Supplementary Information for

Detection of differentially abundant cell subpopulations discriminates biological states in scRNA-seq data

Jun Zhao,Ariel Jaffe,Henry Li,Ofir Lindenbaum,Esen Sefik,Ruaidhrí Jackson,Xiuyuan Cheng,Richard Flavell, and Yuval Kluger

Richard Flavell.
E-mail: richard.flavell@yale.edu

This PDF file includes:

Supplementary text
Figs. S1 to S9 (not allowed for Brief Reports)
Table S1 (not allowed for Brief Reports)
SI References

Supporting Information Text

Supplementary Note 1: 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,

$$DA\text{-}score = \frac{\frac{x(R)}{n_x} - \frac{y(R)}{n_y}}{\frac{x(R)}{n_x} + \frac{y(R)}{n_y}}. \qquad [1]$$

where $x(R), y(R)$ denote the number of cells in the DA region R from sample X and Y respectively; n_x, n_y denote the total number of cells in 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 states (X, Y) and replicated samples in each biological state: $X_1, \ldots, X_{m_X}, Y_1, \ldots, Y_{m_Y}$, 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, we compute the Wilcoxon rank sum test for the vectors $r(X) = \{r(X_1), \ldots, r(X_{m_X})\}$ and $r(Y) = \{r(Y_1), \ldots, r(Y_{m_Y})\}$ to assess whether the difference between them is significant. In cases where m_X and m_Y are small, like in (1) ($m_X = 2, m_Y = 2$), we use a standard two sample t-test instead of Wilcoxon.

Supplementary Note 2: Comparing brain transcriptomic profiles of young and old mice

Ximerakis et al. (2) characterized differences between brain cells of young and old mice (Fig. S5a). As shown in Fig. S5b, they detected 25 distinct cell populations, of which oligodendrocyte precursor cells (1-OPC), neuronal restricted precursors (7-NRP) and immature neurons (8-ImmN) exhibited statistically significant decreases on abundance in old mice.

We applied DA-seq on the merged data and identified several DA subpopulations after retaining cells with significant DA measure, as shown in Fig. S5c,d. DA clusters reported in (2) were identified by DA-seq, specifically, $DA10$ corresponds to 1-OPC, $DA13$ to 7-NRP and 8-ImmN.

DA-seq also identified other DA cell subpopulations. Subpopulations $DA2$ and $DA9$ are two adjacent DA subpopulations and they both overlap with the microglia (21-MG) cluster (Fig. S5d subpanel). However, $DA2$ is enriched with cells from old mouse brains, but in contrast, $DA9$ is enriched with cells from young mouse brains. It is instructive to study the difference between DA cell subpopulations located within the same cluster but have opposite signs of DA measure. We examined the differential expression between cell subpopulations $DA2$ and $DA9$. Characteristic markers for $DA2$ ($DA9$) overlap with shared and microglia specific aging-upregulated (aging-downregulated) genes reported in (2).

To demonstrate the specificity of DA-seq, we compared cell distributions between samples extracted from different young mice (S5e). We verify that DA-seq did not detect any sizable DA subpopulations, as expected. In S5f, we display the sorted DA measure of all cells, both for the actual sample labels (mouse ID) and permuted labels. A negligible fraction of cells (0.6%) were retained as significant DA cells (S5g), and none of them formed sizable DA subpopulations.

Supplementary Note 3: Simulated datasets

In order to test our algorithm with "ground truth" DA subpopulations, we generated two simulated datasets. The first dataset is based on the scRNA-seq data from (3), and the second on a Gaussian mixture model.

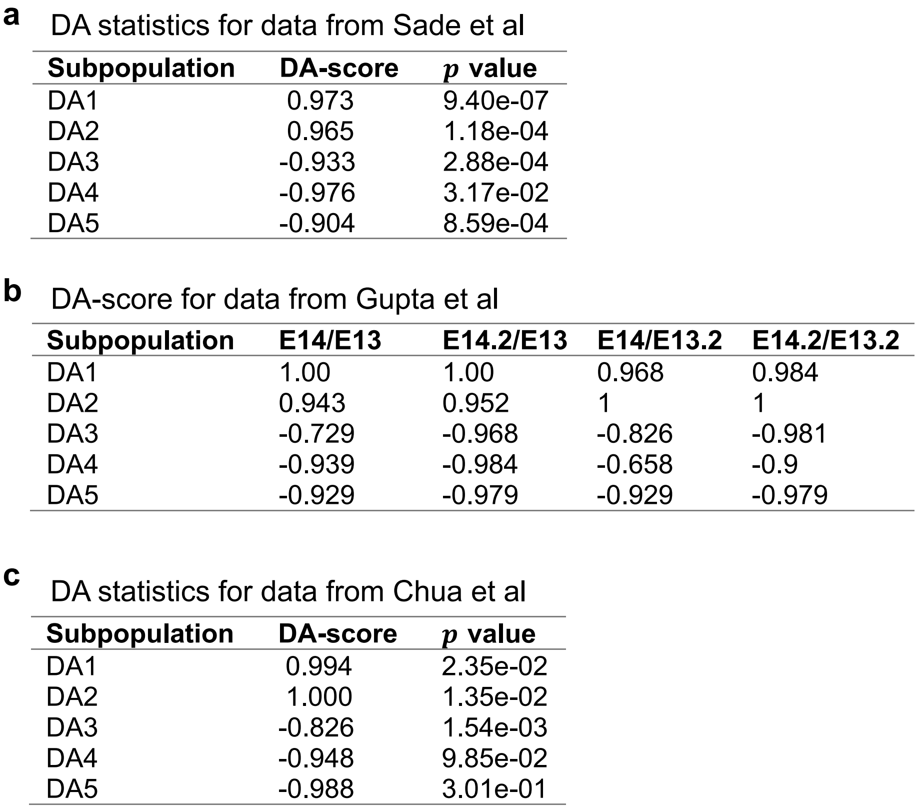
In the first simulation dataset, we used the gene expression profiles from the real data, but assigned labels to cells manually to create artificial DA subpopulations. We selected at random four DA subpopulations based on the kNN graph of the data, as shown in Fig S6a. For cells within the DA subpopulations, imbalanced cells labels were assigned at random, with 90% condition 0, 10% condition 1, or vice versa; labels for cells outside the DA subpopulations were randomly assigned with 50% condition 0 and 50% condition 1 (Fig. S6b). Next, We applied DA-seq to the simulated dataset. The DA measure and final DA subpopulations from our algorithm are shown in Fig. S6c and d, respectively. As can be seen, we successfully recovered all four artificial DA subpopulations and did not introduce false positive subpopulations.

Comparison with Cydar. For comparison, we applied Cydar (4) to the simulated dataset. Briefly, Cydar allocates cells into hyperspheres by randomly selecting a proportion of cells as centers and using a fixed radius. Next, Cydar tests for differential abundance in each hypersphere with a negative binomial model. Since it is hard to evaluate Cydar's performance with just the hyperspheres, we merged all significant hyperspheres (selected based on spatial FDR from Cydar) to obtain DA cells identified by Cydar. In Fig. S6e, receiver operating characteristic (ROC) curves are shown to provide a quantitative comparison. False Positive Rate (FPR) and True Positive Rate (TPR) are calculated by comparing DA cells detected from DA-seq or Cydar to the real DA cells that reside in the artificial DA subpopulations. Cydar was tested with different values of the *tol* parameter, which defines the radius of hyperspheres. 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, especially when the parameter *tol* is small; 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 the second simulation dataset according to a Gaussian mixture model. Let μ_1, μ_2 be 10 dimensional vectors, such that each contains two non zero elements (for instance, elements 1-2 for μ_1 and 3-4 for μ_2). 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 states (Fig. S7a)

$$f(\boldsymbol{x}|y = A) = \begin{cases} \mathcal{N}(0, I) & \text{w.p. } 0.9 \\ \mathcal{N}(\mu_1, I) & \text{w.p. } 0.09 \\ \mathcal{N}(\mu_2, I) & \text{w.p. } 0.01 \end{cases} \qquad f(\boldsymbol{x}|y = B) = \begin{cases} \mathcal{N}(0, I) & \text{w.p. } 0.9 \\ \mathcal{N}(\mu_1, I) & \text{w.p. } 0.01 \\ \mathcal{N}(\mu_2, I) & \text{w.p. } 0.09 \end{cases}$$

where $\mathcal{N}(0, I)$ denotes a 10 dimensional standard Gaussian distribution. Thus, the first two features are differentially expressed, and form a DA subpopulation with high abundance of cells from condition A. Similarly, the other two differentially expressed features form a second DA subpopulation with high abundance of cells from condition B. These two artificial DA subpopulations are highlighted in Fig. S7b. Fig. S7c shows the DA measure, and Fig. S7d shows detected DA subpopulations after clustering significant DA cells. Applying feature selection through STG, we were able to recover all the differentiating features for this dataset (Fig. S7e).



a DA statistics for data from Sade et al

Subpopulation	DA-score	p value
DA1	0.973	9.40e-07
DA2	0.965	1.18e-04
DA3	-0.933	2.88e-04
DA4	-0.976	3.17e-02
DA5	-0.904	8.59e-04

b DA-score for data from Gupta et al

Subpopulation	E14/E13	E14.2/E13	E14/E13.2	E14.2/E13.2
DA1	1.00	1.00	0.968	0.984
DA2	0.943	0.952	1	1
DA3	-0.729	-0.968	-0.826	-0.981
DA4	-0.939	-0.984	-0.658	-0.9
DA5	-0.929	-0.979	-0.929	-0.979

c DA statistics for data from Chua et al

Subpopulation	DA-score	p value
DA1	0.994	2.35e-02
DA2	1.000	1.35e-02
DA3	-0.826	1.54e-03
DA4	-0.948	9.85e-02
DA5	-0.988	3.01e-01

Fig. S1. Statistics of DA subpopulations detected from real scRNA-seq datasets. a Data from Sade-Feldman et al. (3). A DA-score > 0 (< 0) indicates the DA subpopulation is more abundant in samples from non-responders (responders). **b** Data from Gupta et al. (1). A DA-score > 0 (< 0) indicates the DA subpopulation is more abundant in samples from E14.5 (E13.5). **c** Data from Chua et al. (5). A DA-score > 0 (< 0) indicates the DA subpopulation is more abundant in samples from severe (moderate) patients.

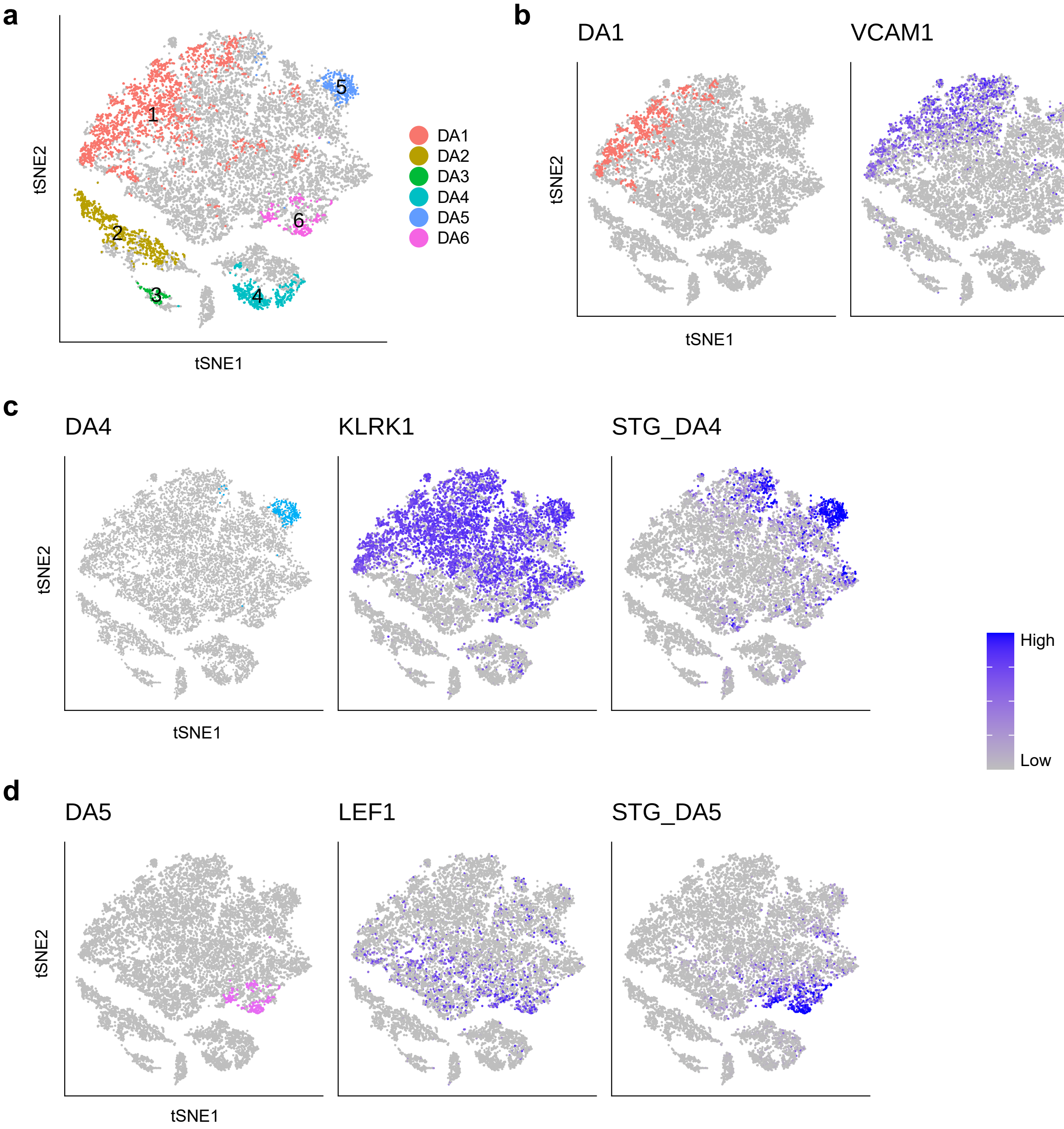


Fig. S2. Characterizing DA subpopulations in data from Sade-Feldman et al. (3). **a** Identified DA subpopulations after relaxing the threshold to $\tau_h = 0.7$. Subpopulation $DA3$ corresponds to the dendritic cell cluster $G4$. **b-d** Marker gene expression and STG prediction score overlay on t-SNE embedding for DA subpopulations. **b** DA1. **c** DA4. **d** DA5.

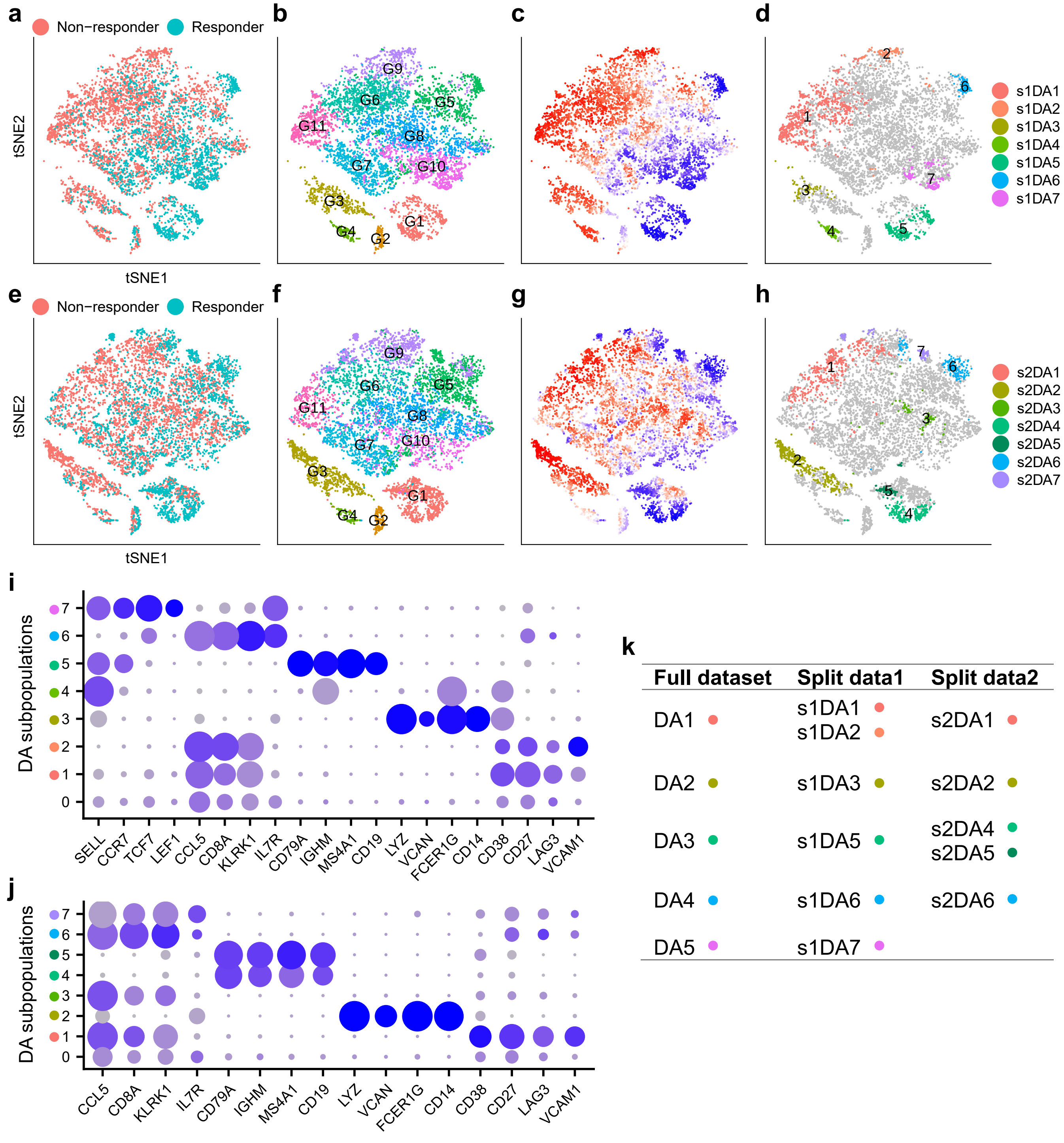


Fig. S3. Cross-validation on data from Sade-Feldman et al. (3). Data was split randomly into two sets, each containing half non-responder samples and half responder samples. **a-d** Show results from split dataset 1 on t-SNE embedding of 7,679 cells. **e-h** Show results from split dataset 2 on t-SNE embedding of 8,612 cells. **a,e** Status of response to immune therapy for each cell. **b,f** Cells colored by cluster labels from (3). **c,g** Cells colored by DA measure. Large/red (small/blue) values indicate a high abundance of cells from the pool of non-responder (responder) samples. **d,h** Distinct DA subpopulations obtained by clustering cells with thresholds $\tau_h = 0.8, \tau_l = -0.8$ on DA measure. **i,j** Dot plots showing marker genes of DA subpopulations. **k** Matched DA subpopulations in the full dataset, split dataset 1 and split dataset 2.

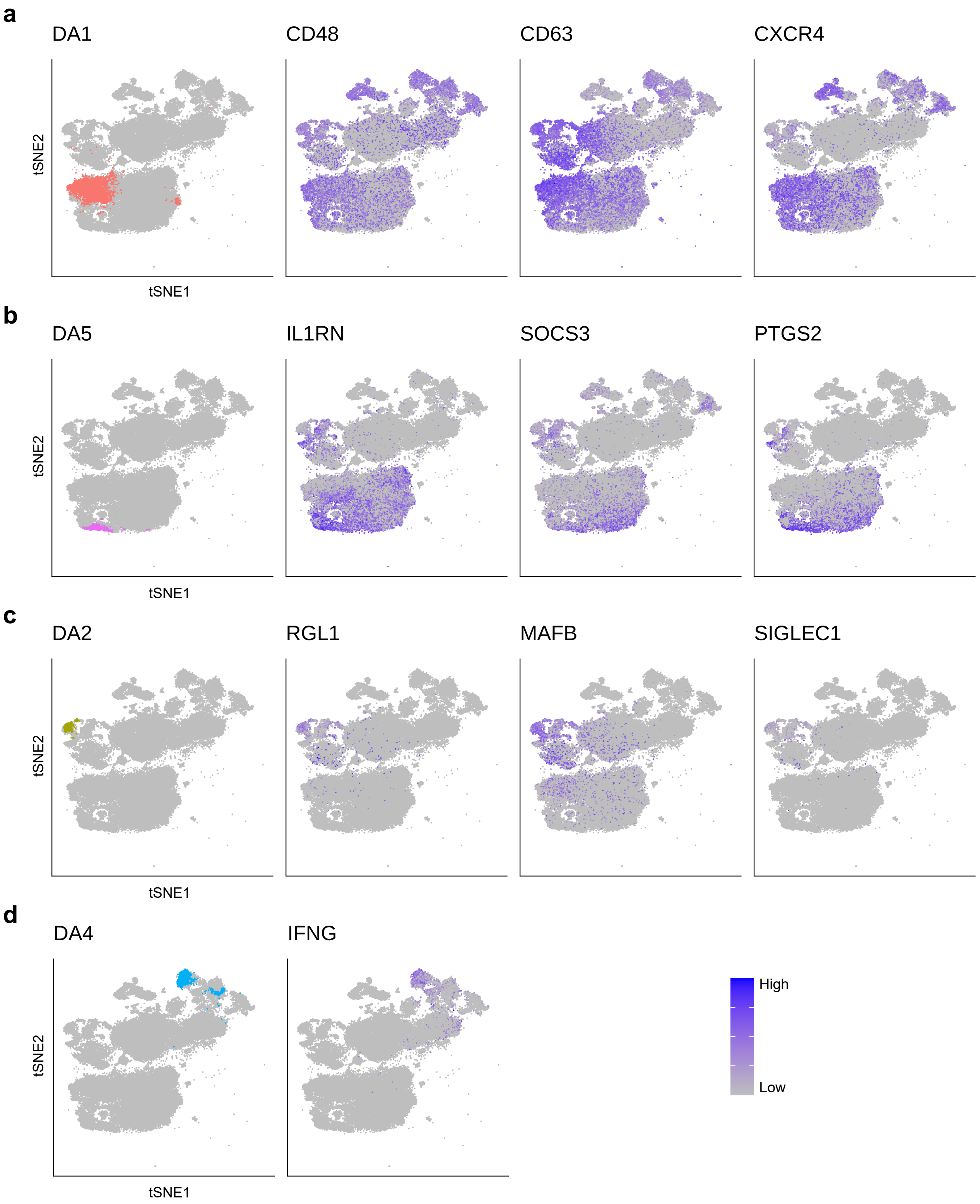


Fig. S4. Characterizing DA subpopulations in data from Chua et al. (5). Marker gene expression overlay on t-SNE embedding for DA subpopulations. **a** DA1. **b** DA5. **c** DA2. **d** DA4.

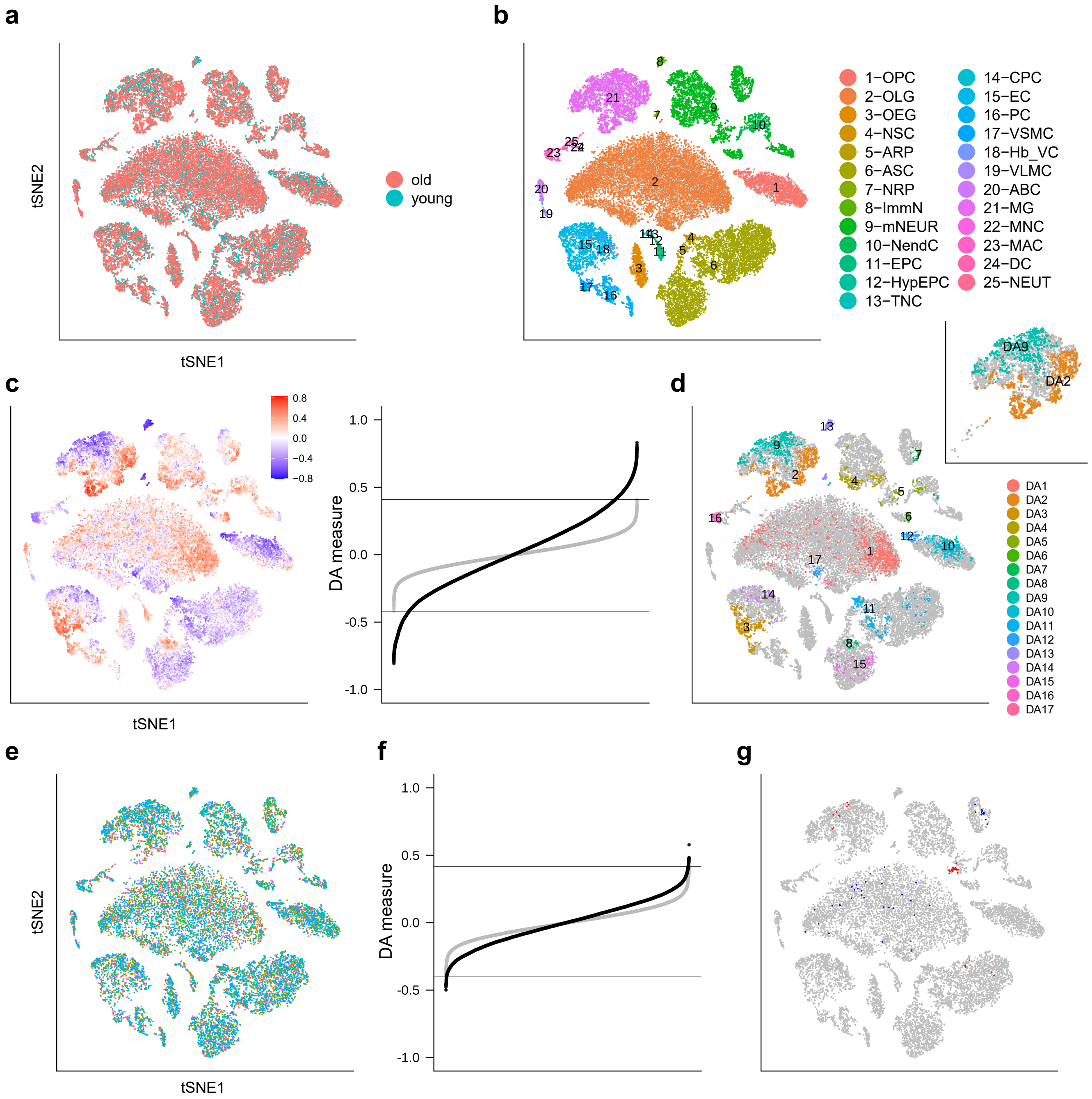


Fig. S5. Comparing brain transcriptomics of young and old mice. a-d t-SNE embedding of 37,069 cells. **a** Cells measured in samples from young mice or old mice. **b** Cluster labels from (2) for each cell. **c** Cells colored by DA measure. Large (small) values indicate a high abundance of cells from old (young) mice. Right panel: DA measure for every cell, ordered monotonically, on real labels (black line) and permuted labels (gray line). **d** Distinct DA subpopulations obtained by clustering cells whose DA measure is above the maximum or below the minimum of permuted DA measure (gray line in **c** right panel). **e** t-SNE embedding of 16,028 cells from eight young mice, cells colored by individual young mice. **f** DA measure for every cell, ordered monotonically, on real labels (black line) and permuted labels (gray line). **g** Highlighted DA cells whose DA measure is above the maximum or below the minimum of permuted DA measure (gray line in **f**).

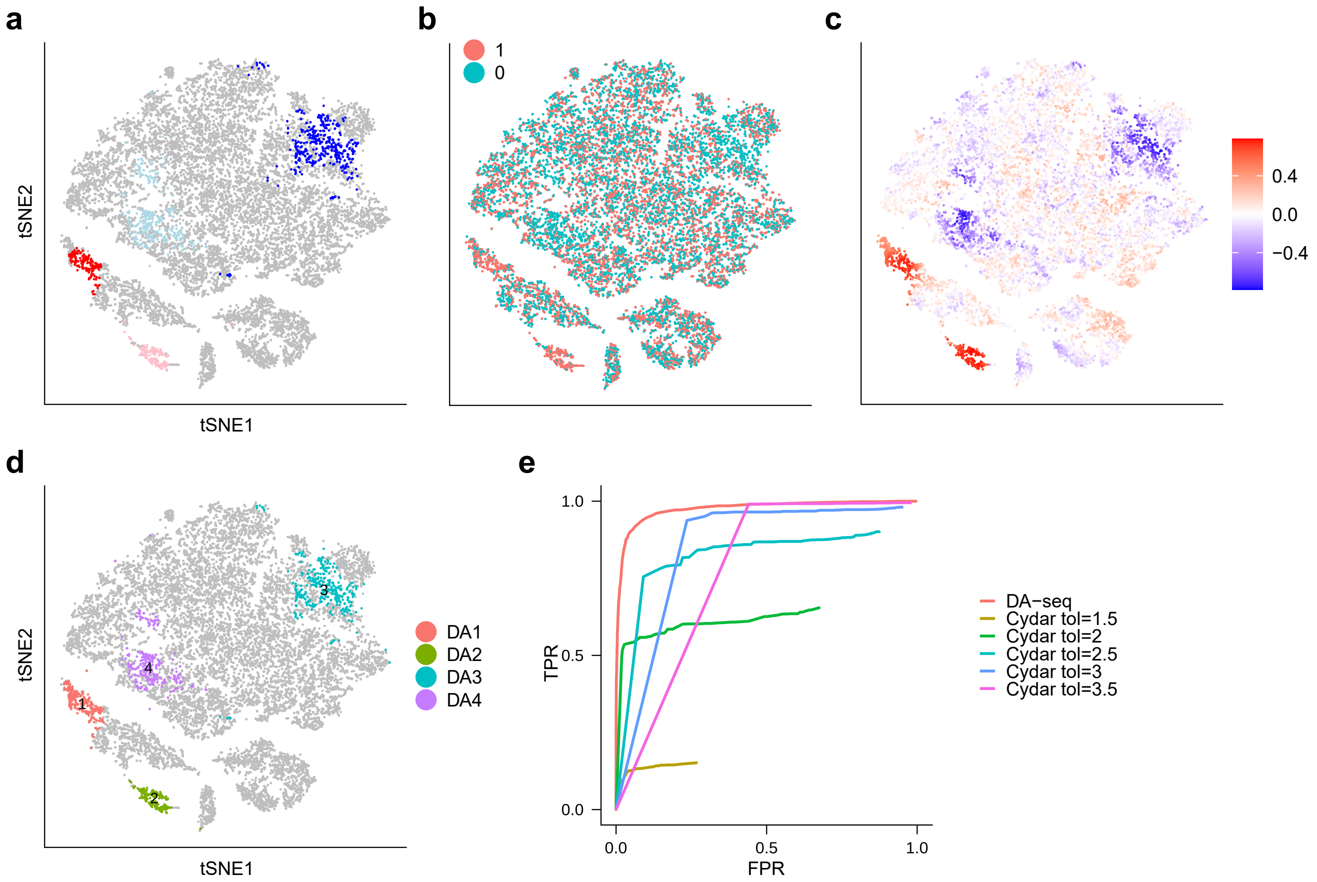


Fig. S6. DA results on simulated dataset. a-d t-SNE embedding of 16,291 cells. **a** Colored cells are four artificial DA subpopulations: two colored with blue hues have more cells from condition 0, while two colored with red hues have more cells from condition 1. **b** Artificial labels for each cell, either condition 0 or condition 1. **c** Cells colored by DA measure. Large (small) values indicate a high abundance of cells from condition 1 (condition 0). **d** Distinct DA subpopulations obtained by clustering cells whose DA measure is above the maximum or below the minimum of permuted DA measure. **e** Receiver operating characteristic (ROC) curves for DA-seq and Cydar results.

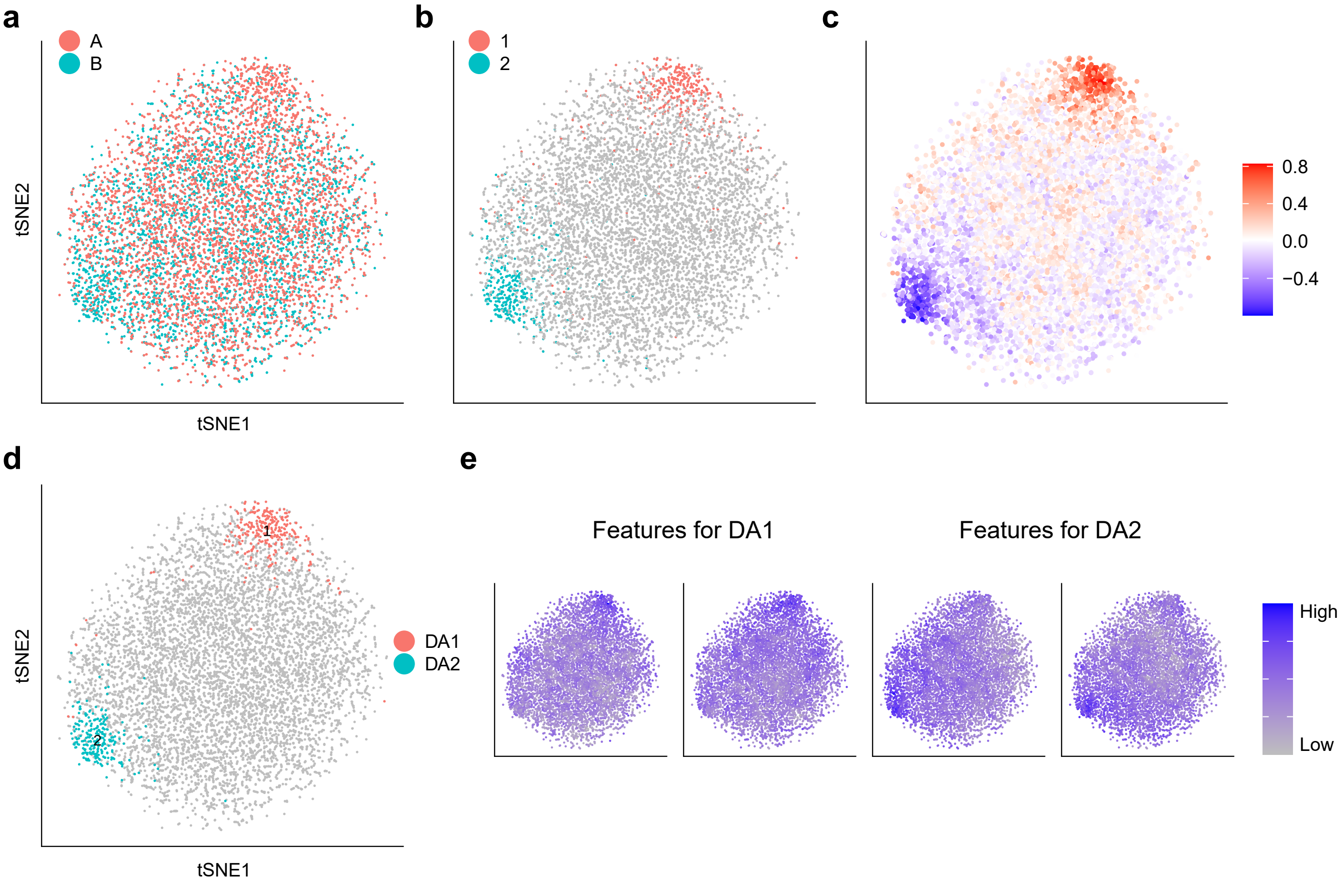


Fig. S7. A Gaussian mixture model. a-d t-SNE embedding of 10,000 cells. **a** Cells colored by its biological condition, A or B. **b** Two artificial DA sites are highlighted. **c** Cells colored by DA measure. Large (small) values indicate a high abundance of cells from condition A (condition B). **d** Distinct DA subpopulations obtained by clustering cells whose DA measure is above the maximum or below the minimum of permuted DA measure. **e** Cells colored by features selected by STG that differentiate the DA subpopulations.

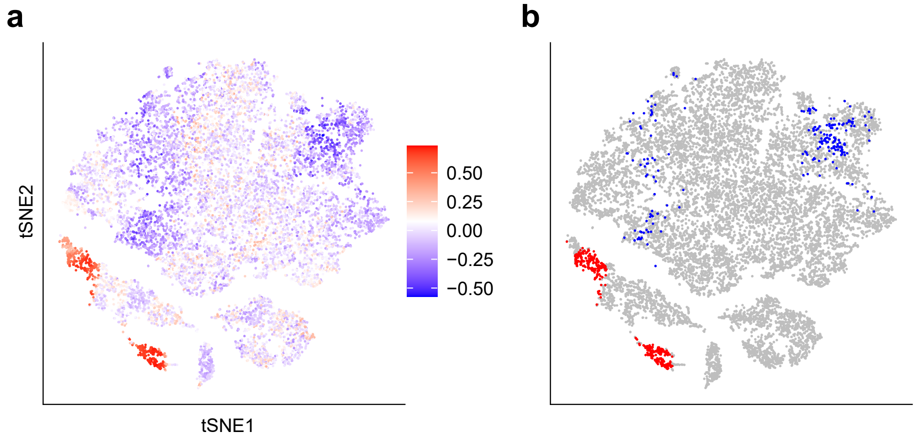


Fig. S8. DA-seq with diffusion distance. t-SNE embedding of 16,291 cells. **a** Cells colored by the DA measure. **b** Highlighted DA cells whose DA measure is above the maximum or below the minimum of permuted DA measure.

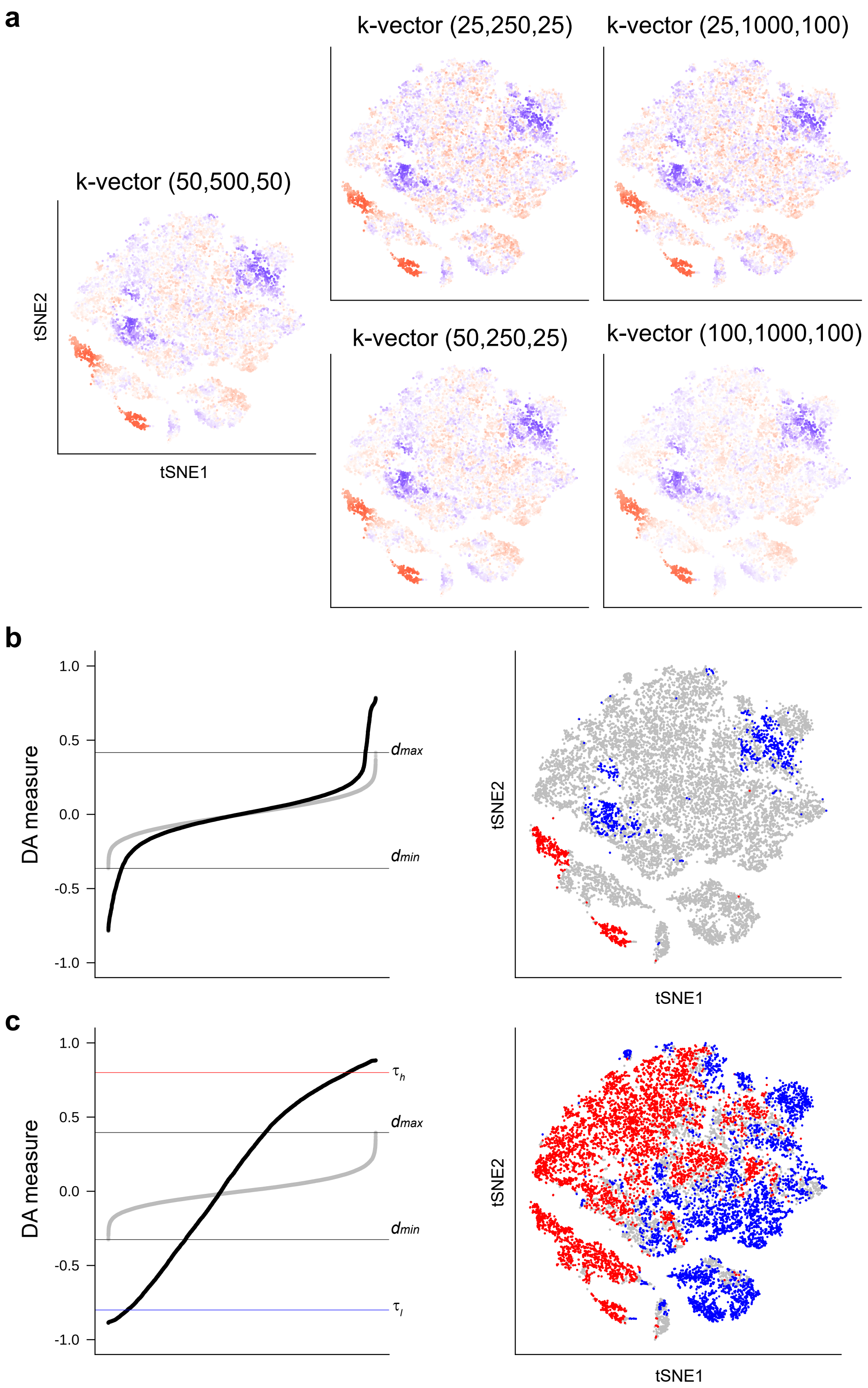
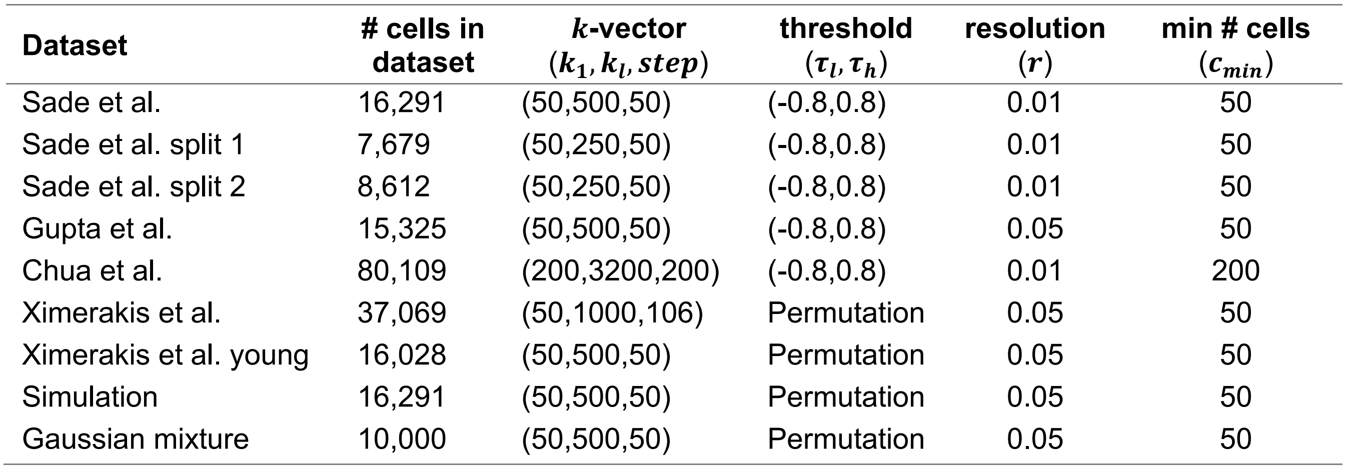


Fig. S9. Choice of parameters in DA-seq. a DA measure of the simulated dataset described in Fig. S6 with different k-vectors. k-vector for each subpanel marked in the format of (k_1, k_l,steps). **b** Results on the simulated dataset described in Fig. S6. Left panel: DA measure for every cell, ordered monotonically, on real labels (black line) and permuted labels (gray line). Black horizontal lines mark significance bounds d_{min}, d_{max}. Right panel: retained DA cells with DA measure above d_{max} or below d_{min}. **c** Results on real scRNA-seq dataset from (3). Two panels structured similar to **b**. Left panel: Black horizontal lines mark significance bounds, red/blue horizontal lines mark the thresholds used in Fig. 2. Right panel: retained cells with DA measure above d_{max} or below d_{min}.



Dataset	# cells in dataset	k-vector $(k_1, k_l, step)$	threshold (τ_l, τ_h)	resolution (r)	min # cells (c_{min})
Sade et al.	16,291	(50,500,50)	(-0.8,0.8)	0.01	50
Sade et al. split 1	7,679	(50,250,50)	(-0.8,0.8)	0.01	50
Sade et al. split 2	8,612	(50,250,50)	(-0.8,0.8)	0.01	50
Gupta et al.	15,325	(50,500,50)	(-0.8,0.8)	0.05	50
Chua et al.	80,109	(200,3200,200)	(-0.8,0.8)	0.01	200
Ximerakis et al.	37,069	(50,1000,106)	Permutation	0.05	50
Ximerakis et al. young	16,028	(50,500,50)	Permutation	0.05	50
Simulation	16,291	(50,500,50)	Permutation	0.05	50
Gaussian mixture	10,000	(50,500,50)	Permutation	0.05	50

Table S1. Parameters used in different datasets.

References

1. K Gupta, et al., Single-cell analysis reveals a hair follicle dermal niche molecular differentiation trajectory that begins prior to morphogenesis. *Dev. cell* **48**, 17–31 (2019).
2. M Ximerakis, et al., Single-cell transcriptomic profiling of the aging mouse brain. *Nat. neuroscience* **22**, 1696–1708 (2019).
3. M Sade-Feldman, et al., Defining t cell states associated with response to checkpoint immunotherapy in melanoma. *Cell* **175**, 998–1013 (2018).
4. AT Lun, AC Richard, JC Marioni, Testing for differential abundance in mass cytometry data. *Nat. methods* **14**, 707 (2017).
5. RL Chua, et al., Covid-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. *Nat. biotechnology* **38**, 970–979 (2020).