embedding such as t-SNE. For robustness, we apply clustering with outliers. Specifically, we used the R package tclust [11], which implements k-means with outliers. The main difference with the standard k-means is that in each iteration the centers of the k clusters are updated according $1 - \kappa$ fraction of the data points in each cluster, where κ is a parameter determined by the user. We note, however, that other clustering methods can be used for this step. The output of this step is a list of regions where each region is assigned a subset of cells. In our next section, we describe a feature selection approach to compute the characterizing genes for each region.

Step 4: Differential Expression Analysis as a Feature Selection Problem. Differential expression analysis (DEA) and feature selection are related tasks. In DEA, one applies univariate statistical tests to discover biological markers that are typical of a certain condition or disease. This approach is typically used for its simplicity and interpretability. As they treat each gene individually, however, they ignore multivariate correlations. Feature selection, on the other hand, seeks an interpretable, simplified, and often superior classification model that uses a small number of genes. Here, we use our recently proposed embedded

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

Input: Merge gene expression profiles of n cells $\{\boldsymbol{x}_i\}$ from two biological conditions (two classes) and assign to each cell a binary class label

Output: A list of DA regions where each region is linked to a subset of cells and a list of genes differentiating it from the remaining cells not included in the region.

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.

Step 1. Calculate score vector $\boldsymbol{s}(\boldsymbol{x};\boldsymbol{k})$ for each cell.
Parameters: a range of l values for k.
for *each cell* i **do**
| compute its score vector s_i for a given range of k values via Eq. (5).
end

Step 2. Obtain a list of DA cells.
Parameters: Low and high thresholds τ_l, τ_h
Train a logistic regression to predict the class labels for all cells using the score vectors $\boldsymbol{s}$ as input feature vectors of the classifier. The cells for which the logistic output is higher than τ_h or lower than τ_l are defined as DA cells.

Step 3. Cluster the DA cells obtained in step 2 into spatial regions in the (dimensionally-reduced) gene space
Parameters: number of clusters k, outlier ratio κ.
Clustering the DA cells with outlier removal with *t-clust* [11]. This clustering is performed on the cells' transcriptomic profiles after dimensionality reduction (**Step 0**).

Step 4. Select genes that differentiate each of the DA regions from the rest of the data.

feature selection [40] method to discover for each DA region a subset of genes that collectively have a profile characteristic for that region, which separates it from the rest of the data.

Given observations $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$, the problem of feature selection could be formulated as an empirical risk minimization

$$\min_{\boldsymbol{\theta}} \frac{1}{N}\sum_{n=1}^{N} L(\boldsymbol{\theta}^T\boldsymbol{x}_n, y_n) \quad \text{s.t. } \|\boldsymbol{\theta}\|_0 \leq r, \tag{6}$$

where r is the number of selected features, L is the loss function and $\boldsymbol{\theta}$ are the parameters of a linear model or more complex neural net model. Due to the ℓ_0 constraint, the problem above is intractable. In practice, the ℓ_0 norm is typically replaced with the ℓ_1 norm, which yields a convex optimization problem as implemented in the popular LASSO optimization approach [35]. Nonetheless, we recently surmounted this obstacle by introducing a stochastic gates (STG) approach to neural networks, which provides a non convex relaxation of the optimization in (6). Each STG is a relaxed Bernouli variable z_d, where $\mathbb{P}(z_d = 1) = \pi_d, d = 1, ..., D$, and D is the total number of genes. Now the risk minimization in (6) could be reformulated by gating the variables in $\boldsymbol{x}$ and minimizing the number of expected active gates. This yields the following objective

$$\min_{\boldsymbol{\theta},\boldsymbol{\pi}} \mathbb{E}_S\left[\frac{1}{N}\sum_{n=1}^{N} L(\boldsymbol{\theta}^T\boldsymbol{x}_n \odot \boldsymbol{z}, y_n) + \lambda\|\boldsymbol{z}\|_0\right]. \tag{7}$$

Objective (7) could be solved via gradient descent over the parameters of the model $\boldsymbol{\theta}$ and the gates $\boldsymbol{\pi}$. To identify genes that are typical to a DA region, we train a model that minimizes (7) by sampling multiple balanced batches from the DA region vs. the backgrounds. Then we explore the distribution of genes that were selected by the model: all d's such that $\pi_d \geq 0$. Note that λ is a regularization parameter that controls the number of selected genes; it could be tuned manually for extracting a certain number of genes or, alternatively, using a validation set by finding λ which maximizes the generalization accuracy.

Preprocessing of scRNA-seq datasets

The R package *Seurat* was used for most preprocessing steps for scRNA-seq datasets discussed in the paper. Details are described below.

Melanoma dataset

Transcripts per million (TPM) scRNA-seq data was obtained from [31]. 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; the top ten PCs were retained for the calculation of t-SNE embedding and DA analysis, following the steps implemented in [31].

Mouse embryonic dataset

Raw count matrices of scRNA-seq data from two time points E13.5 and E14.5 (two replicates each) were obtained from [14]. For each sample, we used *Seurat* to perform data normalization, scaling, variable gene selection, PCA, clustering and t-SNE embedding calculation. As in [14], marker $Col1a1$ was used to select dermal clusters, and only cells in clusters with expression of $Col1a1$ were retained for further analysis. After selecting dermal cells, we first merged the biological replicates for each time point. We then used canonical correlation analysis (CCA) to merge data from two time points and remove batch effect. The top 15 aligned CCA components were used to calculate the t-SNE embedding and diffusion coordinates of the combined dataset.

Aging brain dataset

The normalized expression matrix of scRNA-seq data from young and old mice was generated by Ximerakis et al. [39]. Cell metadata - including ceype,cell sample labels (from young and old mice) - was also obtained from the original paper. As described in [39], PCA was carried out after the identification of variable genes by the "mean variance plot" method from *Seurat*. The top 20 PCs were retained to calculate 2D embedding with t-SNE. DA-seq was performed on the top 20 PCs.

Bone marrow dataset

The raw count data matrix of scRNA-seq from bone marrow cells was obtained and filtered as described in [36]. There are two batches in the data: the niche batch and 5-FU experiment batch (of which there are control and 5-FU treatment samples). For each sample, we performed data normalization, scaling and variable gene identification with *Seurat*. In order to remove batch effect, CCA was applied to merge datasets from the two batches. The top 30 components were aligned and retained for t-SNE embedding calculation and DA-seq analysis.

Results

DA-seq is designed to reveal regions of differential abundance. These regions partially overlap with DA clusters that are found by standard clustering-based DA analysis. We find that DA-seq allows us to reproduce the results obtained by these standard methods. Importantly, it also allows us to uncover markers characteristic to a DA region that might not be detected by standard clustering-based approaches. In this section, we apply our method to four publicly available single cell datasets [14, 31, 36, 39] (refer to the appendix for additional results based on simulated data). In datasets [14, 31], most cells are not distributed in a clear structure with well-separated clusters. Instead, in many regions in the transcriptome gene space, they are distributed in a continuum where cluster partitioning is arbitrary. In this scenario, cluster-based approaches may miss regions of differential abundance. In the other two datasets [36, 39], the cell distributions can be partitioned into distinct clusters. Despite the clear partitioning in this second scenario, cluster-based approaches may fail to detect DA regions that encompass only sub-volumes within their respective clusters.

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

One of the goals of the Sade-Feldman et al. study [31] was to identify factors related to the 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.

Sade-Feldman et al. [31] clustered the 16,291 immune cells into 11 distinct clusters (Fig. 2B). Subsequently, they counted the number of cells from responders and non-responders in each of the predefined clusters and compared the relative abundance between these two groups. Two clusters ($G1, G10$) were enriched in cells from the responders' samples, and four clusters ($G3, G4, G6, G11$) were enriched in cells from non-responders' samples. Finally, the authors composed a list of genes that were highly expressed in these six differentially abundant clusters.

DA-seq versus standard DA analysis.

We applied DA-seq and obtained five DA regions denoted $R1 - R5$ (see Fig. 2D). In Fig. 2C, we color each cell in the t-SNE plot by its prediction probability that its neighborhood has an excess of cells from the pool of non-responder samples relative to cells from the pool of responder samples as computed in step 2 of the algorithm. To compute the score vectors of each cell, we employed k nearest neighborhoods with k ranging from 50 to 500 in steps of 50. The larger the value (see color-bar) associated with a cell, the more likely it is to belong to the pool of cells from non-responder samples. Similarly, the smaller the value, the higher the likelihood the cell belongs to the pool of responder samples.

In contrast to the method applied in [31], the DA regions obtained by our approach are not constrained to any predefined clusters. Thus, there are some important differences between our findings and those of [31] despite some similarities. Five out of the six DA clusters described in [31] have partial overlaps with our DA regions:

$$G1 - R1, \quad G3 - R3, \quad G6 - R2, \quad G10 - R5, \quad G11 - R2.$$

In [31], the clusters $G11$ and $G6$ are reported as two distinct DA regions. In contrast, $R2$, 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 are artificial. Unifying the clusters of exhausted lymphocytes allows us to detect and transcriptionally characterize cell

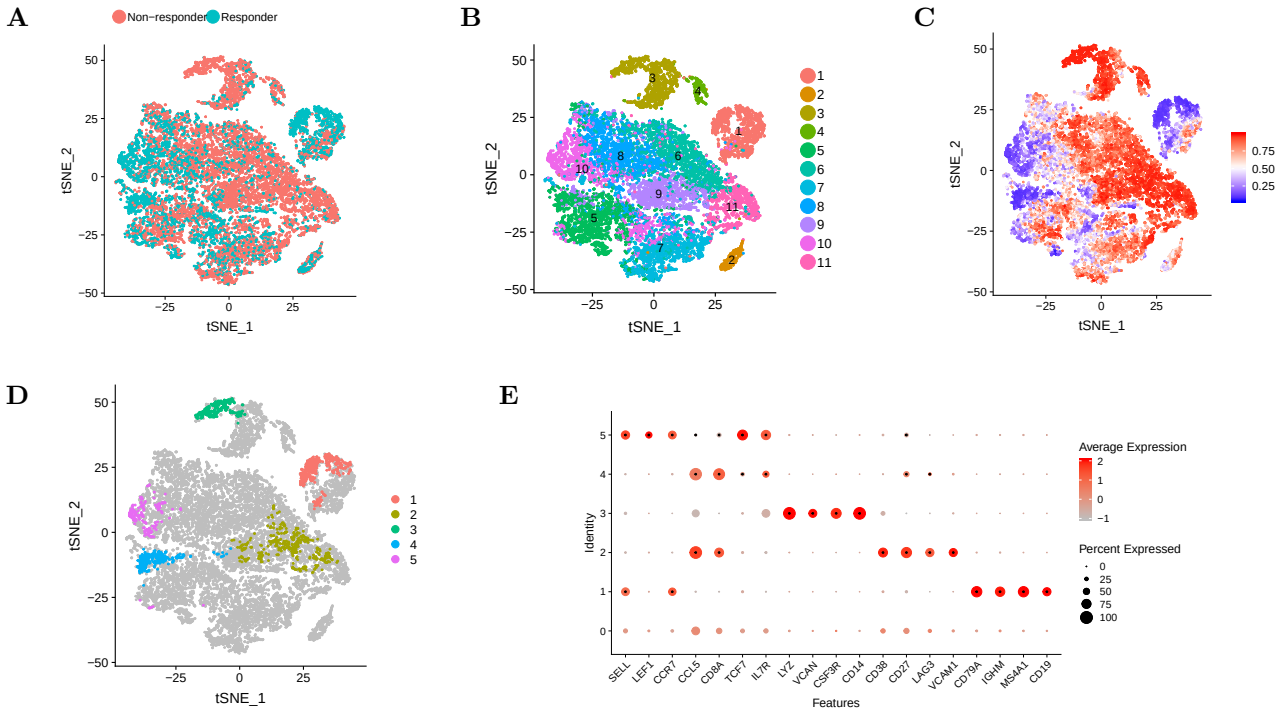

Figure 2. Immune cells from responding and non-responding melanoma patients treated with checkpoint therapy. t-SNE plots of 16,291 cells, colored by (A) status of response to immune therapy; (B) cluster labels from [31]; (C) the predictions of the logistic classifier. Large values indicate a high abundance of cells from the pool of non-responder samples. (D) distinct DA regions obtained by clustering cells from the top and bottom 7.5% quantiles of the predictions in (C); (E) dot plot for markers characterizing the five selected DA regions. The size of each dot represents the percentage of cells expressing the corresponding gene. The color intensity of each dot corresponds to the average gene expression across all cells in the DA region excluding the cells with zero expression values. For each DA region (row), we mark the genes selected by STG with black point in the center of the respective dots. The lowest row in the plot corresponds to the cells in the non DA regions.

subpopulations within this union that are more specific to differences between responders and non-responders. We observe that DA regions $R1, R3$ and $R5$ partially overlap with $G1, G3$ and $G10$, respectively, but they are not identical; furthermore, DA region $R4$ partially overlaps with cluster $G5$ which was not identified as a DA cluster. Unlike the predefined clusters that could bias the differential abundance analysis and result in diffused DA regions, our method simply finds the most discriminative regions regardless of whether the underlying data has either a cluster structure or 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 the appendix. Table 1 shows DA scores and p-values for the five DA regions detected by DA-seq (a positive (negative) DA score indicates that there are more (less) cells from non-responder samples than from responder samples in that DA region). The p-values of four ($R1 - R3, R5$) out of the five regions are indicative of significant differential abundance.

Finally, we identified markers that characterize the DA regions by applying feature selection via stochastic gates. A subset of the identified markers are shown by a dot plot in Fig. 2E. For the DA regions $R1, R3, R4, R5$ and their corresponding clusters in Fig. 2B, DA-seq detected similar lists of characteristic markers. Interestingly, the characteristic $R2$ markers $LAG3$ and $CD27$ define an exhausted lymphocyte population [5, 12] covering three clusters associated with lymphocyte exhaustion. Notably, the most significant gene predicted by DA-seq

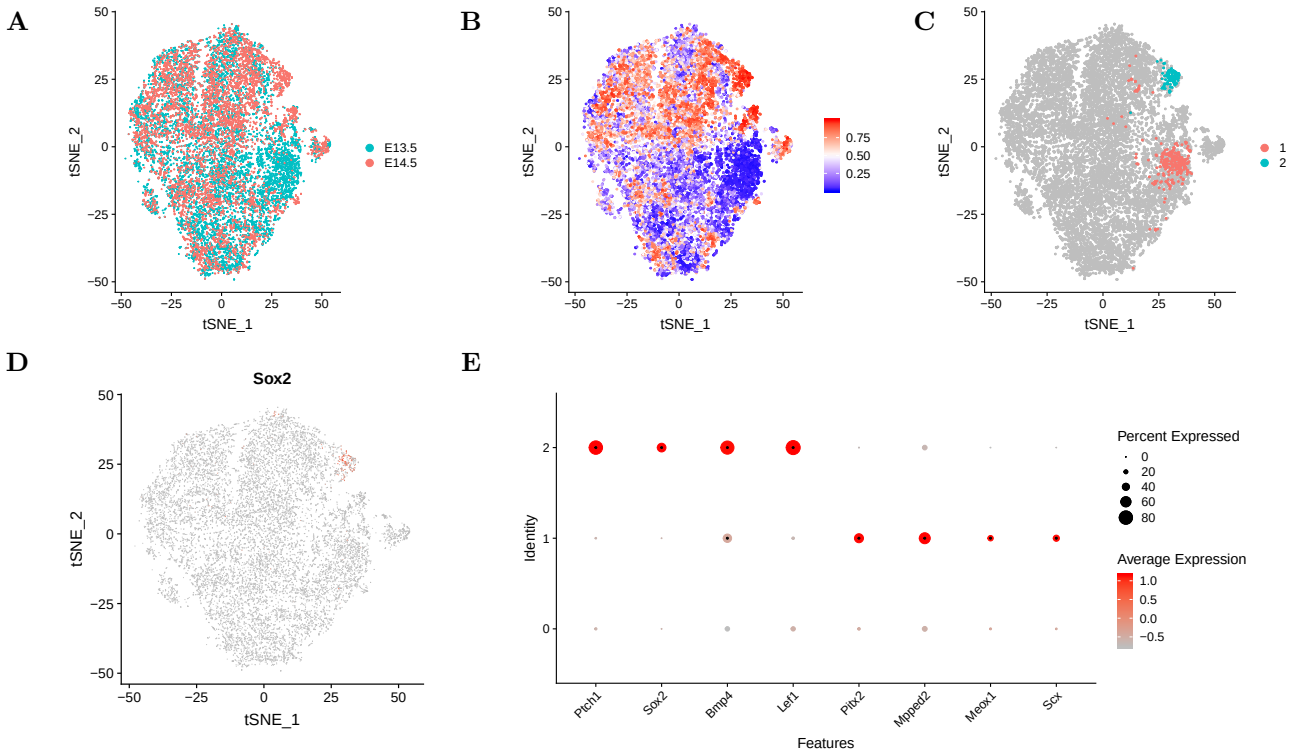

Figure 3. Comparing embryonic mouse dermal cells in embryonic day E13.5 and E14.5. t-SNE plots of 15,426 cells, colored by (A) embryonic day; (B) the predictions of the logistic classifier. High (low) values are indicative that the cell's neighborhood has excess (deficiency) of E14.5 cells relative to E13.5 cells. (C) distinct DA regions obtained by clustering cells from the top and bottom 5% quantiles of the predictions in (B); (D) cells expressing $Sox2$ overlap with the blue DA region in (C); (E) dot plot of several markers that characterize DA regions; details about size, color and non-DA regions information is as in Fig. 2E.

relevant to differential abundance in $R3$ is $VCAM1$, whose importance relative to other genes was ranked much lower in [31]. $VCAM1$ has a potentially an important role in immune escape as has been studied in [17, 21, 24, 32, 38].

Differentiation patterns of early mouse dermal cells

We applied DA-seq to scRNA-seq data obtained from developing embryonic mouse skin [14]. Cells from dorsolateral skin were sequenced at two time points of development, embryonic days E13.5 and E14.5, each with two biological replicates, see Fig. 3A. Dermal clls were selected for analysis by using the marker $Col1a1$

Table 1. Statistics of DA regions in melanoma data

Region	DA score	p-value
1	-0.960	2.23e-05
2	0.990	3.74e-06
3	1.00	4.98e-04
4	-0.972	0.230
5	-0.942	4.26e-05

to study hair follicle dermal condensate (DC) differentiation.

Gupta et al. [14] studied the transcriptional states of the cells by embedding them via diffusion coordinates to capture the manifold structure of the scRNA-seq data 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.

DA-seq versus analysis guided by prior knowledge.

In [14], the authors had prior knowledge that differentiated DC cells express $Sox2$. In contrast to the analysis in [14], our approach does not require prior knowledge. We obtain an unbiased comparison of dermal cells (Fig. 3A) between E13.5 and E14.5, which revealed the differentiated DC cell population discussed in [14]. Fig. 3B colors each cell according to the prediction of the logistic classifier trained on the 10-dimensional score vectors. The number of k nearest neighbors ranges from 50 to 500 in steps of 50. Fig. 3C shows two DA regions identified by DA-seq. Fig. 3E shows the characteristic marker genes associated with these DA regions.

Table 2 shows the 'DA score' of these regions (a positive DA score indicates that there are more cells from E14.5 in this DA region, and a negative DA score indicates more cells from E13.5). Due to lack of replicated samples in this dataset (two replicates for both E13.5 and E14.5), we did not compute a p-value. Instead, we computed the 'DA score' for the two DA regions for every possible pairwise comparison of these samples. Specifically, E13.5 replicate 1 vs. E14.5 replicate 1, E13.5 replicate 1 vs. E14.5 replicate 2, E13.5 replicate 2 vs. E14.5 replicate 1 and E13.5 replicate 2 vs. E14.5 replicate 2. $R2$ is more abundant in E14.5, which corresponds to the $Sox2+$ differentiated DC cells. This can be seen by inspecting the position of $R2$ (blue region) in Fig. 3C and the positions of cells expressing $Sox2$ shown in Fig. 3D. We also identified another significant DA region $R1$ (pink region in Fig. 3C), which is more abundant in E13.5. Interestingly, the Scx and $Meox1$ genes that characterize the cell subpopulation in this DA region link it to early mesodermal cells related to somite development.

Comparing brain RNA profiles of young and old mice.

Ximerakis et al. [39] characterized differences between brain cells of young and old mice, as shown in Fig. 4A. As shown in 4B, they detected 25 distinct cell populations, of which oligodendrocyte precursor cells ($C21$), neuronal restricted precursors ($C17$) and immature neurons ($C14$) all exhibited statistically significant decreases in old mice.

We apply DA-seq on the merged data and identify three DA regions that are more abundant in young mice, as shown in Fig. 4D. Specifically, $R3$ is a mixture of neuronal restricted precursors ($C17$) and immature neurons ($C14$), $R1$ corresponds to oligodendrocyte precursor cells ($C21$) and $R2$ is a subgroup within microglial cells ($C11$). The differential abundance in each of these three DA regions is statistically significant (Table 3). Fig. 4E shows the dot plot obtained by DA-seq which is qualitatively similar to the results obtained from the analysis by [39]. In other words, marker genes characterizing the DA regions obtained by DA-seq agree with

Table 2. Statistics of DA regions in mouse dermal data

Region	DA score	$E13^1/E14^1$	$E13^1/E14^2$	$E13^2/E14^1$	$E13^2/E14^2$
1	-0.783	-0.839	-0.697	-0.859	-0.733
2	1	1	1	1	1

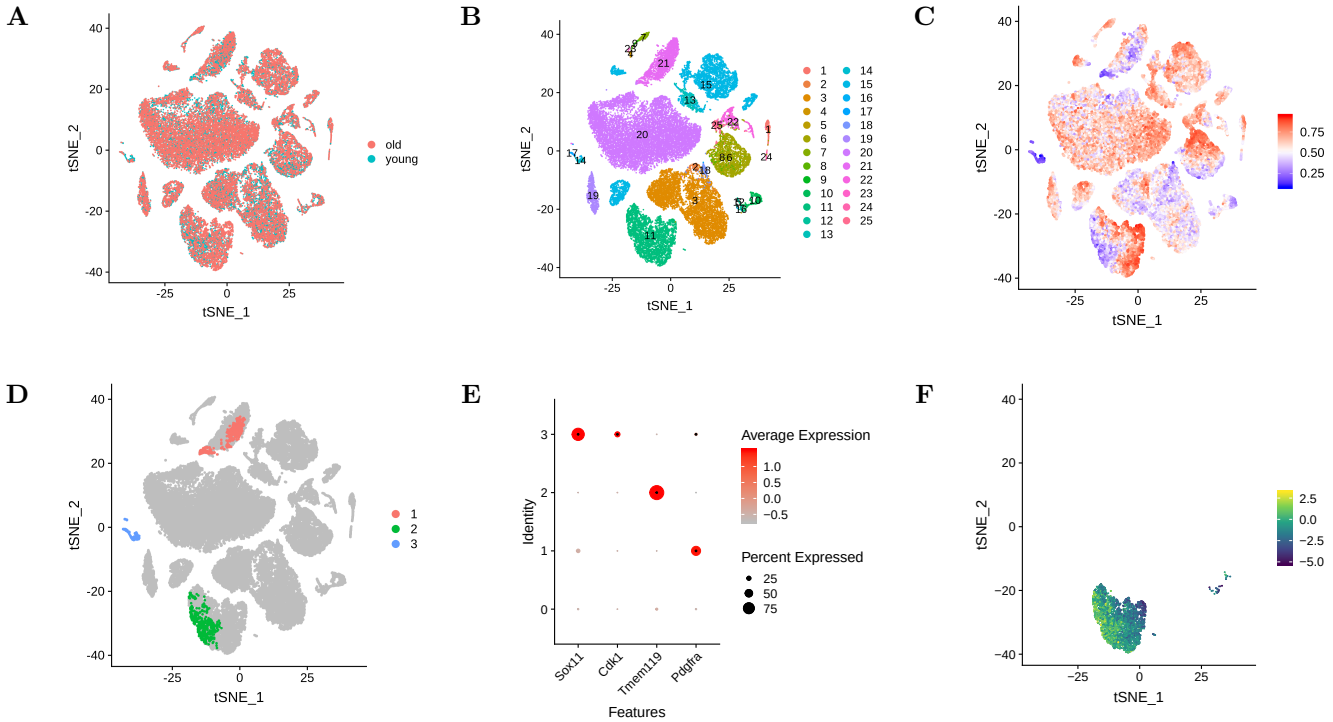

Figure 4. Comparing brain transcriptomics of young and old mice t-SNE plots of 37,069 cells, colored by (A) cells measured in samples from young mice (ceyan) or old mice (red) (B) cluster labels from [39]; (C) the predictions of the logistic classifier. High (low) values are indicative that the cell's neighborhood has excess (deficiency) of cells from old mice relative to cells from young mice; (D) distinct DA regions obtained by clustering cells from the bottom 5% quantile of the predictions in (C); (E) dot plot for markers characterizing the three selected DA regions; details about size, color and non-DA regions information is as in Fig. 2E; (F) cells colored by output of STG differentiating $R2$ and the remaining microglial cells ($C11 - R2$).

marker genes characterizing the clusters detected in [39].

Regions $R1$, $R2$ and $R3$ overlap with the predefined clusters described above. Therefore, the marker genes that characterize each cluster overlap with some of the genes that characterize the corresponding DA region. It is instructive to study the differences between DA regions and the clusters that overlap with them. For instance, we examine below how $R2$ (the subgroup of cells colored in green in Fig. 4D) differs from the remaining cells in the microglia cluster ($C11$ in Fig. 4B). The $Tmem119$ gene differentiates the $R2$ subpopulation from cells in the rest of the clusters but does not distinguish it from the complementary microglial cells (the grey cells that share the same cluster $C11$ as the green cells in Fig. 4D). To better characterize gene sets that separate $R2$ from the remaining microglial cells ($C11 - R2$), we apply STG, our neural network feature selection approach, to differentiate cells from $R2$ and cells from $C11 - R2$. STG identified a total of 115 genes that as a whole differentiate cells in $R2$ from cells in $C11 - R2$. Importantly, the gene combination inferred by STG

Table 3. Statistics of DA regions in aging brain data

Region	DA score	p-value
1	-0.645	0.000155
2	-0.647	0.000622
3	-0.646	0.000310

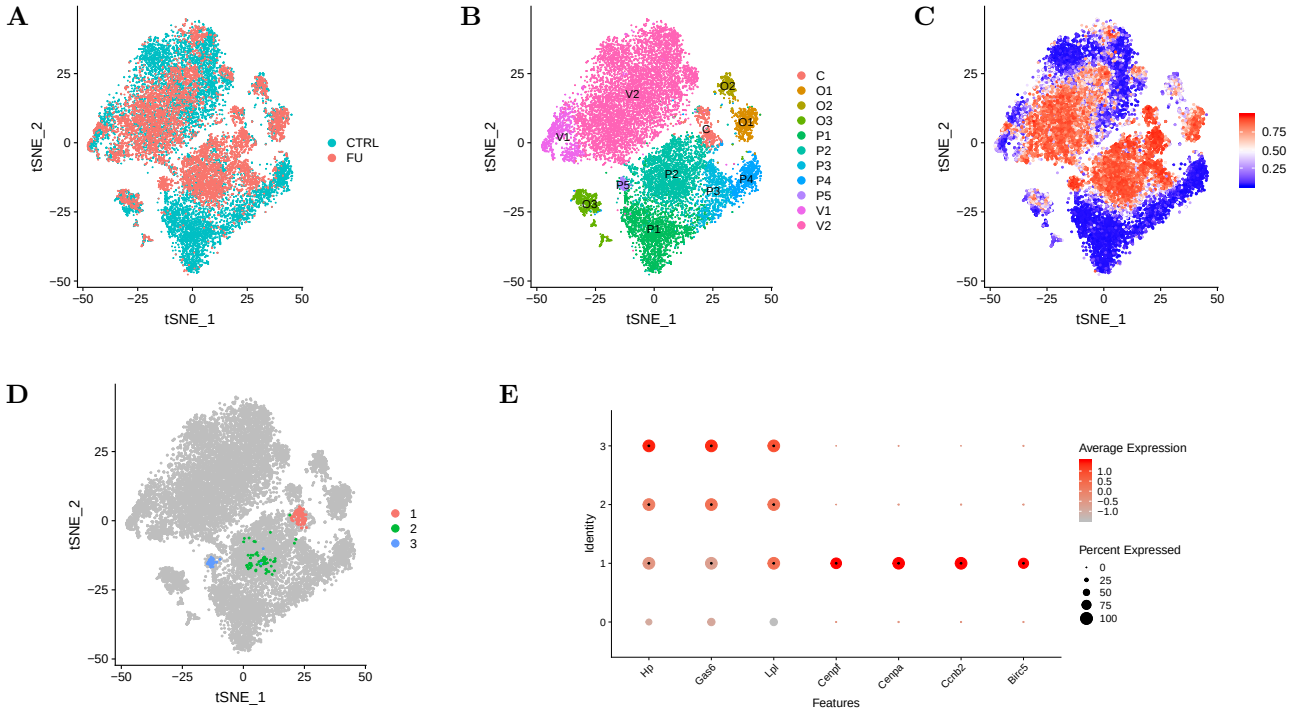

Figure 5. Comparing transcriptomics of bone marrow cells at homeostasis and stress conditions t-SNE plot of 16,291 cells, colored by (A) homeostasis condition (CTRL) or stress condition (FU); (B) cluster labels from [36]; (C) the predictions of the logistic classifier. High (low) values are indicative that the cell's neighborhood has excess (deficiency) of cells from the stress condition relative to cells from the homeostasis condition; (D) distinct DA regions obtained by clustering cells from the top 5% quantiles of the predictions in (C); (E) dot plot for markers characterizing the three selected DA regions; details about size, color and non-DA regions information is as in Fig. 2E

provides a cleaner way to differentiate between the cell subpopulations; this cannot be achieved by individual gene markers. In Fig. 4F, we show the predictions of STG based on the combination of these 115 genes. This specific combination of genes represents a particular gene set that may play an important role in microglial cells during aging.

Cellular landscape change of bone marrow microenvironment at homeostasis and under stress

Tikhonova et al. [36] performed scRNA-seq analysis to investigate the stress response of mouse bone marrow cells, and characterized the regulation of bone marrow microenvironment at homeostasis and under stress conditions. Vascular, perivascular cells and osteoblasts were characterized at homeostasis and under stress-induced (5-FU treatment) haematopoiesis, respectively. Fig. 5A shows the labels for each cell after merging all single cell transcriptomes from both conditions. Cells were clustered and annotated by Tikhonova et al. as displayed in Fig. 5B (with C corresponding to cycling cells, O to osteoblasts, P to perivascular cells and V to vascular cells). Differential abundance between cells from homeostasis and under stress was evaluated for each cluster. They observed that adipocyte-primed clusters expanded in perivascular cells under stress. Specifically, they found a new cell subpopulation (cluster $P5$) and a significant expansion of another cell subpopulation (cluster $P2$) due to treatment with 5-FU. Additionally, they observed an increase in the number of cycling cells; this suggests an increase in cell proliferation under stress.

In [36], cells were first grouped into three major cell types, including vascular cells, perivascular cells and osteoblasts, based on the expression of known cell type markers. Then clustering was performed to further dissect each cell type into smaller subpopulations. Differential abundance between the homeostasis and stress conditions was evaluated by the stress/homeostasis cell ratio for each of these clusters (subpopulations). In our analysis, we merge the single cell trascriptomes from both the stress and homeostasis conditions and apply DA-seq to them. This obviates the need to separate cells into well-characterized cell types based on their known marker genes and the need to further partition these well-characterized cell types into subtypes. We detect three DA regions with a higher abundance of the stress condition (5-FU treatment) as shown in Fig. 5D. These DA regions overlap with the differentially abundant clusters identified in [36]. Specifically, the $R1, R2$ and $R3$ DA regions shown in 5D correspond to the differentially abundant clusters $C, P2, P5$ of 5B, respectively. Characteristic markers expressed in regions $R2$ and $R3$ agree with genes characterizing the corresponding clusters from [36] and are shown in Fig. 5E. Genes related to cell cycle were identified as markers of region $R1$.

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 two samples (e.g., two biological conditions or states). This approach enables us to robustly delineate regions of substantially differential abundance between these two samples. In contrast to existing methods, the subpopulations of cells we discover are not confined to any predefined clusters or cell subtypes. We applied DA-seq to several real scRNA-seq datasets and compared its output to results obtained through conventional methods. DA-seq not only recovers most results obtained by standard approaches but also discovers novel DA regions that could potentially provide valuable biological insights.

Intra-cluster analysis: Practitioners often try to detect intra-cluster differentially expressed genes between two conditions separately for each cluster [7, 31, 36, 39]. If such intra-cluster differentially expressed genes exist, it means that the distributions of cells from these two conditions are shifted with respect to each other and, hence, represent two adjacent DA regions: one enriched by cells from the first condition and the other enriched by cells from the latter. One example is in the comparison between old and young mice shown in Figure 4. Cluster 11 consists of two DA regions, enriched with cells from old and young mouse brains. In this case, intra-cluster analysis will give similar results to differential expression analysis between a DA region (in which all cells have similar transcriptomes) and its immediate neighborhood (in which cells are slightly shifted in the gene space). However, the intra-cluster analysis neither informs us about differential abundance between the states nor is applicable to regions of the distribution that have no cluster-like structure.

Batch effect removal and imputation: It is important to recognize 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: [33], [7] and [15]. Furthermore, imputation for scRNA-seq datasets may also improve downstream analysis and lead to a more accurate differential abundance assessment, as cells are better positioned after imputation [9, 18, 26, 37].

Multi state comparison: DA-seq is focused on comparing two biological states. It is possible, however, to apply DA-seq for multi-state comparisons by considering all pairwise differences in abundance. Alternatively, one can propose a multi-state score vector and replace the binary logistic regression classifier with a multi-class classifier.

Computational complexity: DA-seq is based on computation of a k-nearest neighbor (k-nn) graph, which may be a computational bottleneck for very large datasets. A standard method to compute k-nn graphs is via the application of kd trees [4]. 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 on the order of millions of cells, fast approximate approaches, such as [16, 25], can be applied to increase the scalability of DA-seq.

Direct identification of DA regions in the gene space via neural networks: An alternative approach to detect cells within a DA region is to apply a neural network classifier directly on the input features (genes or PCA coordinates) without computing the score vector in step 1. A network architecture for classification of two clusters often contains a logistic regression as its last layer. The layers preceding the last layer can then be viewed as feature extractors trained in a supervised way. These features may substitute our hand-crafted, multi-scale score-vector features. We conducted preliminary experiments using the full-neural-network approach. The results were comparable to DA-seq for the simulated datasets but inferior for the real-world datasets (results not shown). We conjecture that, for our DA problem, the hand-crafted features allow for a better identification of DA cells because these cells are concentrated in two regions in the score-vector space. On the other hand, the landscape of DA cells in the original gene or PCA space is much more complex. However, it is possible that more sophisticated neural network approaches may outperform DA-seq - especially when a larger number of cell measurements is available.

Application of DA-seq to cells from a single sample: In many biological systems, cell populations could be heterogeneous in terms of the expression status of certain markers. For instance, breast cancer cells from an ER positive patient do not express ER in all her cancer cells. This status can be measured at the transcriptional or translational level. An application of DA-seq to data generated in a single scRNA-seq experiment to compare her ER(+) or ER(-) cancer cells will enable identification of subpopultions of cancer cells enriched by ER(+) or ER(-) cells and, thus, allow exploration of the biological differences between these two populations (beyond their difference in ER status). Essentially, this approach allows us to use cells generated in a single scRNA-seq experiment and compare cells conditioned on the expression status of a single marker. Further potential applications of DA-seq extend to samples in which a proportion of cells are genetically labeled or modified within a single experiment; this would provide a method with which to resolve molecular differences between labeled and unlabeled cells within a population.

DA-seq and cell type identification: In step 4 of DA-seq we characterize each DA region by markers that differentiate it from the remaining cells using our novel neural network embedded feature selection (l_0-based regularization) method. We note that if the system of interest includes cells that have been characterized in earlier studies and documented in cell atlases other methods for cell identification [1] could be utilized to assign cell identification to cells within DA regions. An additional strategy is to correlate each DA subpopulation with a predefined cluster. The most widely-used approach in the community is graph-based clustering, such as shared nearest neighbor (SNN) based clustering implemented in the popular scRNA-seq analysis toolkit Seurat [7]. Alternative unsupervised methods for cell identification can also be utilized to characterize DA regions by overlapping the DA cells and structures revealed by such unsupervised methods [29, 34, 42, 43].

Code availability

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

Acknowledgments

The authors would like to thank Jonathan Levinsohn, Peggy Myung, Manolis Roulis, Shelli Farhadian for many useful discussions; Anastasia Tikhonova and Igor Dolgalev for their help on the bone marrow dataset.

Appendix

DA statistics calculation

For each region detected by DA-seq, we compute a DA-score along with a p-value to indicate its statistical significance . The 'DA score' for a region is computed via,

$$DA\ score = \frac{\frac{x(R)}{n_x} - \frac{y(R)}{n_y}}{\frac{x(R)}{n_x} + \frac{y(R)}{n_y}}. \tag{8}$$

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 [14] ($m_X = 2, m_Y = 2$), 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.

Simulated datasets

In order to test our algorithm with "ground truth", we generated two simulated datasets. The first dataset is based on the scRNA-seq data from [31], and the second on a Gaussian mixture model.

In the first dataset, we used the gene expression profiles from the real data, but assigned labels to cells manually to create artificial DA sites. We selected at random 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. Next, We applied DA-seq to the simulated dataset. The cells with the most DA neighborhood and final DA regions from our algorithm are shown in Fig. S1B and C, respectively. Table S1 shows DA scores and p-values for the four detected DA regions (a positive DA score indicates high abundance of label 2, while a negative DA score indicates high abundance of label 1. As can be seen, we successfully recovered all four artificial DA sites and did not introduce false positive regions.

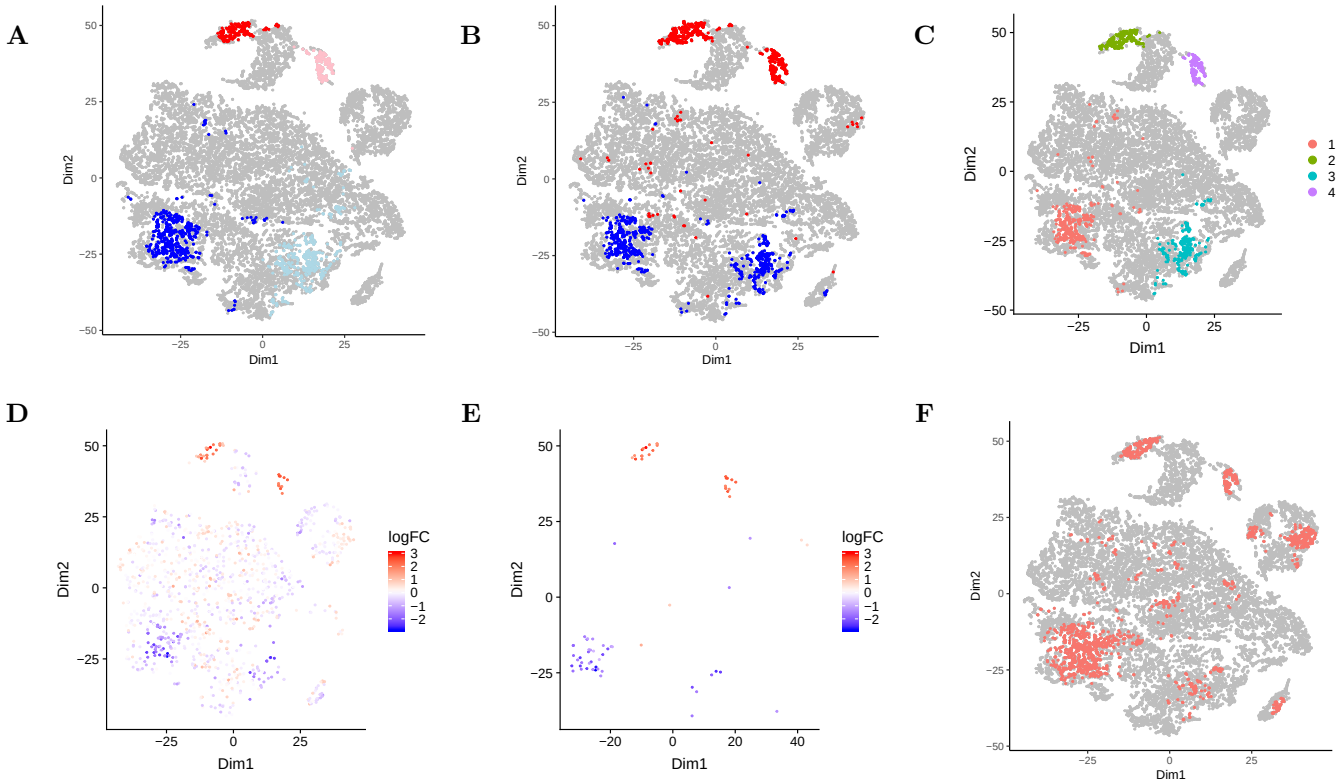

Figure S1. DA results on simulated dataset. t-SNE plot of 16,291 cells: (A) colored cells are four artificial DA sites: two sites on the top have more cells from label 2, while two sites on the bottom have more cells from label 1; (B) the most DA cells selected by our algorithm; (C) DA regions after 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.

Comparison with Cydar

To compare our algorithm with Cydar [27], 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 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 magnitude of 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

Table S1. Statistics of DA regions in simulated data

Region	DA score	p-value
1	-0.694	3.05e-09
2	0.586	5.70e-09
3	-0.762	3.00e-09
4	0.737	3.03e-09

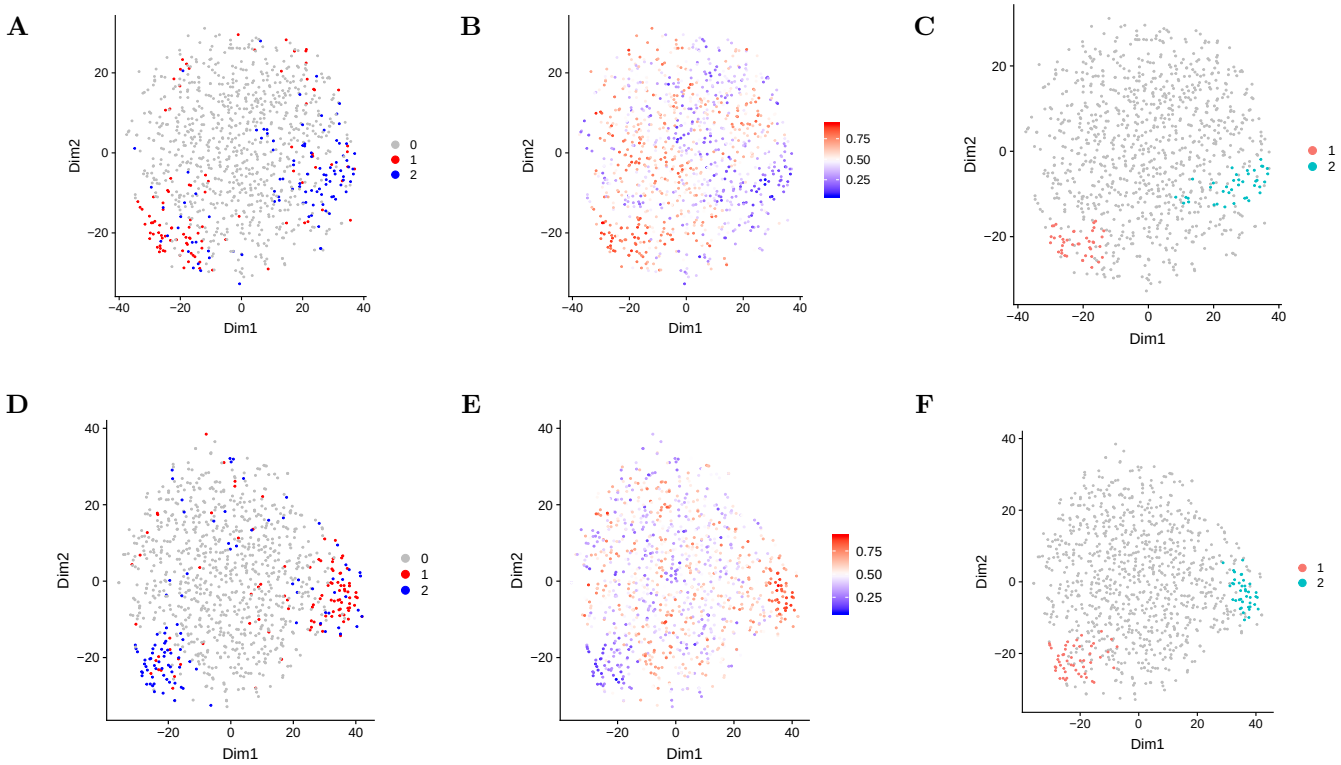

Figure S2. A Gaussian mixture model. t-SNE plot of 1000 cells, panels (A-C) correspond to simulation dataset 1, and (D-F) correspond to simulation dataset 2. In panels (A,D) the red and blue cells form DA regions 1 and 2, respectively, while the grey cells are outside DA region. Panels (B,E) show the prediction of the logistic classifier for every cell. Panels (C,F) color the DA regions detected with DA-seq by clustering the cells in the top and bottom 10% quantiles of the logistic classifier.

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.

Gaussian mixture simulation

We generated two simulated datasets according to a Gaussian mixture model. Let μ_A, μ_B be 10 dimensional vectors, such that each contains 4 non zero elements. (say elements 1-4 for μ_A and 5-8 for μ_B). The 'cells' were generated according to two distribution functions, $f(\boldsymbol{x}|y=A)$ and $f(\boldsymbol{x}|y=B)$ were A and B correspond to two biological conditions

$$f(\boldsymbol{x}|y=A) = \begin{cases} \mathcal{N}(0,I) & \text{w.p. } 0.9 \\ \mathcal{N}(\mu_A,I) & \text{w.p. } 0.08 \\ \mathcal{N}(\mu_B,I) & \text{w.p. } 0.02 \end{cases} \qquad f(\boldsymbol{x}|y=B) = \begin{cases} \mathcal{N}(0,I) & \text{w.p. } 0.9 \\ \mathcal{N}(\mu_A,I) & \text{w.p. } 0.02 \\ \mathcal{N}(\mu_B,I) & \text{w.p. } 0.08 \end{cases}$$

where $\mathcal{N}(0,I)$ denotes a 10 dimensional standard Gaussian distribution. Thus, the first four features are differentially expressed, and form a DA region with high abundance of cells from condition A. The cells in this region are colored in red in Figure S2(A). Similarly, features $5-8$ form a second DA region with high abundance of cells from condition B. The second region is colored in blue in Figure S2(A). Panel B in the figure shows the output of the logistic classifier trained on the score vector, and (C) shows detected

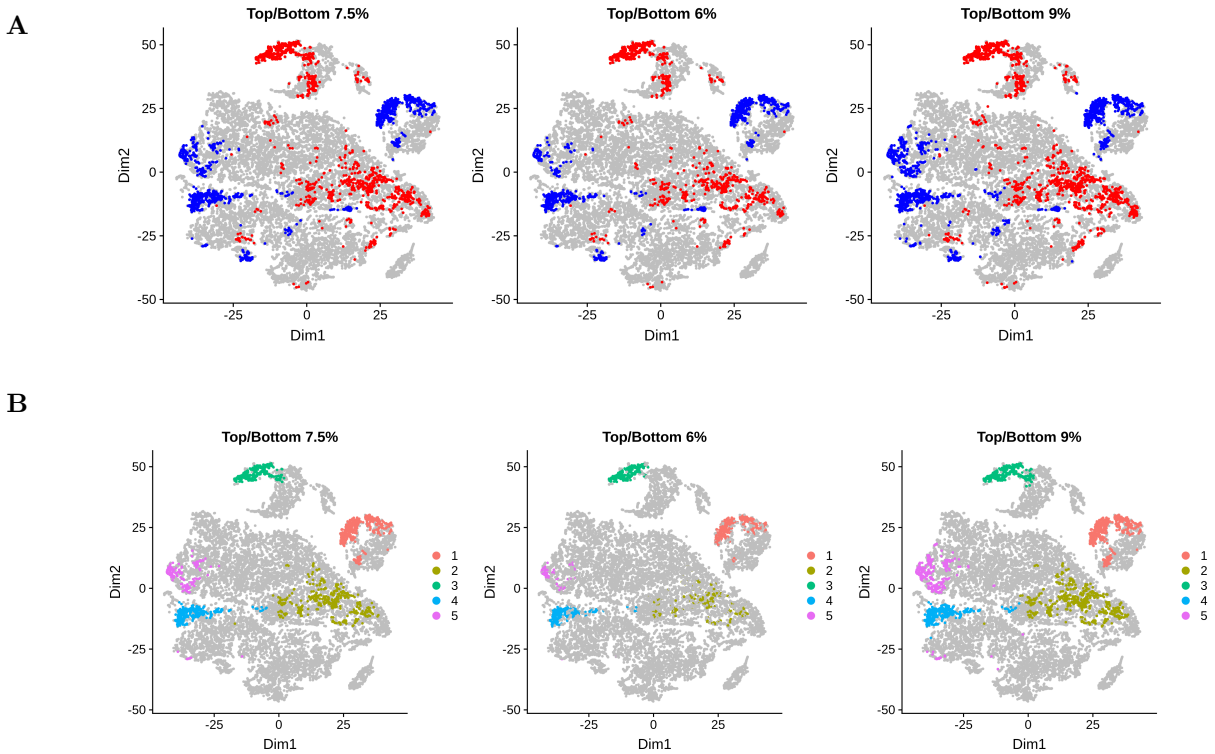

Figure S3. Stability of DA-seq. t-SNE plot of 16,291 cells. (A) Selected DA cells with different threshold settings in Step 2. (B) The corresponding DA regions with fixed parameters in Steps 3-4.

DA regions after clustering cells in the top and bottom 10% quantiles. Applying feature selection through stochastic gates, we were able to recover all the differentiating genes for this dataset.

Our second Gaussian mixtures data was generated similarly to the simulated data above, this time with 2 differentially expressed features that form each of the DA regions. The results of DA-seq for this dataset are shown in Figure S2 panels (D,E,F). For this dataset, we were able to recover all the differentiating genes.

Robustness to changes in thresholds

We exemplify the stability of DA-seq to changes in the total number of DA-cells, determined by the low and high thresholds τ_l, τ_h in Step 2. In other words, we test how a change in τ_l and τ_h affects the outcome of DA-seq: the DA regions, and the list of genes that characterize each of the regions. To that end, we apply DA-seq to the melanoma dataset using three different thresholds in step 2 ($\tau_l = \tau_h = \{6\%,\ 7.5\%,\ 9\%\}$) while keeping the parameters of steps 3-4 fixed.

Fig. S3A shows the selected cells for the three thresholds. Fig. S3B shows the DA regions associated with these thresholds for clustering with $k = 5$. Next, we ranked the genes selected by STG for each region, based on the log fold change between their mean expression within and outside the region. This process was repeated three times for the different thresholds. The agreement between the top n genes across all three thresholds is computed for each of the 5 DA regions and is shown in Table S2, for $n = 10, 20$ and 50.

Table S2. Ratio of common genes across runs with different thresholds

n	1	2	3	4	5
10	0.8	0.8	0.4	0.7	0.7
20	0.8	0.65	0.5	0.75	0.5
50	0.58	0.74	0.44	0.68	0.46

References

1. T. Abdelaal, L. Michielsen, D. Cats, D. Hoogduin, H. Mei, M. J. Reinders, and A. Mahfouz. A comparison of automatic cell identification methods for single-cell rna sequencing data. *Genome biology*, 20(1):194, 2019.

2. R. Basri, M. Galun, A. Geifman, D. Jacobs, Y. Kasten, and S. Kritchman. Frequency bias in neural networks for input of non-uniform density. *arXiv preprint arXiv:2003.04560*, 2020.

3. Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

4. J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

5. S. L. Buchan, T. Manzo, B. Flutter, A. Rogel, N. Edwards, L. Zhang, S. Sivakumaran, S. Ghorashian, B. Carpenter, C. 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.

6. D. B. Burkhardt, J. S. Stanley, A. L. Perdigoto, S. A. Gigante, K. C. Herold, G. Wolf, A. Giraldez, D. van Dijk, and S. Krishnaswamy. Enhancing experimental signals in single-cell rna-sequencing data using graph signal processing. *bioRxiv*, page 532846, 2019.

7. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411, 2018.

8. F. Cazáis and A. Lhéritier. Beyond two-sample-tests: Localizing data discrepancies in high-dimensional spaces. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE, 2015.

9. G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, 10(1):390, 2019.

10. P. Freeman, I. Kim, and A. Lee. Local two-sample testing: a new tool for analysing high-dimensional astronomical data. *Monthly Notices of the Royal Astronomical Society*, 471(3):3273–3282, 2017.

11. H. Fritz, L. A. Garcıa-Escudero, and A. Mayo-Iscar. tclust: An r package for a trimming approach to cluster analysis. *Journal of Statistical Software*, 47(12):1–26, 2012.

12. M. K. Gandhi, E. Lambley, J. Duraiswamy, U. Dua, C. Smith, S. Elliott, D. Gill, P. Marlton, J. Seymour, and R. 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.

13. X. Gao, D. Hu, M. Gogol, and H. Li. Clustermap: compare multiple single cell rna-seq datasets across different experimental conditions. *Bioinformatics*, 2019.

14. K. Gupta, J. Levinsohn, G. Linderman, D. Chen, T. Y. Sun, D. Dong, M. M. Taketo, M. Bosenberg, Y. Kluger, K. 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.

15. L. Haghverdi, A. T. Lun, M. D. Morgan, and J. C. Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421, 2018.

16. K. Hajebi, Y. Abbasi-Yadkori, H. Shahbazi, and H. Zhang. Fast approximate nearest-neighbor search with k-nearest neighbor graph. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

17. H. Harjunpää, M. Llort Asens, C. Guenther, and S. C. Fagerholm. Cell adhesion molecules and their roles and regulation in the immune and tumor microenvironment. *Frontiers in immunology*, 10:1078, 2019.

18. M. Huang, J. Wang, E. Torre, H. Dueck, S. Shaffer, R. Bonasio, J. I. Murray, A. Raj, M. Li, and N. R. Zhang. Saver: gene expression recovery for single-cell rna sequencing. *Nature methods*, 15(7):539, 2018.

19. I. Kim, A. B. Lee, J. Lei, et al. Global and local two-sample tests via regression. *Electronic Journal of Statistics*, 13(2):5253–5305, 2019.

20. J. Kinchen, H. H. Chen, K. Parikh, A. Antanaviciute, M. Jagielowicz, D. Fawkner-Corbett, N. Ashley, L. Cubitt, E. Mellado-Gomez, M. Attar, et al. Structural remodeling of the human colonic mesenchyme in inflammatory bowel disease. *Cell*, 175(2):372–386, 2018.

21. D.-H. Kong, Y. Kim, M. Kim, J. Jang, and S. 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.

22. D. Laehnemann, J. Köster, E. Szcureck, D. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, N. Beerenwinkel, K. R. Campbell, A. Mahfouz, et al. 12 grand challenges in single-cell data science. Technical report, PeerJ Preprints, 2019.

23. M. Li, M. Soltanolkotabi, and S. Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *CoRR*, abs/1903.11680, 2019.

24. K.-Y. Lin, D. Lu, C.-F. Hung, S. Peng, L. Huang, C. Jie, F. Murillo, J. Rowley, Y.-C. Tsai, L. 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.

25. G. C. Linderman, G. Mishne, Y. Kluger, and S. Steinerberger. Randomized near neighbor graphs, giant components, and applications in data science. *arXiv preprint arXiv:1711.04712*, 2017.

26. G. C. Linderman, J. Zhao, and Y. Kluger. Zero-preserving imputation of scrna-seq data using low-rank approximation. *bioRxiv*, page 397588, 2018.

27. A. T. Lun, A. C. Richard, and J. C. Marioni. Testing for differential abundance in mass cytometry data. *Nature methods*, 14(7):707, 2017.

28. E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

29. S. Mallik and Z. Zhao. Multi-objective optimized fuzzy clustering for detecting cell clusters from single-cell expression profiles. *Genes*, 10(8):611, 2019.

30. B. Ronen, D. Jacobs, Y. Kasten, and S. Kritchman. The convergence rate of neural networks for learned functions of different frequencies. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 4761–4771. Curran Associates, Inc., 2019.

31. M. Sade-Feldman, K. Yizhak, S. L. Bjorgaard, J. P. Ray, C. G. de Boer, R. W. Jenkins, D. J. Lieb, J. H. Chen, D. T. Frederick, M. Barzily-Rokni, et al. Defining t cell states associated with response to checkpoint immunotherapy in melanoma. *Cell*, 175(4):998–1013, 2018.

32. M. Schlesinger and G. 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.

33. U. Shaham, K. P. Stanton, J. Zhao, H. Li, K. Raddassi, R. Montgomery, and Y. Kluger. Removal of batch effects using distribution-matching residual networks. *Bioinformatics*, 33(16):2539–2546, 2017.

34. C. Shao and T. Höfer. Robust classification of single-cell transcriptome data by nonnegative matrix factorization. *Bioinformatics*, 33(2):235–242, 2017.

35. R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

36. A. N. Tikhonova, I. Dolgalev, H. Hu, K. K. Sivaraj, E. Hoxha, Á. Cuesta-Domínguez, S. Pinho, I. Akhmetzyanova, J. Gao, M. Witkowski, et al. The bone marrow microenvironment at single-cell resolution. *Nature*, 569(7755):222, 2019.

37. F. Wagner, D. Barkley, and I. Yanai. Accurate denoising of single-cell rna-seq data using unbiased principal component analysis. *BioRxiv*, page 655365, 2019.

38. T. C. Wu. The role of vascular cell adhesion molecule-1 in tumor immune evasion. *Cancer research*, 67(13):6003–6006, 2007.

39. M. Ximerakis, S. L. Lipnick, B. T. Innes, S. K. Simmons, X. Adiconis, D. Dionne, B. A. Mayweather, L. Nguyen, Z. Niziolek, C. Ozek, et al. Single-cell transcriptomic profiling of the aging mouse brain. *Nature neuroscience*, 22(10):1696–1708, 2019.

40. Y. Yamada, O. Lindenbaum, S. Negahban, and Y. Kluger. Feature selection using stochastic gates. *arXiv preprint arXiv:1810.04247*, 2018.

41. G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8:14049, 2017.

42. X. Zhu, T. Ching, X. Pan, S. M. Weissman, and L. Garmire. Detecting heterogeneity in single-cell rna-seq data by non-negative matrix factorization. *PeerJ*, 5:e2888, 2017.

43. X. Zhu, H.-D. Li, Y. Xu, L. Guo, F.-X. Wu, G. Duan, and J. Wang. A hybrid clustering algorithm for identifying cell types from single-cell rna-seq data. *Genes*, 10(2):98, 2019.