OXFORD

*Briefings in Bioinformatics*, 2024, **25(6)**, bbae494
https://doi.org/10.1093/bib/bbae494
**Problem Solving Protocol**

# Detecting expressed genes in cell populations at the single-cell level with scGeneXpress

Sascha Jung[1], Céline Barlier[2], Aitor Martinez Perez[1], Antonio del Sol[1,2,3,*]

[1]Computational Biology Group, CIC bioGUNE-BRTA (Basque Research and Technology Alliance), Parque Cientifico Tecnologico de Bizkaia 801A, 48160 Derio, Spain
[2]Computational Biology Group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 avenue du Swing, L-4367 Esch-sur Alzette, Luxembourg
[3]Ikerbasque, Basque Foundation for Science, Plaza Euskadi 5, 48009 Bilbao, Bizkaia, Spain
*Corresponding author. Computational Biology Group, CIC bioGUNE-BRTA (Basque Research and Technology Alliance), Parque Cientifico Tecnologico de Bizkaia 801A, 48160 Derio, Spain. Tel: (+352) 46 66 44 6982; E-mail: antonio.delsol@uni.lu

## Abstract

Determining whether genes are expressed or not remains a challenge in single-cell RNAseq experiments due to their different expression spectra, which are influenced by genetics, the microenvironment and gene length. Current approaches for addressing this issue fail to provide a comprehensive landscape of expressed genes, since they neglect the inherent differences in the expression ranges and distributions of genes. Here, we present scGeneXpress, a method for detecting expressed genes in cell populations of single-cell RNAseq samples based on gene-specific reference distributions. We demonstrate that scGeneXpress accurately detects expressed cell markers and identity genes in 34 human and mouse tissues and can be employed to improve differential expression analysis of single-cell RNAseq data.

**Keywords**: single cell RNAseq; gene expression; discretization; bioinformatics; cell identity; cell markers

## Introduction

Multicellular organisms are composed of diverse cellular phenotypes displaying different morphologies and specialized functions, which are controlled by specific gene expression programs. In this regard, the advent of single-cell RNAseq (scRNAseq) technologies has offered an unprecedented view on the cellular phenotypes that led to the generation of large-scale cellular atlases [1, 2]. Characterizing the observed expression programs, including the detection of expressed genes, is crucial to understand which genes play critical roles in cell type specific cellular processes, disease progression, or development. However, the detection of expressed genes is impeded by the different expression ranges of each gene, which is influenced by the genetic composition of their regulatory regions such as promoters and enhancers, the extracellular cues received from the microenvironment as well as their lengths [3–5].

Currently, two types of methodologies are employed to detect expressed genes in scRNAseq data, namely absolute and comparative approaches. Absolute approaches consider all genes to be expressed whose transcript count is above a fixed threshold. Although the choice of this threshold is arbitrary, most studies set it to one transcript in lieu of sound alternatives [6, 7]. Moreover, for a gene to be expressed in a cell population, it has to be expressed in a predefined fraction of cells. As a result, absolute approaches neglect the specific expression ranges of each gene and is in most cases too restrictive or too permissive in the determination of expressed genes. On the other hand, in the absence of specific absolute thresholds, researchers revert to comparative approaches to identify significant changes in expression with respect to a baseline [8, 9]. In the context of scRNA-seq data, comparison is carried using differential expression analysis where the baseline and consequently the detected expressed genes depend on what cell populations are compared, which limits the interpretability of the results. Moreover, when comparing to two or more populations, a pooling approach is typically followed in which a population of interest is contrasted with a baseline composed of all other cells in the dataset. However, this renders the results to be dependent on the fractions of each population in the sample, which does not necessarily resemble the true composition in the tissue.

To address this issue, we developed scGeneXpress, a computational method for detecting expressed genes in cell populations of scRNAseq samples. In contrast to existing approaches, scGeneXpress detects expressed genes in a statistical framework based on gene-specific reference distributions to derive individual thresholds above/below which a gene can be considered to be expressed/not expressed. As a result, scGeneXpress is, to our knowledge, the first method that provides gene-specific thresholds for detecting whether a gene is expressed or not, without the need for comparing to other cell populations in a scRNAseq sample. We demonstrate that the detected expressed genes are biologically meaningful and that leveraging this statistical framework improves the recovery of known marker as well as cell identity genes across 73 human and 32 mouse cell types from Tabula Sapiens and Tabula Muris [1, 2]. Moreover, we show that scGeneXpress can be employed to more accurately detect

**Received:** May 8, 2024. **Revised:** August 8, 2024. **Accepted:** September 20, 2024

differentially expressed genes when compared to widely-used traditional approaches. In summary, we demonstrate that scGeneXpress is a versatile tool for detecting expressed genes in scRNAseq datasets that overcomes the limitations of current approaches.

# Materials and Methods

## scGeneXpress workflow

We developed scGeneXpress to detect expressed genes and quantify gene expression in low, medium and high levels. The quantification strategy is based on RefBool [10] and was adapted for single-cell UMI data. scGeneXpress takes as an input the single cell UMI matrix of a cell population and a reference dataset to estimate the expression thresholds for a gene to be expressed. In particular, the quantification strategy is composed of two main parts: (i) the construction of threshold distributions for each gene and (ii) the gene quantification of a query cell population. In fact, the construction of threshold distributions is optional as we built pre-compiled backgrounds for mouse and human in this study to detect expressed genes in a query cell population.

## Constructing threshold distributions for each gene

The construction of threshold distributions for each gene constitutes the initial step of scGeneXpress. These distributions are later used to quantify gene expression of a query cell population. However, the first step of scGeneXpress is to remove all genes with non-zero UMI counts in less than 10 cells as well as low quality cells using the strategy provided in the Scuttle R package. In addition, cell populations with less than 50 cells are removed. Subsequently, the single cell matrix is normalized using scTransform [11, 12]. In fact, normalization factors used for each gene are saved and re-used to normalize query samples. After normalization, the expression of each gene in the reference dataset is scaled to the unit interval. Then, lower and upper-threshold distributions were computed using optimization functions combined with a bootstrap approach. In particular, scGeneXpress performs the following steps:

1) **Bootstrapping.** Given a normalized reference expression dataset $G \in \mathbb{R}^{n \times m}$ with $n$ genes and $m$ cells, scGeneXpress creates for each gene $G_{i,\cdot}$ a set of 1000 random samples $G_i^{rand} = \{g' = c(g'_1, \ldots, g'_{100}) \,|\, \forall_{j \in [1,..,100]} g'_j \in G_{i,\cdot} \wedge g'_j \neq 0\}$. These samples will be used in the following steps to assess the variability of the derived gene expression thresholds for considering a gene to be expressed or not expressed, respectively.
2) **Creation of empirical distribution functions.** For each random sample $g' \in G_i^{rand}$ of gene $G_{i,\cdot}$, an empirical cumulative distribution function is created ($ecdf_{g'}$). It is important to note that $ecdf_{g'}(1) = 1$ and $\forall_{x<1} ecdf_{g'}(x) < 1$, since all genes are scaled to the unit interval.
3) **Identification of lower and upper expression thresholds.** scGeneXpress identifies thresholds for a gene to be in one of three expression levels, namely not expressed (denoted '0'), mediumly expressed ('0.5') and highly expressed ('1'). For that, two thresholds are being derived to separate not expressed and mediumly expressed genes as well as mediumly expressed and highly expressed genes, respectively. For the sake of brevity, we refer to these thresholds as 'lower' and 'upper' expression thresholds. For a gene $G_{i,\cdot}$, both thresholds are derived for all $g' \in G_i^{rand}$ by solving the following optimization problems:

$$tr_{low} = \underset{0 \leq x \leq 1}{\mathbf{argmax}} \left(x \cdot \left(1 - ecdf_{g'}(x)\right)\right)$$

$$tr_{high} = \underset{0 \leq x \leq 1}{\mathbf{argmax}} \left((1 - x) \cdot \left(ecdf_{g'}(x)\right)\right)$$

Thus, the lower threshold represents an optimal point for calling a gene not expressed whereas the upper threshold represents an optimal point for calling a gene highly expressed.

Following the three previous steps results in the required threshold distributions $T_{lower}$ and $T_{upper}$, which are composed of the thresholds $tr_{low}$ and $tr_{high}$ for each bootstrapping sample, respectively. From a conceptual point of view, it is impossible to determine whether a gene should be considered expressed ('1') if it is expressed at the mean level across cells without additional information. Therefore, we aim to derive step functions that best approximate the actual expression distribution across cell types. For instance, deriving the threshold $tr_{low}$ can be viewed as defining a step-function with jump discontinuity at an expression value of $tr_{low}$ which minimizes the error with the actual probability distribution from the left side of the mean. Conversely, deriving the threshold $tr_{high}$ can be viewed as defining a step-function with jump discontinuity at an expression value of $tr_{high}$ which minimizes the error with the actual probability distribution from the right side of the mean.

## Gene level quantification

scGeneXpress quantifies the gene expression of a query cell population into three levels of expression: low, medium and high compared to a reference. Like in the case of the reference data, low quality cells and underrepresented genes are removed from the query data using the same strategy. Next, each gene is normalized and scaled using the corresponding factors from the reference dataset (c.f. sub-section 'Constructing Threshold Distributions for each Gene'). Using this pre-processed data, the method computes p-values for each gene in each cell to determine in which category its expression falls, i.e., not expressed, mediumly expressed or highly expressed, based on the derived threshold distributions and the genes expression of the query cell population. In this regard, for gene $i$ in cell $c$, two p-values are computed based on the lower and upper expression threshold distributions $T_{lower}$, and $T_{upper}$ namely:

$$p_{i,c}^{lower} = ecdf_{T_{lower}}(G_{i,c})$$

$$p_{i,c}^{upper} = 1 - ecdf_{T_{upper}}(G_{i,c})$$

Given these p-values, gene $i$ in cell $c$ is quantified as follows:

$$d_{i,c} = \begin{cases} 1, p_{i,c}^{upper} \leq 0.05 \\ 0.5, p_{i,c}^{upper} < 0.05 \wedge p_{i,c}^{lower} > 0.05 \\ 0, otherwise \end{cases}$$

Of note, the significance threshold may be changed by the user. Finally, scGeneXpress identifies the overall expression of each gene in the query cell population. For that, the number of cells in which each gene has been classified to be not expressed ('0') is determined and compared it to a binomial distribution with mean $n*p$, where $n$ is the number of cells in the query sample and $p$ is the fraction of expressed genes over all possible genes across cells. In this context, we assume independence of genes such that

the binomial distribution signifies the number of cells in which a gene is expected to be expressed when randomly re-assigning the discretized expression values of the query population. Genes that have been determined to be not expressed in more than $2 \cdot \sqrt{n \cdot p \cdot (1-p)} + n*p$ cells are considered to be not expressed at the population level. Otherwise, the classification will be performed based on whether the gene is mediumly or highly expressed in the majority of cells. Importantly, for the purpose of this manuscript, we subsumed both mediumly and highly expressed genes as being expressed.

### Differential expression analysis

Differential expression analysis was performed for all tissues in the Tabula Muris and Tabula Sapiens cell atlases using Seurat v4 [9]. In particular, the FindAllMarkers function was employed to detect significantly upregulated genes in each cell population compared to the other cell populations in a tissue. The function was invoked with default parameters, i.e., the minimum number of cells expressing a gene must amount to at least 10% of all cells ('min.pct = 0.1') and the log-fold change must be larger than 0.25 to be tested for differential expression ('logfc.threshold = 0.25').

## Results

### Building gene-specific reference distributions for discretizing scRNAseq data

Estimating gene-specific reference distributions is the cornerstone of scGeneXpress (Fig. 1a). For that, we leveraged the large-scale scRNAseq expression atlases Tabula Sapiens and Tabula Muris for human and mouse that contain samples from 22 and 12 organs, respectively [1, 2]. Next, we sampled cells from every cell type such that an equal number of cells from all organs in which a cell type is present is selected and all cell types are equally represented. Subsequently, we applied variance-stabilizing transformations and normalized the expression of each gene by its maximum observed value to estimate empirical cumulative distribution functions for each gene. The thresholds for classifying each gene as expressed is then determined by minimizing the theoretical rate of false positive and false negative categorizations (see Methods for details).

Before assessing the performance of scGeneXpress, we set out to confirm its underlying hypothesis that different genes have different expression ranges and distributions. In this regard, we first observed that the maximum expression in the reference distributions, i.e., after applying variance-stabilizing transformations to the raw scRNAseq data, differs widely between genes ranging from 0.69 to 9.26 (average: 2.89) for human and from 0.69 to 8.71 (average: 2) for mice, respectively (Fig. 1b). Similarly, dispersion estimates of the reference distributions underscores this observation (Fig. 1c). The dispersion for human genes ranges from 8.7∗10^-6 to 16 (average: 0.05) whereas a significantly larger range can be observed for mouse genes (range: 6.2∗10^-5–14783.39; average: 3.93) (Fig. 1c). In summary, we indeed observed a high level of heterogeneity in both the expression ranges and distributions of genes in human and mouse, which supports the hypothesis scGeneXpress is built upon.

### scGeneXpress recovers known cell markers and identity genes

To demonstrate that scGeneXpress accurately detects expressed genes, we first set out to assess its ability to detect the expression of known marker genes. In this regard, we obtained all human and mouse markers from Cellmarker 2.0 [13] and selected those that were obtained from reviews, low throughput experiments or have been suggested by companies. Since scGeneXpress is not exclusively designed to predict markers but more generally detect expressed genes, we assessed the fraction of recovered markers across 73 human and 32 mouse cell types (Fig. 2a, b). As a result, scGeneXpress was able to recover on average 51.5% (SD: 23.9) and 46.5% (SD: 23.4) of human and mouse marker genes, respectively. Importantly, we compared the performance of scGeneXpress against three widely employed statistical differential expression tests implemented in Seurat [9] and observed that scGeneXpress recovered significantly more known marker genes than any other method (Fig. 2a, b). In particular, scGeneXpress showed equal or better performance across cell types in 79.3%, 79.3% and 93.1% of cases compared to Poisson-, Wilcox- and T-test (Supplementary Fig. 1a). Finally, we interrogated the relation between the number detected expressed genes by scGeneXpress and differentially expressed genes computed by other methods. Due to the inherent conceptual difference between differential expression analysis and the detection of expressed genes, we unsurprisingly found that scGeneXpress generally considers more genes to be expressed than there are differentially expressed genes (Supplementary Fig. 1b). Only the Poisson-test resulted in a higher number of differentially expressed genes in six cell types. Nevertheless, exemplified by Keratinocytes of the tongue, the higher number of differentially expressed genes does not necessarily translate into a better recovery of known marker genes. Moreover, the number of genes detected to be expressed is significantly smaller compared to the use of ad hoc thresholding approaches (Supplementary Fig. 1c). Interestingly, in some cases, only a low number of genes where detected to be differentially expressed. Since differential expression methods rely on the comparison with other cell populations in the same sample, they fail in cases where the majority of cells is similar, such as in the thymus (Supplementary Fig. 1b). As another independent validation strategy, we collected identity genes, i.e., genes that have been linked to the function of a cell type, from AmiGO [14] and compared the performance of scGeneXpress to differential expression-based methodologies. As a result, we again observed a significantly better performance of scGeneXpress recovering on average 27.8% (SD: 15.2) and 32.9% (SD: 19.5) of known mouse and human identity genes (Fig. 2c, d). Like in the case of recovered markers, scGeneXpress is performing equal or better in every cell type, which cannot solely be explained by the higher number of detected expressed genes (Supplementary Fig. 2a, b). In addition, the number of detected expressed genes is again significantly lower compared to an ad hoc thresholding approach.

### scGeneXpress detects functionally relevant genes and improves differential expression analysis

Encouraged by the recovery of known marker and cell identity genes, we sought to interrogate the relationship between the detected expressed genes by scGeneXpress and differentially expressed genes obtained using the Wilcox-test, the best performing differential expression method, in 73 human cell types. As we expected, scGeneXpress considered differentially expressed genes to be expressed (Fig. 3a). However, in a few cases, more than 50% and up to 96% of differentially expressed genes were not found to be expressed. Coincidentally, these cases occurred in cell populations having only a low number of cells, which indicates that statistical significance could not be achieved for most cases. For instance, there were only 10 myofibroblasts in the lung and 10 Schwann cells in the tongue, which hindered

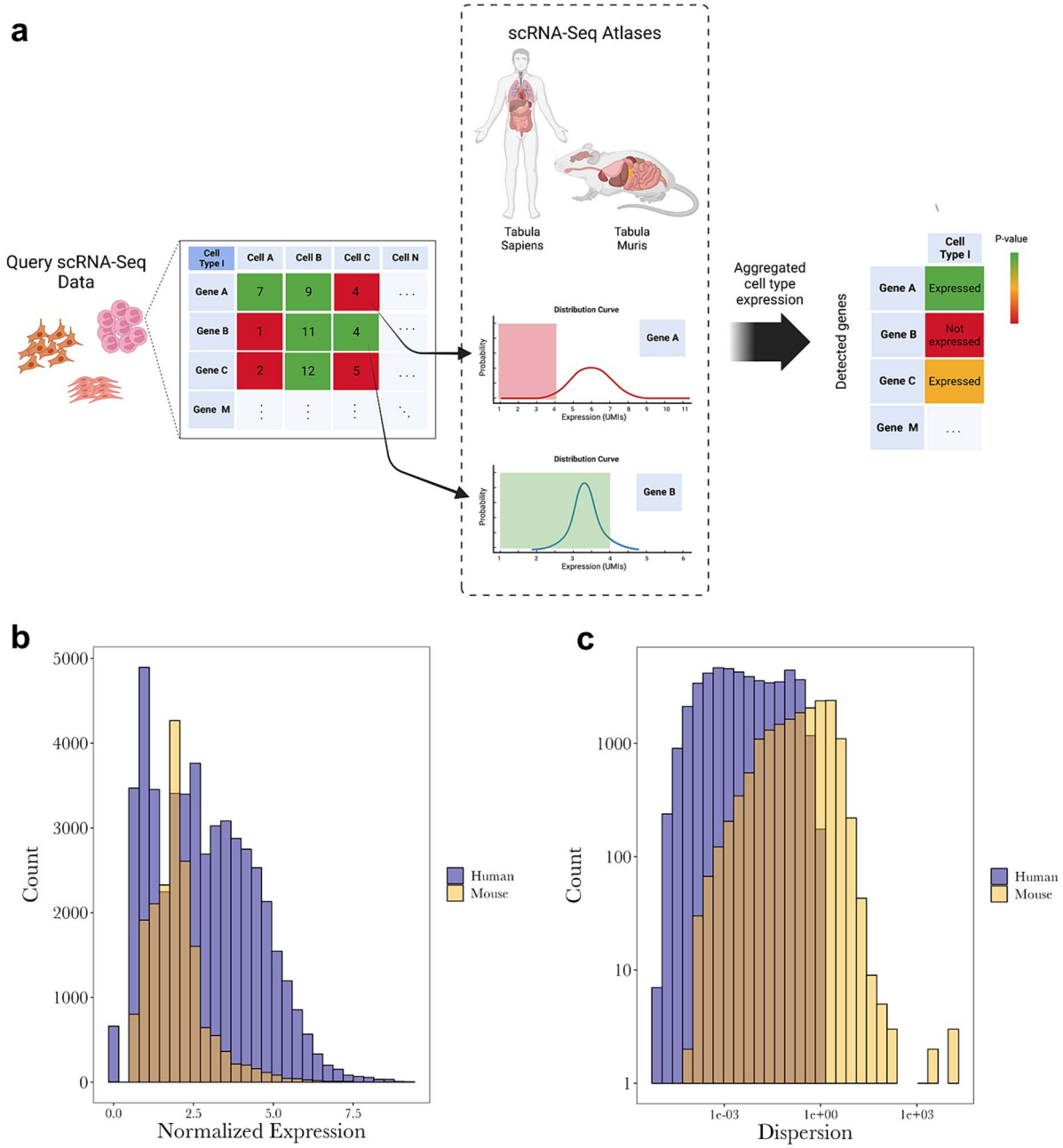


Figure 1. **scGeneXpress workflow and validation.** (**a**) Overview of the scGeneXpress workflow. Given an expression matrix of a single population from an scRNAseq sample, scGeneXpress detects whether expressed genes in each cell by comparing their expression to gene-specific reference distributions generated from tabula Muris and tabula sapiens. Finally, a population-level assessment of the expression of the gene is obtained. (**b**) Histogram of the maximum observed expression values in the tabula Muris (yellow) and tabula sapiens (blue) atlases for all genes after variance-stabilizing transform. Genes have been divided into 30 bins. (**c**) Histogram of the dispersion parameters of fitted negative binomial distributions for each gene after normalizing gene expression data in the tabula Muris (yellow) and tabula sapiens (blue) atlases. Genes have been divided into 30 bins. Both x- and y-axis are on log-scale.

the detection of significantly expressed genes (see Methods for details). Nevertheless, Gene Ontology (GO) enrichment of detected expressed genes by scGeneXpress shows high cell type specificity. In particular, we collected GO term-gene-cell relationships from AmiGO and performed GO enrichment of expressed genes for all cell types in all tissues of the Tabula Sapiens atlas [1, 14]. Subsequently, we selected significantly enriched GO terms that are specific to the respective cell types. For instance, in the case of liver hepatocytes, we identified five terms that correspond to key cell type functions, such as lipid metabolism, response to fatty acids and iron homeostasis (Fig. 3b). This pattern appeared to be consistent across cell types and tissues (Supplementary Figs. 3-24). In addition, as expected, we also found GO terms corresponding to general cellular functions carried out by many or all cell types, such as RNA splicing, protein localization or protein complex assembly, when considering all significantly enriched categories.

Finally, we interrogated whether our method can also improve the performance of traditional differential expression methods. To do so, we collected a gold standard dataset of naïve and specialized T cells for which the transcriptional landscape was profiled at the bulk and single-cell level [15]. When applying

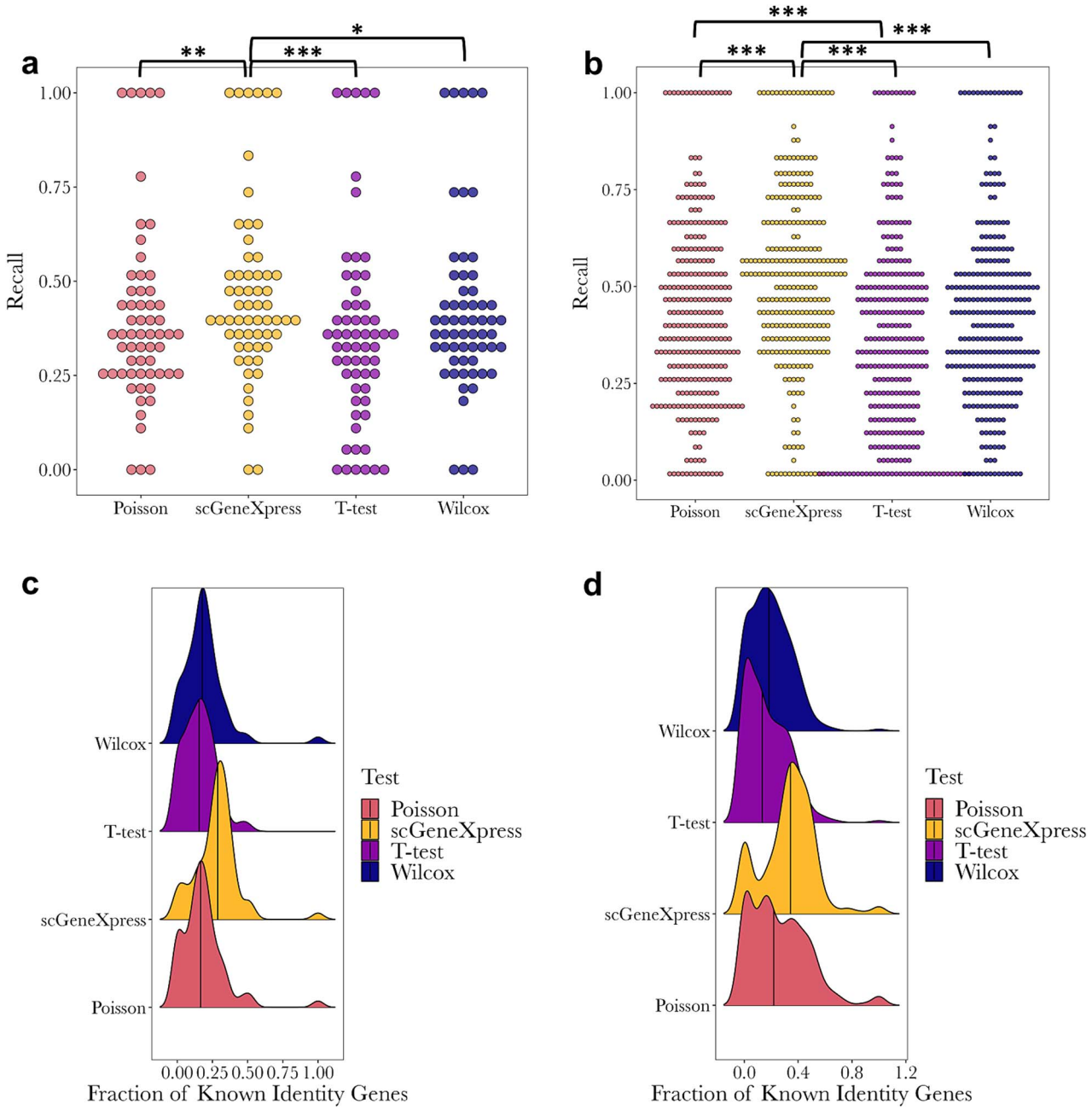


Figure 2. **Detection of cell markers and identity genes using scGeneXpress.** (**a, b**) Dotplot of the fraction of recovered marker genes for 73 and 32 (**c**) human and (**b**) mouse cell types for scGeneXpress and three differential expression-based methods. Significant differences in the average recovery were assessed using a paired t-test (significance levels: ∗: $P < 0.01$; ∗∗: $P < 0.001$; ∗∗∗: $P < 0.0001$) (**c, d**) Ridgeplot of the fraction of recovered functional identity genes for 73 and 32 (**c**) human and (**d**) mouse cell types for scGeneXpress and three differential expression-based methods. Vertical lines in each distribution indicate the mean. scGeneXpress recovers significantly more identity genes than any other differential expression-based method (paired t-test, $P < 0.0001$).

scGeneXpress as well as traditional differential expression methods on the single-cell data, we observed wide discrepancies in the number of detected differentially expressed genes (Fig. 3c). Generally, scGeneXpress and a poison test detected the highest number of differentially expressed genes across comparisons. However, when compared to differential expression analysis results on bulk data, the number of detected genes was significantly lower in almost all cases. Especially when unstimulated T cells were involved, at most 50% of genes detected in bulk RNA-seq have been found in the single-cell samples. We then set out to quantify the concordance between the differentially expressed genes detected by single-cell methods, including scGeneXpress, and in the bulk gold standard analysis (Fig. 3d). Strikingly, we observed that scGeneXpress showed the highest performance with an average Cohen's Kappa of 0.189 whereas the Wilcox test, which showed the best performance of the traditional differential expression methods in recovering identity and marker genes, performed worst (Average Kappa: 0.145). Moreover, especially in cases where Th2 cells were involved, scGeneXpress has a demonstrably higher concordance with bulk data than any other method. Moreover, when comparing the differentially expressed genes of all methods, we observed that

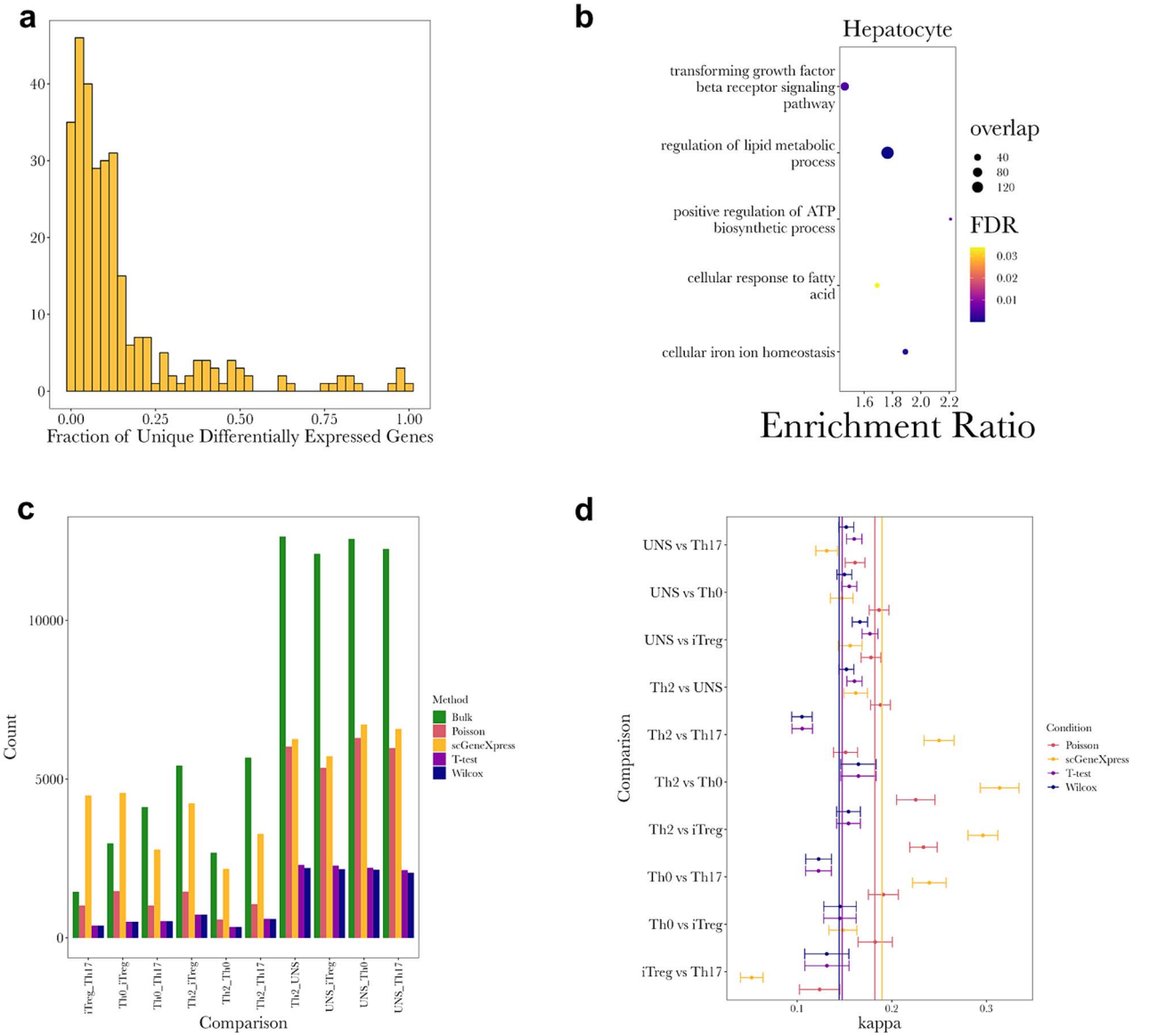


Figure 3. **Detection of functional genes and improvement of differential expression analysis.** (**a**) Histogram of the fraction of genes uniquely detected to be differentially expressed. (**b**) Top enriched cell type specific GO terms resulting from the expressed genes in liver hepatocytes as detected by scGeneXpress. The size of each circle is proportional to the number of overlapping genes and the color encodes the false discovery rate (FDR). (**c**) Number of differentially expressed genes in a dataset of human T cells based on bulk gold standard data ('bulk'), traditional single cell RNA-seq differential expression tests and scGeneXpress. (**d**) Concordance of differential expression analysis in a dataset of human T cells between a bulk gold standard, traditional single-cell RNA-seq differential expression tests and scGeneXpress. Concordance was measured using Cohen's kappa. Each dot represents the computed kappa value and the depicted interval corresponds to the 95% confidence interval.

more than 48% of genes are uniquely detected by scGeneXpress (Supplementary Table S1). In contrast, only up to 3.8% of differentially expressed genes are uniquely detected by other methods. Thus, we conclude that detecting expressed genes in cell populations of single-cell RNA-seq experiments using scGeneXpress can significantly improve the detection of differentially expressed genes.

## Computational requirements of scGeneXpress

We evaluated the runtime and memory requirements of scGeneXpress. Since the quantification of query samples is conducted within seconds, we specifically assessed the computational requirements for computing the threshold distributions depending on the size and composition of the reference data as well as the number of bootstrap samples. For this task, we employed heart tissue data from Tabula Sapiens [1] and varied (i) the number of initial cells per cell type while selecting a total of 2000 cells (with replacement), (ii) the number of total cells, and (iii) the number of bootstrap samples. In addition, for the first two cases, 1000 bootstrap samples were drawn whereas in the third case a total of 2000 cells were considered. All assessments have been carried out on a High-Performance Computing cluster using 15 cores on Intel Xeon CPU. As expected, the number of cell types has little influence on the runtime and memory usage, which only increase slightly after 50 cells per cell type (Fig. 4a, b). The reason for the abrupt increase in runtime and memory usage between 10 and 50 cells per cell type is due to the number of genes passing the filter of being expressed in at least 10 cells. In contrast, the number of total cells is almost linearly related to peak memory consumption and sub-additively to runtime (Fig. 4c, d). Finally, increasing the

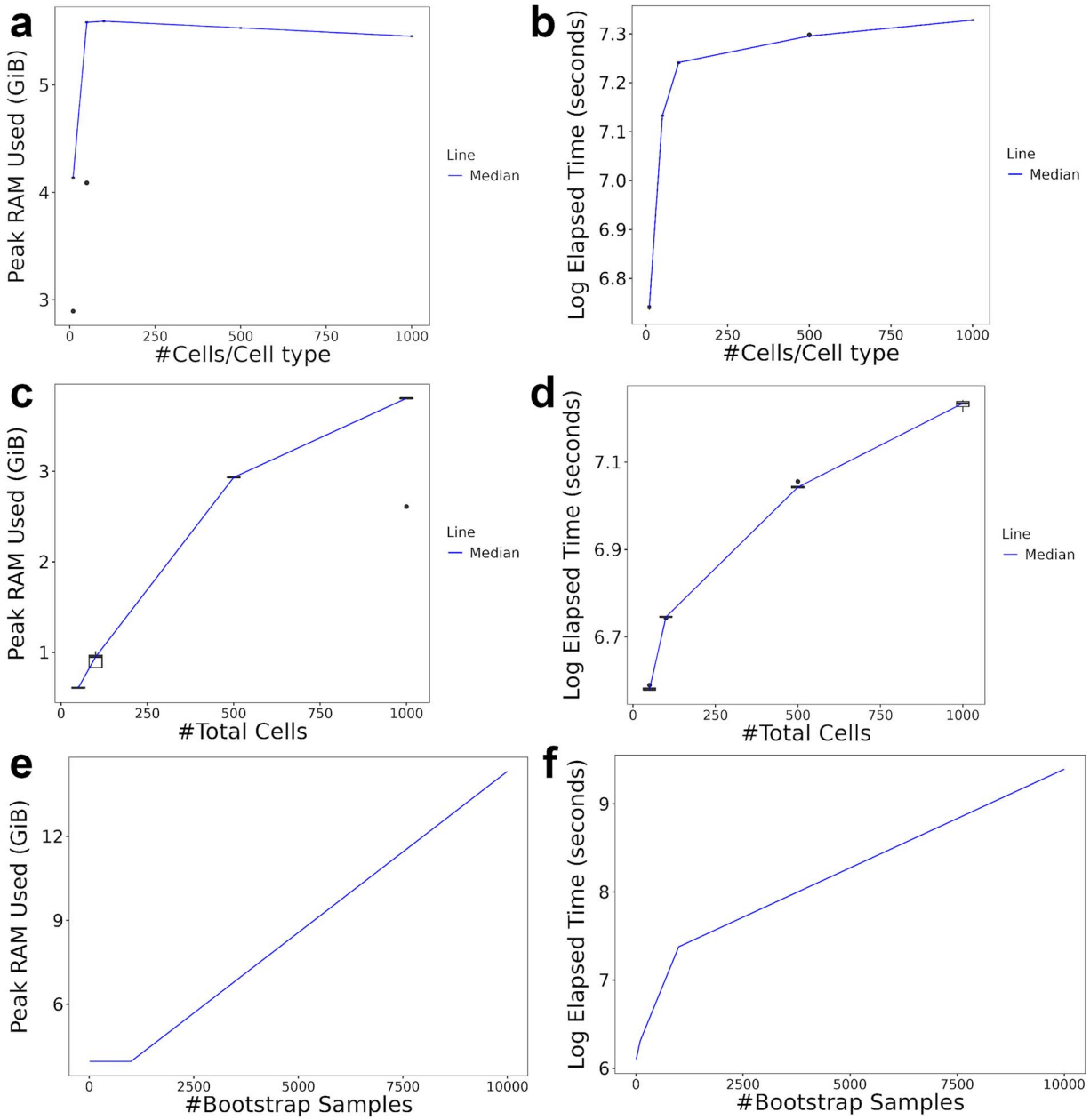


Figure 4. **Computational resources required by scGeneXpress.** Runtime and peak memory usage for generating threshold distributions in three scenarios: (i) varying the number of cells per cell type, subsequently sub-sampling to 2000 total cells and 1000 bootstrap samples (**a, b**), (ii) 1000 cells per cell type, varying the number of totals cells and 1000 bootstrap samples (**c, d**) and (iii) 2000 total cells and varying the number of bootstrap samples (**e, f**).

number bootstrap samples is linearly increasing peak memory consumption and sub-exponentially runtime (Fig. 4e, f).

## Discussion & Conclusion

In this study, we have developed scGeneXpress, a computational method for quantifying single-cell RNAseq expression. One of the key aspects of scGeneXpress is its reliance on data-derived reference distributions for each gene, which allow the identification of gene-specific thresholds to determine whether a gene can be considered to be expressed or not. The method offers several advantages compared to traditional differential expression testing. Namely, using these gene-specific reference distributions allows to account for the individual expression ranges of each gene and considers their usual expression across cell types. Consequently, genes that are specifically expressed in only a few cell types will have lower thresholds to consider them expressed compared to, for instance, housekeeping genes. Moreover, our reference-based approach is resilient to variation in the cell type composition of a query sample, since cell types are treated independently. In fact, all cell types together are only considered for estimating the parameter of a binomial distribution that is used to assess whether a gene can be considered to be expressed at the cell type level. In addition, our approach is agnostic of cluster annotations in the query data as long as the clusters have been correctly identified. Another important aspect of scGeneXpress is its generality. Although we have created the reference distributions across different cell types in this study, it can work

with any background dataset. For instance, reference distributions based on subtypes of the same cell type would enable the detection of expressed genes in a subtype specific manner. Despite the theoretical framework behind scGeneXpress, we demonstrated that the use of gene-specific reference distributions improves multiple aspects of single-cell data analysis, including marker gene detection and differential expression analysis. Although scGeneXpress offers a wide range of advantages over traditional differential expression testing, it has some limitations. In particular, its performance depends on the clustering of the data that is used to assemble the reference distributions for each gene. While it is not necessary to annotate the identified clusters to cell types or subtypes a priori, they need to have the right granularity. In particular, clusters with varying granularity, i.e., belonging to both cell types and subtypes, could introduce biases towards certain cell types. Similarly, inaccurately clustered query samples, in which cells of different types are grouped together, adversely affect the detection of expressed genes for each cluster. Indeed, since statistical significance is assessed by comparing the number of cells expressing a gene in a given cluster against random permutations of the query data, grouping different cell types will eventually lead to insignificant results at the cluster level. Nevertheless, it should be noted that the obtained result is valid given the data provided to scGeneXpress. Moreover, although scGeneXpress can effectively remove batch effects characterized by mean shifts by leveraging scTransform [12], it may be affected by complex batch effects. In particular, SCTransform normalizes data using a generalized linear model, which relates sequencing depth to gene expression counts. This method computes Pearson residuals, providing a variance-stabilized measures of gene expression. The main idea is to reduce the dependency between a gene's average expression and its variability across cells, cutting down on technical noise while preserving the biological signal [12]. scGeneXpress keeps a record of the normalization factors derived for each gene after normalizing the single-cell matrix with SCTransform. These factors are then used to normalize query datasets, ensuring that different datasets remain consistent and comparable. This approach effectively handles mean-shift batch effects, a common issue in single-cell RNA sequencing experiments. However, we acknowledge that our method might not fully address more complex batch effects, which often arise in the analysis of complex experimental designs, such as those involving multiple sample categories or matched samples [16, 17]. Given this limitation, we suggest that users who think their datasets might have these complex batch effects take extra pre-processing steps. Specifically, they should consider correcting batch effects in both the background and query datasets to suit their unique needs before applying our method.

In summary, we expect that scGeneXpress will be a useful tool to obtain insights into key cellular processes in a wide variety of research domains including the study of development and disease progression that can be incorporated into existing single-cell analysis workflows.

**Key Points**

- scGeneXpress is a novel method for quantifying single-cell gene expression data without the need to compare against other cell populations contained in the same dataset
- scGeneXpress considers real gene expression distributions as background references, thus accounting for the different expression scales of genes
- Expression data quantified with scGeneXpress improves differential expression analysis and marker gene detection

## Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

## Acknowledgements

Figure 1a was created with Biorender.com.

## Author contributions

S.J. developed the method, performed the analysis, created figures, wrote the manuscript and supervised the computational work; C.B. developed the method and wrote the manuscript; A.M.P. performed the analysis and created figuress. A.d.S supervised the project, conceived the idea and wrote the manuscript.

## Conflict of interest

The authors declare that they have no competing interests.

## Funding

This work was supported by the Luxembourg National Research Fund [PRIDE17/12244779/PARK-QC to C.B.]; the Spanish Ministry of Science and Innovation [PID2020-118605RB-I00 to S.J.]; and the New Frontiers in Research Fund (NFRFT-2022-00327 to A.d.S).

## Data availability

The data underlying this article are available in the article and in its online supplementary material. scGeneXpress is available as an R package at: https://github.com/saschajung/scGeneXpress

## References

1. Tabula Sapiens Consortium*, Jones RC, Karkanias J. *et al.* The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* 2022;**376**:eabl4896. https://doi.org/10.1126/science.abl4896.
2. Pisco AO, McGeever A, Schaum N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a tabula Muris. *Nature* 2018;**562**:367–72. https://doi.org/10.1038/s41586-018-0590-4.
3. Keren L, Hausser J, Lotan-Pompan M. *et al.* Massively parallel interrogation of the effects of gene expression levels on fitness. *Cell* 2016;**166**:1282–1294.e18. https://doi.org/10.1016/j.cell.2016.07.024.
4. Brown JC. Role of gene length in control of human gene expression: Chromosome-specific and tissue-specific effects. *Int J Genomics* 2021;**2021**:8902428. https://doi.org/10.1155/2021/8902428.
5. Harrison PW, Wright AE, Mank JE. The evolution of gene expression and the transcriptome-phenotype relationship.

*Semin Cell Dev Biol* 2012;**23**:222–9. https://doi.org/10.1016/j.semcdb.2011.12.004.

6. Taylor SR, Santpere G, Weinreb A. *et al.* Molecular topography of an entire nervous system. *Cell* 2021;**184**:4329–4347.e23. https://doi.org/10.1016/j.cell.2021.06.023.
7. Replogle JM, Saunders RA, Pogson AN. *et al.* Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell* 2022;**185**:2559–2575.e28. https://doi.org/10.1016/j.cell.2022.05.013.
8. Squair JW, Gautier M, Kathe C. *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* 2021;**12**:5692. https://doi.org/10.1038/s41467-021-25960-2.
9. Hao Y, Hao S, Andersen-Nissen E. *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.
10. Jung S, Hartmann A, del Sol A. RefBool: A reference-based algorithm for discretizing gene expression data. *Bioinformatics* 2017;**33**:1953–62. https://doi.org/10.1093/bioinformatics/btx111.
11. Choudhary S, Satija R. Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biol* 2022;**23**:27. https://doi.org/10.1186/s13059-021-02584-9.
12. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* 2019;**20**:296. https://doi.org/10.1186/s13059-019-1874-1.
13. Hu C, Li T, Xu Y. *et al.* CellMarker 2.0: An updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res* 2023;**51**:D870–6. https://doi.org/10.1093/nar/gkac947.
14. Carbon S, Ireland A, Mungall CJ. *et al.* AmiGO: Online access to ontology and annotation data. *Bioinformatics* 2009;**25**:288–9. https://doi.org/10.1093/bioinformatics/btn615.
15. Cano-Gamez E, Soskic B, Roumeliotis TI. *et al.* Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4+ T cells to cytokines. *Nat Commun* 2020;**11**:1801. https://doi.org/10.1038/s41467-020-15543-y.
16. Rahman MA, Tutul AA, Sharmin M. *et al.* BEENE: Deep learning-based nonlinear embedding improves batch effect estimation. *Bioinformatics* 2023;**39**:btad479. https://doi.org/10.1093/bioinformatics/btad479.
17. Kharchenko PV. Publisher correction: The triumphs and limitations of computational methods for scRNA-seq. *Nat Methods* 2021;**18**:835. https://doi.org/10.1038/s41592-021-01223-2.