Unbiased visualization of single-cell genomic data with SCUBI

Graphical abstract

Highlights
- Cell masking biases the visualization of low-dimensional representations
- SCUBI addresses the masking bias by aggregating information in non-overlapping squares
- Unbalanced cell numbers bias the visual comparisons across samples
- SCUBI enables visualization of cell-type proportion and gene expression without bias

Authors
Wenpin Hou, Zhicheng Ji

Correspondence
zhicheng.ji@duke.edu

In brief
The visualization bias of low-dimensional representations in single-cell genomic data is often overlooked. Hou et al. demonstrate two types of biases leading to data misinterpretation and present SCUBI, a tool that addresses these biases by aggregating information in non-overlapping squares or by controlling the total number of cells across samples.
Unbiased visualization of single-cell genomic data with SCUBI

Wenpin Hou and Zhicheng Ji

1Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
2Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA
3Lead contact

SUMMARY

Visualizing low-dimensional representations with scatterplots is a crucial step in analyzing single-cell genomic data. However, this visualization has significant biases. The first bias arises when visualizing the gene expression levels or the cell identities. The scatterplot only shows a subset of cells plotted last, and the cells plotted earlier are masked and unseen. The second bias arises when comparing the cell-type compositions across samples. The scatterplot is biased by the unbalanced total number of cells across samples. We developed SCUBI, an unbiased method that visualizes the aggregated information of cells within non-overlapping squares to address the first bias and visualizes the differences of cell proportions across samples to address the second bias. We show that SCUBI presents a more faithful visual representation of the information in a real single-cell RNA sequencing (RNA-seq) dataset and has the potential to change how low-dimensional representations are visualized in single-cell genomic data.

INTRODUCTION

Single-cell sequencing measures thousands of features for each cell, which is hard to perceive directly. Dimension reduction techniques such as principal-component analysis (PCA; Wold et al., 1987), t-distributed stochastic neighbor embedding (t-SNE; Van der Maaten and Hinton, 2008), and uniform manifold approximation and projection (UMAP; Becht et al., 2019) thus become crucial to map the cells with high-dimensional information to a low-dimensional space so that the essential information in the dataset can be directly perceived. The low-dimensional representations are then visualized with scatterplots to understand the latent structure of the data, to identify cell subpopulations or developmental trajectories, or to compare across different samples (Butler et al., 2018; Wolf et al., 2018). This visualization procedure has become an essential part of most single-cell genomic studies.

However, we observe significant biases in such visualization procedures that could lead to problematic interpretations of the data in many real applications. Here, we report two major types of biases related to the most basic and commonly used exploratory data analysis procedures. We demonstrate these two biases in a real single-cell RNA sequencing (RNA-seq) dataset and discuss their impact on data interpretations. We introduce a visualization method, SCUBI, to address these biases. We show that SCUBI more faithfully reveals the biological information and can change how the low-dimensional representations of single-cell genomic data are visualized.

RESULTS

The first type of bias arises when visualizing the gene expression levels or cell identities with scatterplots. Since cells are plotted consecutively, cells plotted earlier are masked by cells plotted later and become unseen. Thus, the scatterplot only shows a subset of cells plotted last, which may not represent the information of the entire dataset and may lead to biases (Figure 1A). We

MOTIVATION

With the increased number of cells and samples in single-cell genomic data, visualizing the low-dimensional representations with scatterplots is often biased. Two common types of biases happen either due to cells being masked by other cells or an unbalanced total number of cells when comparing across samples. These biases are often overlooked and rarely reported in the literature but may negatively impact downstream analyses or even lead to misinterpretation of data. We report these two types of biases and present a software tool, SCUBI, which overcomes these biases.

SUMMARY

Visualizing low-dimensional representations with scatterplots is a crucial step in analyzing single-cell genomic data. However, this visualization has significant biases. The first bias arises when visualizing the gene expression levels or the cell identities. The scatterplot only shows a subset of cells plotted last, and the cells plotted earlier are masked and unseen. The second bias arises when comparing the cell-type compositions across samples. The scatterplot is biased by the unbalanced total number of cells across samples. We developed SCUBI, an unbiased method that visualizes the aggregated information of cells within non-overlapping squares to address the first bias and visualizes the differences of cell proportions across samples to address the second bias. We show that SCUBI presents a more faithful visual representation of the information in a real single-cell RNA sequencing (RNA-seq) dataset and has the potential to change how low-dimensional representations are visualized in single-cell genomic data.
demonstrate this bias on a real single-cell RNA-seq dataset where peripheral blood mononuclear cells (PBMCs) were collected from 15 healthy donors and 171 COVID-19 patients with different severity levels (70 with mild diseases, 58 with moderate diseases, and 43 with severe diseases) (Su et al., 2020). To identify the locations of the major cell types in PBMCs, we first used scatterplots to visualize the expression levels of CD3D, CD14, CD19, and CD56, which are marker genes for T cells, monocytes, B cells, and natural killer (NK) cells, respectively (Figure 1B). Compared with the distribution of cells with different strata of expression levels (Figure 1C), the scatterplots fail to fully capture the gene expression information, showing different levels of biases and posing difficulties in locating these four major cell types in PBMCs. The bias is the worst for the CD56 gene, where the cells with high CD56 expression levels can be barely seen in the scatterplot. The stratified view (Figure 1C) also has several
limitations despite its recovery of four major cell types. First, each stratum has only one color and cannot reveal the expression variation across the cells within that stratum. Second, visual comparison across strata increases the difficulty of gaining an overall view of the data. Third, stratified views take more space and may not be suitable to visualize many genes’ expressions simultaneously. We also include an alternative strategy of increasing the transparency levels of the cells, commonly used to visualize dense plots (Figures 1D and S1A). The results are very similar to the original scatterplots (Figure 1B) and still cannot identify the majority of CD56 + NK cells.

SCUBI addresses this first bias by partitioning the coordinate space into small non-overlapping squares. Each square is colored by the averaged gene expression levels of the cells falling in the square (Figure 1E; STAR Methods). This approach aggregates the information of all cells and is not biased toward the cells plotted last on the scatterplot. When applied to the same dataset, SCUBI more faithfully visualizes the gene expression information and marks the distinct locations of the four major cell types in PBMCs indicated by the four marker genes, especially for B cells and NK cells (Figure 1F).

In another discrete case, the scatterplot suffers from the same type of bias and fails to correctly visualize a region where the cells from males are more enriched (Figures 1G and 1H). Used as validation, Figure 1H relies on predefined regions and cannot discover new regions of interest. Plus, increasing the transparency level cannot fully alleviate the visualization bias, and the plot becomes blurry (Figures 1I and S1B). In comparison, SCUBI addresses this bias by visualizing the relative abundances of cells within each square and clearly shows the enrichment of cells from males (Figure 1J; STAR Methods).

Of note, SCUBI visualization is affected by the size of the squares. If the squares are too small, the signals are diluted within each square, and the visualization may appear fragmented. If the squares are too large, the signals are oversmoothed, and the resolution is low (Figures 2, S2A). To balance between oversmoothing and fragmentation, SCUBI calculates the averaged numbers of cells in squares for different square sizes and uses a heuristic elbow method to guide the users to choose the size of the squares (STAR Methods; Figures 2 and S2A). SCUBI allows users to alter the size of the squares based on the actual needs of the analysis.

The second type of bias arises when comparing cell-type compositions across samples to identify enriched or depleted cell types in certain sample groups. Scatterplots confound the cell-type compositions with the total number of cells in each sample, which is often an unwanted technical effect. To illustrate, Figure 3A shows an example comparing cell-type compositions in the COVID-19 dataset, where the marker gene information in Figure 1F is used to infer the cell types (Figure 3C). T/NK cells appear to be more enriched in mild patients, and monocytes appear to be more enriched in moderate/severe patients. However, when we mimic a change in the study design and randomly exclude 90% of cells from mild patients, all cell types

Figure 2. Resolution is chosen with a heuristic elbow method when addressing the first type of bias
Left: elbow plots to guide the choice of resolution factor. Each row corresponds to the full data or data with different proportions of data points subsampled. The elbow points are marked as dotted vertical lines, and their values are displayed. Right: SCUBI visualizations of the CD56 gene with different proportions of data points subsampled (rows) and different choices of resolution factors (columns).
appear to be more enriched in moderate/severe patients (Figure 3B), inconsistent with Figure 3A or the stratified distributions of cells (Figures 3D and 3E). Thus, scatterplots can be hugely biased by the total number of cells and may not reflect the cell-type composition information of interest.

SCUBI addresses the second bias by controlling the total number of cells in each sample. Within each non-overlapping square, it calculates the proportion of cells of each sample falling in the square, fits a linear regression of cell proportions against the sample covariate across samples (disease severity in this case), and visualizes the coefficient of the regression model (Figure 3F). By default, SCUBI smooths the coefficients across squares using loess regression to increase the robustness for squares with cells from a small number of samples. The choice of the square size is guided by the same heuristic elbow method in Figure 2 (Figure 4). We apply SCUBI to study the enriched cell types associated with each disease severity level in the COVID-19 dataset. For example, to identify the enriched cell types in healthy donors, SCUBI uses a binary variable as the covariate, where healthy donors are coded as 1, and other samples are

Figure 3. SCUBI addresses the bias of comparing the cell-type compositions across different types of samples
(A) Scatterplots showing the distribution of cells with different severity levels.
(B) Same as (A) but with 90% of cells from mild COVID-19 patients excluded.
(C) The overall locations of major cell types on UMAP inferred by marker gene information in Figure 1F.
(D) The UMAP space is split into four regions.
(E) Proportion of cells with different severity levels in each region in (D).
(F) Cartoon demonstrating the workflow of SCUBI. To address the bias, SCUBI calculates the proportion of cells falling in each non-overlapping square, fits a linear regression with the sample covariate, and visualizes the regression coefficient.
(G) SCUBI visualizes the comparisons of cell-type compositions across different severity levels with the default resolution factor. Each column shows the cell-type enrichment in one type of samples (e.g., healthy donors) compared to other types of samples (e.g., COVID-19 patients). The enrichment is smoothed using loess regression.
(H) Same as (G) but with 90% of cells from mild COVID-19 patients excluded.
(I) SCUBI visualizes the correlation between S100A8 gene expression levels and disease severity levels with the default resolution factor.
(J) Distributions of S100A8 gene expression levels with different disease severities in the highlighted region in (G).
coded as 0. SCUBI then visualizes the cell enrichment, which is defined as the difference of cell proportions between healthy donors and other samples. The difference is calculated as the fitted regression coefficient for the covariate. With the cell-type information inferred by Figure 1F, SCUBI identifies that T/NK cells are enriched in mild patients, B cells are enriched in moderate patients, and monocytes are enriched in moderate and severe patients (Figure 3G). The pattern is highly consistent after randomly excluding 90% of cells from mild patients (Figure 3H) and agrees with the stratified distributions of cells (Figures 3D and 3E), showing that SCUBI is not confounded by the total number of cells and unbiasedly visualizes the cell-type compositions.

SCUBI also provides a feature to visualize the association of gene expression and the sample covariate of interest. Figure 3I shows an example of the S100A8 gene whose expression is positively correlated with disease severity levels in a monocyte subpopulation. Here, the ordinal variable of disease severity is converted to a numerical variable that takes on values of 1, 2, 3, and 4, corresponding to healthy donors and COVID-19 patients with mild, moderate, and severe symptoms, respectively. The association is confirmed by a stratified view of the gene expression (Figure 3J). The choice of the square size is guided by a different heuristic elbow method based on the proportion of squares that contain cells from samples with at least two different covariate values (STAR Methods; Figure S2B). This is to ensure that the regression is feasible in most of the squares.

The SCUBI framework allows the flexibility of including multiple sample covariates in the linear regression to control for confounding effects such as age and sex.

In addition to visualizing single-cell genomic data, SCUBI can be generally applied to other types of data that require the visualization of the low-dimensional representations with many subjects.

**DISCUSSION**

Visualizing the low-dimensional representations with scatterplots is often the first step toward understanding the single-cell genomic data. We have demonstrated the two common yet often unnoticed biases in such visualizations that lead to problematic interpretations of the data in a real setting. We show that our method, SCUBI, can address these biases and provide an unbiased view of the data, which will benefit the study design and follow-up data analysis and, ultimately, the soundness and reproducibility of the scientific conclusions.

**Limitations of the study**

SCUBI depends on the single-cell genomic data that have been adequately processed by other software tools or pipelines. While data processing does not fall within the scope of this work, improper processing of the data may lead to problematic visualization results. The visualization by SCUBI provides an overview of the data and is useful for designing downstream analysis strategies. However, SCUBI itself does not provide comprehensive...
quantitative analysis results, which should be done by other more dedicated software tools.

**STAR+METHODS**

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - Lead contact
  - Materials availability
  - Data and code availability
- **METHOD DETAILS**
  - SCUBI formulation
  - Assigning cells to non-overlapping squares
  - Tuning size of the squares
  - Mode 1: SCUBI for aggregated gene expression
  - Mode 2: SCUBI for averaged continuous measure
  - Mode 3: SCUBI for discrete categories
  - Mode 4: SCUBI for comparing cell type compositions across samples
  - Mode 5: SCUBI for comparing gene expression across samples
  - Mode 6: SCUBI for comparing continuous measure across samples
  - Single-cell RNA-seq data processing
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

**SUPPLEMENTAL INFORMATION**

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2021.100135.

**ACKNOWLEDGMENTS**

Z.J. is supported by the Whitehead Scholars Program at Duke University School of Medicine. W.H. is supported by the National Institutes of Health grant 1K99HG011468. W.H. would like to acknowledge Dr. Andrew P. Feinberg, Dr. Stephanie C. Hicks, and Dr. Hongkai Ji for their mentorship.

**AUTHOR CONTRIBUTIONS**

Z.J. conceived the study. W.H. and Z.J. developed the SCUBI software package, conducted the analysis, and wrote the manuscript.

**DECLARATION OF INTERESTS**

The authors declare no competing interests.

Received: June 29, 2021
Revised: September 1, 2021
Accepted: December 3, 2021
Published: January 4, 2022

**REFERENCES**

Becht, E., McInnes, L., Healy, J., Dutertre, C.-., Kwok, I.W., Ng, L.G., Ginhoux, F., and Newell, E.W. (2019). Dimensionality reduction for visualizing single-cell data using umap. Nat. Biotechnol. 37, 38–44.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. 36, 411–420.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-y., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. Nat. Methods 16, 1289–1296.

Papadimitriou, C.H., and Steiglitz, K. (1998). Combinatorial Optimization: Algorithms and Complexity (Courier Corporation).

Su, Y., Chen, D., Yuan, D., Lausted, C., Choi, J., Dai, C.L., Voilet, V., Duvvuri, V.R., Scherler, K., Troisch, P., et al. (2020). Multi-omics resolves a sharp disease-state shift between mild and moderate covid-19. Cell 183, 1479–1495.

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-sne. J. Mach. Learn. Res. 9, 2579–2605.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. Chemom. Intell. Lab. Syst. 2, 37–52.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). Scanpy: large-scale single-cell gene expression data analysis. Genome Biol. 19, 1–5.
**STAR METHODS**

**KEY RESOURCES TABLE**

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| Deposited data      |        |            |
| Single-cell RNA-seq datasets of peripheral blood mononuclear cells from COVID-19 patients or healthy donors | https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-9357/ | ArrayExpress: E-MTAB-9357 |
| Software and algorithms | SCUBI R Package | This paper | https://doi.org/10.5281/zenodo.5696986 |

**RESOURCE AVAILABILITY**

**Lead contact**
Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Zhicheng Ji (zhicheng.ji@duke.edu).

**Materials availability**
This study did not generate new materials.

**Data and code availability**
- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- All original code has been deposited at Zenodo repository and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

**METHOD DETAILS**

**SCUBI formulation**
SCUBI visualizes the reduced dimension representations of single-cell genomic data by first assigning cells to non-overlapping squares partitioning the coordinate space, and then aggregating and visualizing the information across cells within each square. SCUBI comes with six modes designed for six different scenarios in single-cell data visualization. The first three modes analyze all cells together without considering the situation where cells may come from multiple samples. The last three modes are designed for multi-sample single-cell datasets and compare cells across different samples. They accept both numerical and categorical variables as sample covariates. The categorical variables can be analyzed in two ways. If the variable is ordinal, it can be converted to a numerical variable in the regression model. Alternatively, SCUBI will generate one plot for each category, comparing the samples from that category with all other samples.

**Assigning cells to non-overlapping squares**
For a single-cell genomic dataset with \( N \) cells, denote \( x_n \) and \( y_n \) as the coordinates of the \( n \)-th cell \( (n = 1, 2, \ldots, N) \). \( x_n \) and \( y_n \) correspond to the two dimensions to be visualized, such as the first two dimensions of PCA, t-SNE, or UMAP. SCUBI assigns the \( n \)-th cell to square \( b_{jk} \), where \( j = \lfloor x_n \times a \rfloor \), \( k = \lfloor y_n \times a \rfloor \), and \( j \) and \( k \) are both integers. Here \( \lfloor \cdot \rfloor \) means rounding to the nearest integer and \( a \) is a resolution factor. By default, \( a = 10 \times 20 / \max(R_1, R_2) \), where \( R_1 \) and \( R_2 \) are the ranges of the two reduced dimensions. This design deals with the potentially different scales of the reduced dimensions. For ordinary UMAPs with ranges of the two dimensions around 20, the default factor corresponds to squares with widths and heights of around 0.1. \( b_{jk} \) is centered at \( \left( \frac{j}{a}, \frac{k}{a} \right) \) and has both length and width of \( \frac{1}{a} \), thus not overlapping with other squares. Denote \( c_{jk} \) as the set of the indices of the cells assigned to \( b_{jk} \).

**Tuning size of the squares**
The default choice of \( a \) works well in the example and in most cases. For users who wish to fine tune the resolution, SCUBI uses a heuristic elbow method to guide users to choose the resolution factor \( a \), thus the size of squares. For Mode 1-4, SCUBI iterates...
through a vector of $a$: $1 \times 20 / \max(R_1, R_2), 2 \times 20 / \max(R_1, R_2), \ldots, 50 \times 20 / \max(R_1, R_2)$, and calculates the averaged number of cells falling in squares for each $a$. SCUBI uses a scatterplot to visualize how the averaged numbers of cells change with different choices of $a$, and users will visually identify an elbow point as the recommended choice of $a$ (Figures 2, 4, and S2A). For Mode 5-6, SCUBI iterates through the same vector of $a$. For each choice of $a$, it calculates the proportion of squares which contain cells from samples with at least two different covariate values, thus feasible to perform regression (Figure S2B).

**Mode 1: SCUBI for aggregated gene expression**

This mode visualizes the aggregated expression of a single gene or a set of genes. It can be used to identify groups of cells with high or low expression levels of one or multiple marker genes.

For $n$-th cell, denote $l_n$ as the total number of reads (library size) and $e_{gn}$ as the number of reads for gene $g$, respectively. For gene set $G$ (may have one or multiple genes) in square $b_{ij}$, SCUBI calculates the aggregated log-scaled normalized expression using a pseudobulk approach: $\log_2 \left( C \times \frac{\sum_{g \in G} \sum_{n \in \text{cells}} e_{gn}^{ps}}{\sum_{n \in \text{cells}} l_n} + 1 \right)$ where $C$ is a constant determined by the library size of the data ($10^5$ by default). Squares are then drawn on the coordinate space and colored by the normalized gene expression values.

**Mode 2: SCUBI for averaged continuous measure**

This mode visualizes the arithmetic mean of a continuous measure of a cell, such as the total number of reads or the proportion of reads mapped to the mitochondrial genome. It can be used to identify groups of cells with high or low values of the continuous measure.

For each square, SCUBI calculates the averaged value of the continuous measure for cells assigned to the square. Squares are then drawn on the coordinate space and colored by the averaged values.

**Mode 3: SCUBI for discrete categories**

This mode visualizes the distribution of cells with discrete categories, such as the cell types, cell clusters, or the types of samples from which the cells were collected. It can be used to identify the location of a cell cluster, or regions where cells from certain samples are enriched.

The visualization of this mode is designed to resemble the original data. For a square that contains cells from a single category, the color corresponding to that category is assigned to the square. For a square that contains a total of $k$ cells from more than one categories, the square is divided into multiple sub-squares with equal sizes. Specifically, denote $f(k)$ as the largest factor of $k$ that is smaller than or equal to $\sqrt{k}$, and let $l = \arg\max_{f(l)} f(n), n = k, k + 1, k + 2$. The square will be split into $f(l)$ rows and $f(l)$ columns, resulting in $l$ sub-squares. Each cell in the square will be assigned to a nearby sub-square by solving a linear sum assignment problem (Papadimitriou and Steiglitz, 1998) (using the solve_LSAP function in R’s clue package), so that each sub-square receives at most one cell. Each sub-square with a cell will then be assigned the color corresponding to the category of the cell assigned to that sub-square.

**Mode 4: SCUBI for comparing cell type compositions across samples**

This mode compares the cell type compositions across different types of samples, such as samples with different levels of disease severity. It can be used to identify the regions where cells from a certain type of samples are enriched. This mode can deal with a continuous covariate, such as identifying enriched cell types with increasing age. Additional sample covariates can be included as well to control for potential confounding effects.

For sample $s$, denote $p_{b,s} = \frac{d_{s,b}}{\text{deg}}$ as the proportion of the cells assigned to the square $b$. Here $d_{s,b}$ is the number of cells in sample $s$ assigned to square $b$, and $\text{deg}$ is the total number of cells in sample $s$. In each square, a linear regression is fitted where the response variable is the proportion $p_{b,s}$, and the independent variables are the covariate of interest as well as other potential confounders to be controlled for. The coefficient associated with the covariate of interest is obtained. Squares are then drawn on the coordinate space and colored by the coefficients of the linear regression models. By default, the coefficients are smoothed with loess regression (base::loess() in R).

**Mode 5: SCUBI for comparing gene expression across samples**

This mode compares the aggregated expression of a single gene or a set of genes across different types of samples. It can be used to identify the regions where gene expression is positively or negatively correlated with the covariate of interest. It can also deal with a continuous covariate or include additional covariates.

Similar to Mode 1, a pseudobulk approach is used to calculate the aggregated normalized expression for the gene set in each square and each sample. In each square, a linear regression is fitted where the response variables is the aggregated normalized gene expression levels and the independent variable are the covariate of interest as well as other potential confounders. Squares are then drawn on the coordinate space and colored by the coefficients of the linear regression models, and the coefficients are smoothed with loess by default.
**Mode 6: SCUBI for comparing continuous measure across samples**

This mode is similar to Mode 5 except it compares a continuous measure of cells across different types of samples. It can be used to identify the regions where the continuous measure is positively or negatively correlated with the covariate of interest.

For each square and for each sample, the continuous measure is averaged across all cells assigned to the square. A linear regression is then fitted in each square where the response variable is the averaged continuous measure, and the regression coefficients are visualized subsequently.

**Single-cell RNA-seq data processing**

Single-cell RNA-seq raw read counts of COVID-19 peripheral blood mononuclear cells (PBMCs) were downloaded from a recent study (Su et al., 2020). We filtered out cells with less than 500 genes expressed, less than 2000 reads, or more than 10% reads mapped to mitochondrial genome. Samples with less than 500 retained cells were excluded. Genes that are expressed in at least 1% of all cells were retained. Mitochondrial genes were removed. Harmony (Korsunsky et al., 2019) with default parameters was performed to integrate cells from different samples. A Seurat (Butler et al., 2018) pipeline with default parameters was performed for dimension reduction and cell clustering.

**QUANTIFICATION AND STATISTICAL ANALYSIS**

Sample size of the single-cell RNA-seq dataset is described in the Results section. Scatterplot were generated using the R statistical software.